潜在的なコミュニティの可視化

吉田 仙　　亀井 剛次　　横尾 真　　大黒 毅　　船越 要　　服部 文夫

NTT コミュニケーション科学研究所
619–02 京都府相楽郡精華町光台 2–2
phone: 0774–95–1821　　fax: 0774–95–1871
{yoshida,kamei,yokoo,ohguro,kf,hattori}@cslab.kecl.ntt.co.jp

梗概

This paper explains the development of a system that graphically presents potential communities (groups with common interests) for a certain mass of people to support in the forming of new communities. The system locates users in a two-dimensional space where the distance between users reflects the relevance between them. The characteristics of this system is that it obtains user profiles automatically using information published on the WWW; it provides a customized view for each user by a personal agent; and a user can broadcast messages to relevant people using the provided view. We also describe a method where each agent learns the appropriate weights of keywords that represent user profiles by using user feedback. It is shown that agents can speed up this learning process by communicating the results they learned with each other. The effectiveness of the developed methods is shown by experimental evaluation.

1 はじめに

インターネットやモバイルコンピューティングシステムのような公共コンピュータネットワークの普及にともない，個人や企業の活動は現実社会だけでなくネットワーク上の仮想社会でも行われるようになりつつある．この仮想社会をより魅力的で便利なものにするために，ネットワーク上でのグループの形成や相互理解，インフォーマルなコミュニケーションを支援する環境が望まれている [5]．このような環境の総称として，コミュニティウェアという言葉が提唱されている [4]．

コミュニティウェアとグループウェアを比較すると，グループウェアはすでに組織化された人々（企業におけるプロジェクトのメンバーなど）における協調作業を支援するものであり，そのユーザは LAN を通して固定的な相手と通信する．これに対しコミュニティウェアは，より多様で組織化されていない集団を対象にする．コミュニティウェアの目的は，

- 共通の興味を持つ人を探しコミュニティを形成すること
- コミュニティ内でのインフォーマルな対話や集団活動

- コミュニティ内の情報統合，コミュニティの組織化

といったことの支援である．

本論文では，コミュニティウェアの例として，上の一番目の目的，すなわち新たなコミュニティの形成を支援するシステムについて説明する．このシステムは，国際会議の参加者やボランティア活動の志願者のような集団を対象に，潜在的なコミュニティ，すなわち共通の興味を持つ人々を可視化する．具体的には，システムは興味の近さの度合い（関連度）をそれぞれの人と人の間について求める．そして，ユーザ間の距離がそれら関連度を反映するように，各ユーザを二次元平面上に配置する．本システムは，各ユーザに特化したビューを持つ個人エージェントを提供する．さらに，特化されたビューを用いたメッセージの同報も可能である．

本システムの特徴の一つは，各ユーザが WWW 上に公開している情報をもとに，システムが自動的に各ユーザのプロファイルを作成することである．これによりユーザが自らのプロファイルを入力する手間を軽減でき，またプロファイルを公開することにともなうプライバシーの問題を避けられる．こうして得たプロファイルからキーワード集合を抽出し，ユーザ間の関連度を計算する．

潜在的なコミュニティを可視化する上で，適切なキーワードを選びだすことは，本システムを実装する上で重要な点の一つである．プロファイルから抽出されたキーワード集合は，しばしば，ユーザを特徴づけるのに重要でなかったり関係ないようなキーワードを含む．したがって，ユーザ間の関連度の計算において，すべてのキーワードを等しく用いることは適当でない．この問題を解決するため我々は，ユーザからのフィードバックを用いて，各エージェントがユーザのプロファイルを表すキーワード間の適切な重み付けを学習する手法を開発した．各エージェントが学習結果を交換することによって，学習過程を高速化できることが示される．ある国際会議の発表者のデータを用いた評価実験では，エージェント間の協調を導入することにより，60% の recall と 59% の precision を得るために必要なユーザからのフィードバックの回数を，約 85 回から 50 回に減らすことができることが示される．

