インターネットにおける情報の知的利用
－情報収集から情報統合へ－

奈良先端科学技術大学院大学
武田英明
takeda@is.aist-nara.ac.jp

１．　はじめに

WWW は急速に社会に浸透している。多くの情報提供源が WWW を介して情報を提供し、WWW はもはやそれなしにはこの情報化社会が成り立たないほど、社会システムの一部となっている。

WWW はとにもかくにも情報流通機構として画期的である[1]。これは WWW がシステムとして抜きんでた存在であったというよりは、世の中のニーズとインフラの現状に時期的にうまく適合したことによることが大きい。

本稿では WWW における情報利用について利用者の立場から何が必要とされているか、そしてどんな試みがなされているかについて論じる[2]。

２．　WWW のよさ・悪さ

我々は日常的に WWW を利用し、そのメリット、そしてデメリットを日々感じている。

A. WWW はいかに使えるか

では、その WWW のメリットとは何であろうか。

（１） とにかく多量の情報がある。

WWW ページの総量は実際のところ不明である。サーチエンジン goo[3]では 1 億 2000 万ページ、国内だけでも 1700 万ページあるという[4]。大量の情報の蓄積は、組織内のデジタル情報の蓄積の公開に加え、既存の情報提供源の WWW 化によって加速されている。

（２） 多種多様な情報がある。

これまでの情報源は特定の目的の情報の提供を目的としていた。WWW はそういったこれまでの情報源が対象としてきた情報に限らず、ニッチ情報とでもいえる分類不能な情報も提供している。これがまた WWW の情報の増大を促している理由でもある。

（３） 検索が使える。

この多量かつ多様な情報を一括して検索することができる。サーチエンジンを利用した検索は画期的であり、これほど情報の世界を身近にしたものはない。たいていのキーワードを入れればとりあえず何かが返ってくるのである。

B. WWW はいかに使えないか。

こういったメリットがあるものの、使えば使うほど不満が溜まるものでもある。なぜだろうか。

（１） とにかく雑多である。

多種多様な情報があることは反面、よい分類がないということでもある。これはいざ使うときには面倒なことである。

（２） 屑情報が多い

またジャンク情報、すなわち役に立たない情報が多いのも WWW の特徴である。これは情報提供が誰でもできるということの裏面である。

（３） 検索がうまくいかない。

サーチエンジンによる検索の結果からほしいものを見つけるのは至難の業である。それは WWW の情報が増えるにつれ、ますます加速されている。

３．　何が必要とされているか －利用者の視点－

WWW の抱える問題点の起源の一つは、WWW が情報提供者のためにデザインされており、情報利用者のためのデザインが欠けている点であろう。

---

[1] 本稿で WWW という場合は、WWW サーバ、WWW ブラウザ、さらにはキャッシュサーバなど WWW を通じて情報を伝達させる仕組み全体を指す。

[2] ネットワーク情報の流通そのものについては[1]を参照されたい。

[3] http://www.goo.ne.jp

[4] 1998 年 10 月 14 日プレスリリースによる。

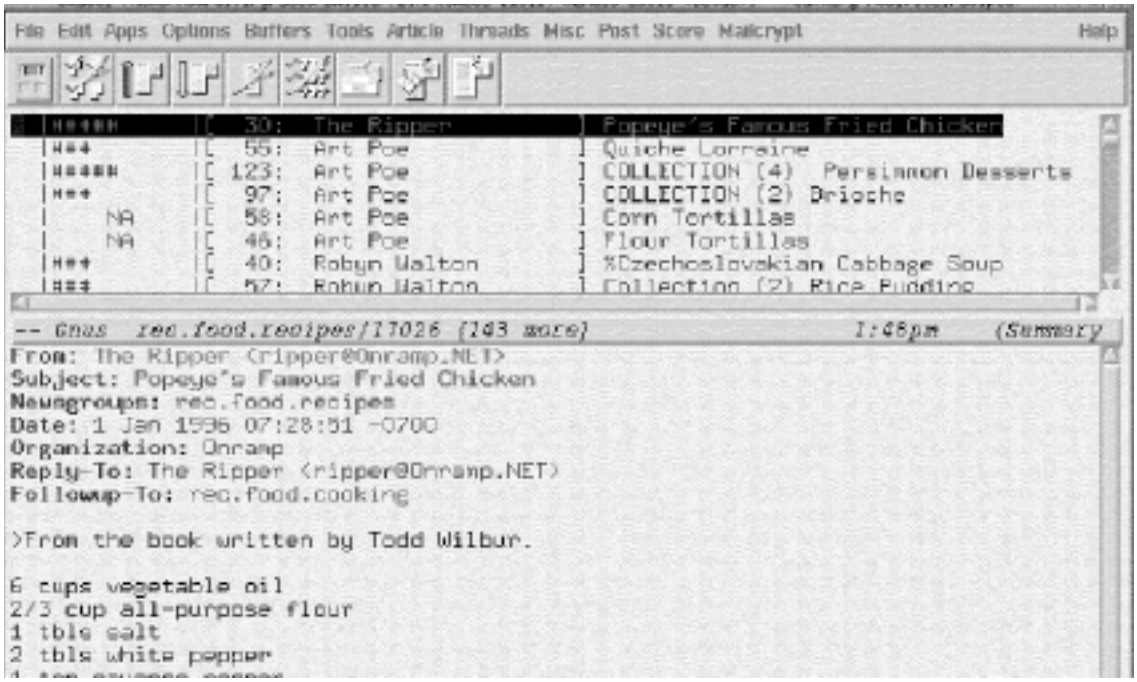

図１ GroupLens の表示例[3]

WWW は多様かつ多数の情報提供者の情報提供を可能にする仕組みである。多様かつ多数の情報提供者があるということは、それ以上に多様かつ多数の情報利用者がいることである。この情報利用者の多様性や規模といった点に対して WWW はシステムとして対処していない点が前節で挙げたような不満を生むもとになっている。

ではどうすればよいのか。まず、情報利用者の立場を分析しよう。

（１） 情報利用者はそれぞれ異なる。

情報利用者はそれぞれ自らの意図を持って情報を利用しているので、それぞれ情報利用者としての立場は異なる。WWW の場合はその違いは、単に好みの違いといえるものから情報利用形態（ブラウジングしているのか、特定の情報を探しているかなど）の違い、対象領域の違いまである。

（２） 情報利用者は相互依存的である。

その反面、情報利用者の情報利用の立場はそれぞれが単独で存在するものではなく、多くの情報利用者と明示的／暗黙的にあるいは直接的／間接的に関係している。広く社会での関心事に興味を持つ人間間の関係は暗黙的かつ間接的な関係であるし、ある現実のコミュニティで俎上にあがる話題を共有する人々は明示的かつ直接的な情報利用者の関係をもっているといえる。これらの人間間の関係が情報利用の立場に影響を与えている。

（３） 話題は情報提供者と情報利用者の間にある。

また情報利用の立場当然、提供される情報によって変化する。提供される情報の内容の変化は情報利用者の立場の変化を導く。WWW は現実には日々変化する情報源であり、情報内容の変化は日々生じている。

（４） 情報利用者は情報提供者でもある。

多くのアクティブな情報利用者は(潜在的な)情報提供者である。単に情報提供者か情報利用者という分類ではなく、情報世界への参加者である。

（１）（２）（３）という問題に関しては情報フィルタリングという技術がひとつの解決の方法を提供している。しかし、情報フィルタリングは比較的均質な情報源を対象にしてきたので情報利用者の立場の違いは好みの違いという範囲に留まっていることが多い。また、（４）の問題には対処していない。

これに対して人のネットワークに着目するという研究が近年盛んになっている。これは情報フィルタリングのように、情報提供者と情報利用者あるいは、情報と情報利用者の関係を求めることが解なのではなく、人と人との関係を求めることが解であるという点で、新しい視点を提供している。具体的には直