同様な目的を持つシステムが，ICMAS'96 モーバイルアシスタントプロジェクト [4, 5] という国際会議の参加者を支援する実験において提供された [7]．我々のシステムは，

- WWW から得られる情報からのプロファイルの作成
- 各ユーザに特化したビューを持つ個人エージェントの提供
- 特化されたビューを用いて興味の近さにもとづくメッセージの同報を実現

という点において [7] のシステムとは異なる．

本論文では，まず第 2章で我々が開発したシステムの詳細について述べる．次に第 3章で各エージェントがキーワード間の適切な重み付けを学習する手法を提示し，第 4章では学習過程を高速化するためにエージェント間の協調機構を導入する．本システムの手法の評価結果は第 5章で示す．

2 システムの詳細

本システムは，ユーザ間の興味の近さに基づき，潜在的なコミュニティを可視化する．システムは興味の近さの度合い（関連度）をそれぞれの人と人の間について求める．そして，ユーザ間の距離がそれら関連度を反映するように，各ユーザを二次元平面上に配置する．つまり，関連の深いユーザ同士は近くに，浅いユーザからは遠くに配置される．ユーザは，自身の周

りの分布を見ることによって，潜在的に自分に関係するコミュニティを見つけることができる．このようなシステムによって，ユーザ間での出会いや情報交換が促進されることが期待できる．

システムがユーザ間の関連度を求めるには，ユーザのプロファイルを取得しなければならない．ユーザにとっては，プロファイルを記入するのは面倒で時間のかかる作業である．また，プロファイルを公開するのは，ユーザのプライバシーを侵害しかねない．このような問題を避けるため，システムはすでにWWW上に公開されている情報を用い，情報検索 [6] の技術を用いてユーザのプロファイルを取得する．

WWW上に公開されている個人情報からその人の特徴を把握するには，WWWのページからキーワードを抽出する必要がある．WWWのページのような文書からそれを特徴づけるキーワードを抽出する過程は，“indexing” として知られている [6]．しかしながら，そのような手法を単独でWWWページに適用しても，WWWのページは研究分野，経歴，趣味など雑多な情報を含んでいるため，望ましい結果は得られない．そこで我々は，システムが扱うドメイン（“ボランティア活動”，“研究・プロジェクト” など）に関する情報を，ユーザに関連するWWWページの選択やindexingに利用する．

キーワード抽出の手順は以下の通りである．システムはまず，与えられたURLからリンクをたどり，WWWページを自動的に収集する．このとき，より関連があると思われるリンク，例えばそのドメインに関連する語句によってタグ付けされたリンクなどを，優先的にたどる．つぎに，収集されたページから語句[1]が統計的手法を用いて抽出され，それら語句のうちドメインに関連するものだけがキーワードとして選択される[2]．このときstemming dictionary[6] を用いて品詞や綴り，単数複数，簡略形と複合形の変化を吸収する．

各キーワードの各ユーザのWWWページ中での出現頻度を調べることにより，そのユーザの特徴ベクトルが作成される．その特徴ベクトルから各ユーザ間の関連度がcosine measure [8, 1] を用いて計算され，関連度行列が出来上がる．

システムはこの関連度行列をユーザに見せるのであるが，このとき，サーチエンジンのように単に自分との関連度が高いものから順番に並べただけでは，自分以外のユーザ同士の関係はわからない．潜在的コミュニティ，すなわち興味を同じくする人々の集団を表示するには，他人同士の関連度も考慮にいれる必要がある．

そこで本システムでは，自分と他のユーザとの間の関連度だけでなく，他のユーザ同士の関連度をも可視化するために，ばねモデル [2] を用いる．具体的には，各ユーザ間に仮想的なばねが置かれる．ここで，各ばねの自然長は，それがつながっているユーザ間の関連度に比例する．そして，すべてのばねの張力が均衡するように各ユーザが配置される．