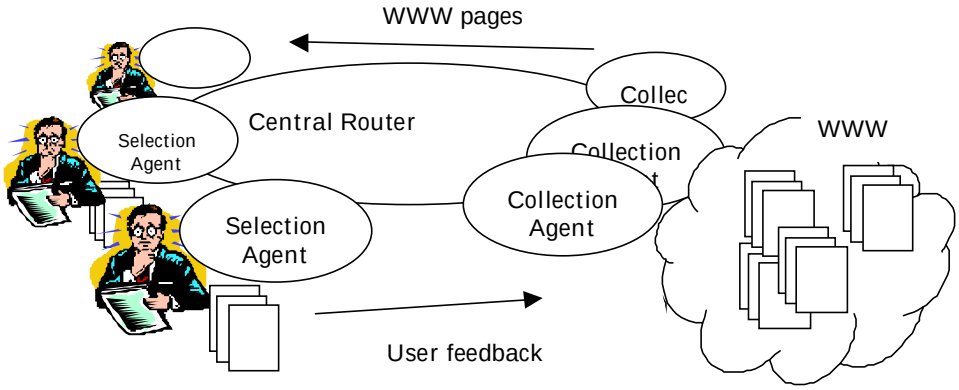

図 2 Fab のアーキテクチャ

接的な関係を求めるのか、コミュニティの形成など間接的な関係を求めるかによって異なってくる。
以下では、情報フィルタリングから人と人のネットワークの研究まで現在行われている研究を概観する。

４．　ユーザの好みを知る

ユーザの好みを知る古典的な方法は情報フィルタリングである。情報フィルタリングとは多量の情報からユーザが必要とする情報だけを選別してユーザに提示するという方法である。近年は選別というよりは推薦であるという意味で、情報推薦 (information recommendation)と呼ばれることもある。情報フィルタリングには情報からキーとなる特徴を抽出してその特徴を利用して行う内容に基づくフィルタリングと、情報とユーザ間の関係を収集することによりユーザの好みを推定する社会的フィルタリングがある。

４．１ GroupLens[2][3]

ネットワークの情報に対する情報フィルタリング・システムの先駆的な研究としては GroupLens がある。これは netnews(ネットワーク上の議論グループ)の記事を対象にしている。Netnews の情報の特徴は（１）議論毎にグループが分けられているものの、実際にはかなり雑多な記事が多いこと、（２）多量であること、（３）動的であること（日々記事は投稿される)、（４）読者は誰でも投稿者になれる[1]、であり、WWW の情報がもつ特徴と似ている。GroupLens ではこの netnews の情報をソシアル・フィルタリング(social filtering)あるいは協調的フィルタリング(collaborative flitering)という方法で選別する。ソシアル・フィルタリングとはユーザの好みを集め、あるユーザの好みと類似するほかのユーザの好みを探し、それらを用いて情報を選択／推薦するというしくみである。この方法の特徴は一切情報の内容に触れない点であり、このため原理的にはどんな情報源に対しても適用可能である。
ソシアル・フィルタリングのしくみを簡単な例題で説明する。今、ユーザの好みは 1 から５で与えられるとし、ある時点で表１のような好みが収集できているとする。このとき、A さんに対して 6 の記事をどう推薦したらよいか考える。
まず A さんと他の人の相関係数を求める。ユーザ数 m、記事数 n において相関係数は次式である。

$$r_{kk'} = \frac{Cov(k,k')}{\sigma_k \sigma_{k'}} (k = 1 \ldots m, k' = 1 \ldots m)$$

$$r_{kk'} = \frac{\sum_{l=1}^{n'} (x_{kl} - \bar{x}_k)(x_{k'l} - \bar{x}_{k'})}{\sqrt{\sum_{l=1}^{n'} (x_{kl} - \bar{x}_k)^2} \sqrt{\sum_{l=1}^{n'} (x_{k'l} - \bar{x}_{k'})^2}}$$

ただし、ｎ’は記事からどちらかの値のない記事を除いたものである。例えば、

$$r_{AB} = \frac{-2 \cdot 1 + 2 \cdot (-1) + (-1) \cdot 2 + 1 \cdot (-2)}{\sqrt{4+4+1+1}\sqrt{1+1+4+4}} = -0.8$$

$$r_{AC} = \frac{-2 \cdot (-1) + 2 \cdot 1}{\sqrt{4+4}\sqrt{1+1}} = 1$$

となる。そこで、評価見積りは重み付け平均でもとめることができる。

表１ 好みの例

記事	A さん	B さん	C さん	D さん
1	1	4	2	2
2	5	2	4	4
3			3	
4	2	5		5
5	4	1		1
6	?	2	5	

[1] そうでない議論グループもある。

図３ PHOAKS の画面

$$x'_{kl} = \bar{x}_k + \frac{\sum_{k' \neq k} (x_{k'l} - \bar{x}_{k'}) r_{kk'}}{\sum_{k' \neq k} |r_{kk'}|}$$

先の場合、

$$x'_{A6} = 3 + \frac{(-1)\cdot(-0.8) + 2\cdot 1}{0.8 + 1} = 4.56$$

となる。評価値の予測は他にも近いユーザだけを利用するなどいろいろ考えられる。

実際の表示は図２のようになる。各行の先頭についている部分が推薦度合いを示している。

協調的フィルタリングの方法は内容的な問題に立ち入らないため情報内容の変化などにロバストである。しかし、誰も評価しないデータには決して推薦されないなど、ある程度規模が大きくならないと効果がでないなどの問題がある。

４．２ Fab[4]

Fabは協調的フィルタリング／推薦だけでなく、内容に基づくフィルタリング／推薦(content-based filtering/recommendation)も合せて用い、WWWのページを収集提示するシステムである。

図 2 に示すようにユーザ毎の選択エージェント(selection agent)と複数の収集エージェント(collection agent)からなる。選択エージェントはユーザのプロファイルを持ち、収集エージェントはトピックスのプロファイルを持つ。

選択エージェントの持つユーザのプロファイルはユーザのフィードバックから内容に基づく推薦によって生成する。このプロファイルはまた類似のユーザを判定するのにも使われる。すなわち、ユーザは自分のプロファイルで高く評価される情報と類似するユーザが高く評価する情報の両方を受けとる。

収集エージェントの持つプロファイルはあるトピックスに関するものであり、収集エージェント総体としてユーザの興味を表現する。このため、収集エージェントは消滅やコピーをしながら全体として適応的に振る舞う。これは情報と収集エージェントの関係と収集エージェントとユーザの関係を最適にするということである。WWW においては netnews の場合と異なり、情報全体を対象に推薦の対象にするわけにはいかない。このため、収集側にもユーザの好みを反映することが必要になる。

４．３ Lifestyle Finder[5]

Lifestyle Finder では協調的フィルタリングではユーザのクラスタリングがうまくいかないことを指摘し、人口統計学的データを用いて、ユーザを分類する方法を提案している。あるデータではアメリカの人口調査、雑誌購読、カタログ利用者と嗜好によってアメリカの人口を 62 のクラスターに分けている。ある少数の質問をすることで、ユーザをいずれかのクラスターに分類する。人口統計学的データは実際には詳細すぎることがあるので、その場合抽象

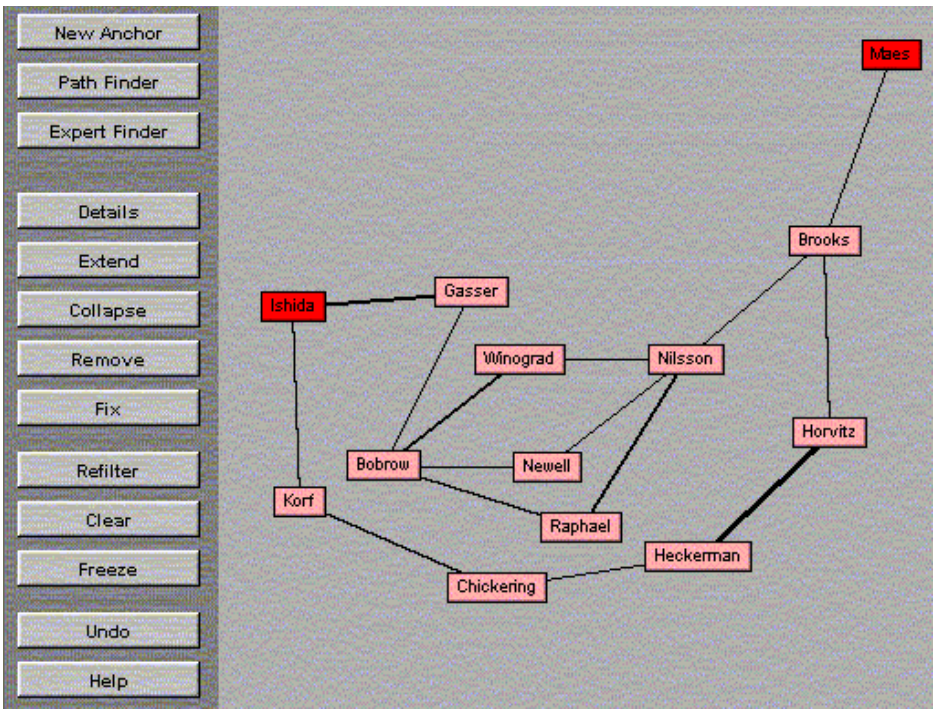

図４ Referral Web の表示

化を行うことで質問を限っている。例えばあるクラスターのデータで各ニュース番組視聴の割合がある程度あれば、そのクラスターは「ニュース番組視聴を好む」と特徴づけることができる。
このシステムの結果は「何を買うべきか」「どこへ行くべきか」「どこで買うべきか」についてのWWW ページの推薦である。ユーザのフィードバックはクラスターの部分集合をより適合させるのに使われる。

５．　話題を探し当てる

前節では個人の好みの推定が主眼であった。個人の好みには明示的あるいは暗黙的に必要とする話題を推定しているとはいえるが、ここではより直接的に話題を探し当てる方法の研究を概観する。

５．１ PHOAKS[6]

WWW の検索エンジンで「いい」URL をみつけることは困難である。ディレクトリ型サービス（Yahoo!など）では一般的な分類であり、特定の話題はないことが多い。こういう時はその分野をよく知っている人たちに聞くの一番いいわけで、実際netnews にはそういう質問と回答がよくある。それを自動化したのが PHOAKS である。
PHOAKS は netnews の記事においてよく推薦される URL を集めて提示するシステムである。図３にその表示例を示す[1]。しくみとしては（１）記事の選別（spam 記事などの排除)、（２）URL の選別(署名･宣伝 URL の排除、推薦 URL の発見)からなる。（２）の推薦 URL の発見には典型的な言い回しなどのヒューリスティックスを用いている。
この他 Siteseer[7]では人の bookmark そのものを使って類似 URL を発見しようとする方法もある。

５．２ FAQ Finder[8][9][10][11]

Netnewsでは多くの議論グループでFAQ(Frequent Asked Questions)と呼ばれる良くある質問とその回答のリストが用意されている。これをうまく検索しようというのが FAQ Finder である。このシステムの技術的にユニークな点は、浅い自然言語処理に限って利用することでも一定の機能が果たしていることである。
FAQ Finder ではユーザの質問に対して、(1)適切な FAQ ファイルの選択、(2)適切な QA の組の選択、２つの段階からなる。(1)では特徴ベクトルを用いて判定している。(2)でも特徴ベクトルを用いているが、さらに WordNet[12]の単語のネットワークを用いて単語の類似性を計算して用いている。

６．　人と人のネットワークを発見する

ネットワークにある情報は我々の持つ知識のほんの一部である。むしろ本当に知りたいことは個人的に

[1] http://www.phoaks.com//index.html

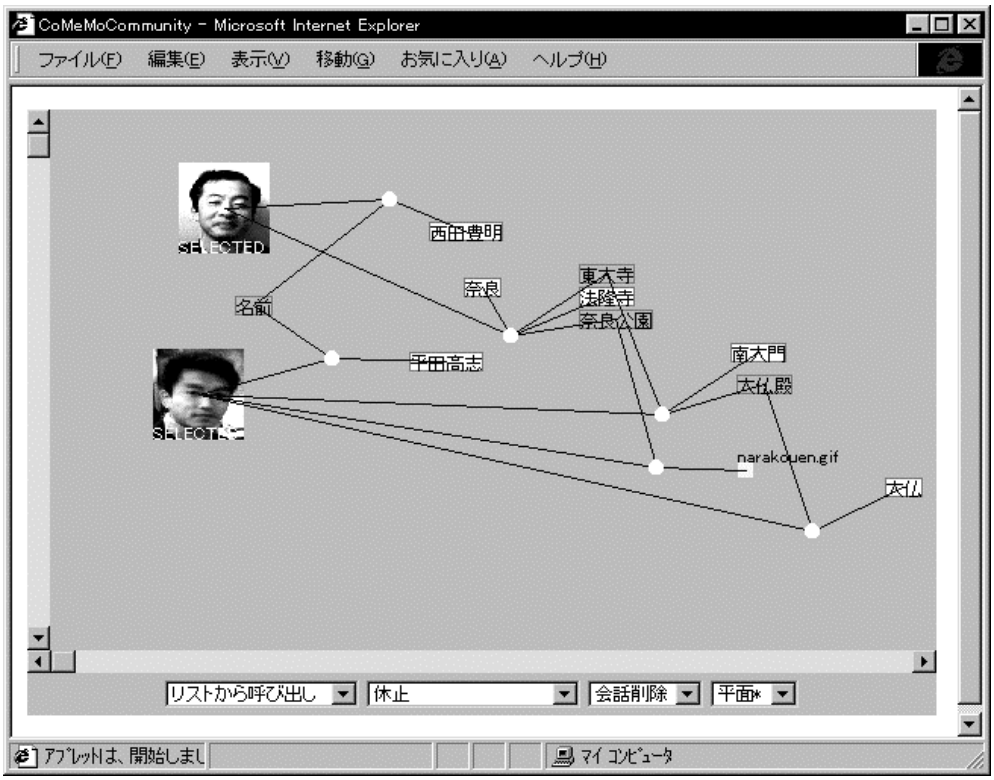

図５ CoMeMoCommunity の表示

聞くことでわかる。ふつう我々はわからないことがあれば、近くのわかる人に聞きに行く。こういった過程をサポートしようするアプローチがある。人と話題の関係、人と人の関係これらを探すのここでのテーマである。

６．１ Referral Web[13][14]

これは人のつながりを利用して情報を獲得するという日常生活の仕組みを情報ネットワークに適用したものである。すなわち、ネットワーク上の情報（ここでは論文などの書誌データや WWW ページ）を利用して、人と人の関連性を調べ、知りたいトピックスがあった場合、その人のネットワークを探索して、自分と関係している人でそのトピックスを知っている人を推薦するというシステムである。
図４はその例で[1]、「Ishida」と「Maes」という人の関連を表示させたものである。この場合、６つのリンク４経路で関連があることがわかる。
同様に起点の人物とトピックスを指定することで起点からそのトピックスに近い人までの経路が表示される。

[1] http://www.research.att.com/~kautz/referralweb/

６．２ Contact Finder[15]

このシステムはネットワーク上の情報からトピックスに詳しい人を推薦するシステムである。このシステムでは Intranet の netnews を対象にして、そこでの問答の記事から答えている人を推定することで、トピックスと人物を結び付けている。回答記事の推定は答えるときの典型的な言い回しを利用し、トピックスの推定は大文字で始まるなどのヒューリスティックスを利用している。

６．３ CoMeMoCommunity[16][17]

このシステムでは連想表現で表現された各人の知識から共通性を動的に提示することで人と人との関連性を認識させるというシステムである。連想表現はある概念が他の概念を想起するとだけの簡単なものである。この連想表現を各人に自由に作ってもらう。そのあと、人の代理人であるエージェントが共通の場において順次共通の話題を提示していく。このシステムの特徴は関係の発見を仮想的な対話として時間経過をもって提示することである。図５にその例を示す。

６．４ ICMAS Mobile Assistant Project[18]

ここまでは情報世界での人の awareness であった。しかし、人は情報世界の中だけに存在するわけ

図６ C-MAP の画面例

ではない。むしろ、現実の世界での存在の方が重要である。ICMAS Mobile Assistant Project は現実のコミュニティのメンバーがどう情報ネットワークによって支援可能化を実験したプロジェクトであり、京大、奈良先端大、NTT が共同で行ったものである。これは物理的な参加者の awareness と情報空間での awareness をいかに結び付けられるかという意味で興味深い実験である。

国際会議の参加者に携帯端末(MagicLink)を持たせ、その上でコミュニケーションを促進するような実験プログラムを走らせた。具体的には通常の電子メイルや掲示板の他に、周辺情報案内 (Action Navigator、NTT)、個人カスタマイズ型情報共有(InfoCommon、奈良先端大)、コミュニティサポートサービスとしては出会い支援 (Social Matchmaking、京大)を用意した。InfoCommon では個人の興味に応じた情報検索、発信の支援、情報のゆるやかな関連づけを用いた静的な情報と動的な情報の統合･構造化といった機能をグラフィカルユーザインタフェースで提供した。これはコミュニティにおける情報活動「調べる→尋ねる→人を知る→発言する」の過程を支援するためであった。

６．５ C-MAP[19]

前項と同様に物理的な awareness と情報的 awareness を融合するシステムとして C-Map がある。これは研究展示の場において、ユーザの物理的存在（どの展示場所にいるか)、ユーザの興味、展示物の内容相互の関連を提示するというシステムである。

このシステムではユーザの興味と展示内容は 2 次元空間での力学的リンクとしてあらわされ、ユーザの物理的存在とこれらとの関連は life-like なエージェントの振る舞いとして表示されている。図６にその例を示す。

７． まとめ

本稿ではネットワーク情報をいかに使うかという立場から、関連する研究を概観した。使う立場を考えるということは実は情報世界への参加者としてのユーザを再定義することであった。しかしまだ多くのことが未解決である。例えば以下のことは今後さらに考察していかなければならないであろう。

（１） 情報世界において personality とはなにか
（２） 情報世界においてコミュニティ形成とはなにか
（３） 情報世界においてコンテクストとはなにか

本稿が以上のことを考える上で多少でも参考になれば幸いである。

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] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. GroupLens: Applying collaborative filtering to usenet news. Communications of the ACM, Vol. 40, No. 3, pp. 76-87, 1997.

[3] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In In Proceedings of the 1994 Computer Supported Cooperative Work Conference, pp. 175-186, New York, 1994. ACM.

[4] Marko Balabanovic and Yoav Shoham. Fab: Content-based, collaborative recommendation. Communications of the ACM, Vol. 40, No. 3, pp. 66-72, 1997.

[5] Bruce Rulwich. Lifestyle finder. AI Magazine, Vol. 18, No. 2, pp. 37-45, 1997.

[6] Loren Terveen, Will Hill, Brian Amento, David McDonald, and Josh Creter. PHOAKS: A system for sharing recommendations. Communications of the ACM, Vol. 40, No. 3, pp. 59-62, 1997.

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

[8] Robin D. Burke, Kristian J. Hammond, Vladimir Kulyukin, Steven L. Lytinen, Noriko Tomuro, and Scott Schoenberg. Qeustion answering form frequent asked question files. AI Magazine, Vol. 18, No. 2, pp. 57-66, 1997.

[9] Robin Burke, Kristian Hammond, and Julia Kozlovsky. Knowledge-based information retrieval from semi-structured text. In 1995 AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval, pp. 15-19, 1995.

[10] Kristian Hammond, Robin Burke, and Steven L. Lytinen. A case-based approach to knowledge navigation. In Proceedings of IJCAI-95, pp. 2071-2072, 1995.

[11] Kristian Hammond, Robin Burke, and Charles Martin. FAQ Finder: A case-based approach to knowledge navigation. In AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Envrionments, pp. 69-73, 1995.

[12] George A. Miller. WordNet: A lexical database for english. Communications of the ACM, Vol. 38, No. 11, pp. 39-41, 1995.

[13] Henry Kautz, Bart Selman, and Mehul Shah. Referral web: Combining social networks and collaborative filtering. Communications of the ACM, Vol. 40, No. 3, pp. 63-65, 1997.

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

[15] Bruce Rulwich and Chad Burkey. ContactFinder agent: Answering bultin board questions with referrals. In AAAI-96, pp. 10-15, 1996.

[16] Takashi Hirata, Harumi Maeda, and Toyoaki Nishida. Facilitating community awareness with associative representation. In Proceedings Second International Conference on Knowledge-Based Intelligent Electronic Systems (KES'98), volume 1, pp. 411-416, 1998.

[17] 平田高志, 前田晴美, 西田豊明. コミュニティにおける連想表現を用いた自己開示と知識共有. 人工知能学会研究会資料(SIG-FAI-9802), pp. 63-68, 1998.

[18] 石田亨, 西村俊和, 八槇博史, 後藤忠広, 西部喜康, 和氣弘明, 森原一郎, 服部文夫, 西田豊明, 武田英明, 沢田篤史, 前田晴美. モバイルコンピューティングによる国際会議支援. 情報処理学会論文誌, Vol. 39,, 1998. (採録決定).

[19] 角康之, 江谷為之, SidneyFels, NicolasSimonet, 小林薫, 間瀬健二. C-map: Contextaware な展示ガイドシステムの試作. 情報処理学会論文誌, Vol. 39, No. 10, pp. 2866-2878, 1998.