この配置は各ユーザに対応するエージェントによって行われ，そのユーザはビューの中央に配置される．また，ばねの強さは，中央のユーザとそれ以外のユーザの間の自然長がより保たれるよう調節される．すなわち，中央のユーザとそれ以外のユーザの間のばねは強く，まわりのユーザ間のばねは弱くする．このため，関連度行列がすべてのエージェントで共通であっても，作成されるビューは各ユーザに特化したものとなる．

図1は，ボランティア活動の志願者のためのビューの例である．ビュー上の各アイコンはそれぞれ一人のユーザを表し，ビュー中央のアイコンはそのエージェントが対応しているユーザを表す．アイコンをクリックすると，それに相当するユーザの情報が，図の下部のように表示される．

[1] 単語または句

[2] 現在，この選択は手動で行われているが，シソーラスを用いて自動化することも可能．

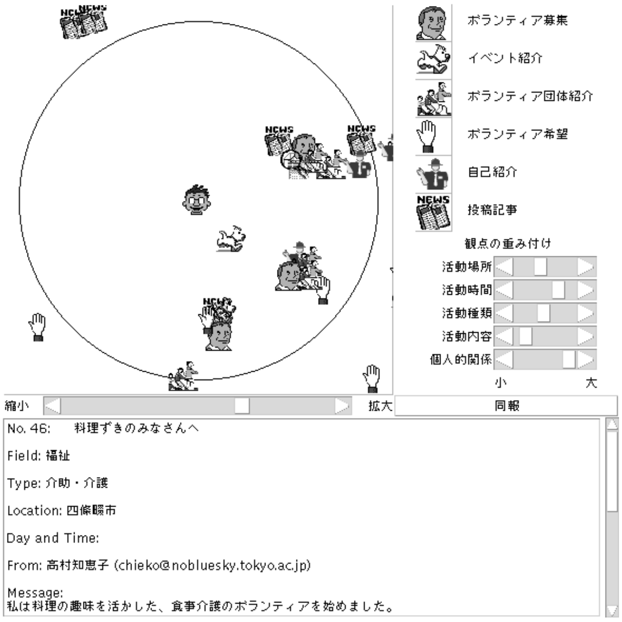

図 1: システムの動作イメージ

本システムは興味の近さにもとづく同報の機能を持つ．すなわち，ユーザは，自分からある一定の範囲内にいる全ユーザに対してメッセージを同報することができる．同報の範囲（ビュー中に円で示される）はズームのスライドバーで調節できる．この機能によって，メーリングリストのメンバーのような固定的なグループに対する通信でなく，境界が曖昧な潜在的コミュニティのメンバーに対する通信が実現される．

3 キーワードの重み付けのフィードバック

前章において，ユーザの WWW ページからキーワードを半自動的に抽出する手法について述べた．しかしながら，プロファイルから抽出されたキーワード集合は，ユーザを特徴づけるのに重要でなかったり関係ないようなキーワードを含みうる．不要なキーワードはユーザ間の関連度の計算において誤差を生み，ビューの配置に悪影響をおよぼす．したがって，ユーザ間の関連度の計算において，すべてのキーワードを等しく用いることは適当でない．このため我々は，関連度の計算において，キーワード間に重みを付け，この重み付けの望ましい値をユーザからのフィードバックによって学習する機構を導入する．

ユーザからのフィードバックは次のように与えられる．ユーザははじめに，自分が中央に配置されているビューを見る．そこで自分と興味を同じくしないと思われる人が自分のすぐ近くに配置されていると，それをエージェントに伝える．するとエージェントは，示された人と自分との間の関連度が減少するよう，キーワードの重み付けを変更する．そして，新しい重み付けを用いて関連度を再計算し，変化したビューを表示する．

詳細な手順は以下の通りである．ユーザの人数を n，キーワードの数を m，ユーザ i の特徴ベクトルを

$$\vec{f_i} = \langle f_{i,1}, \ldots, f_{i,m} \rangle$$

キーワードの重み付けを

$$\vec{w} = \langle w_1, \ldots, w_m \rangle$$

とする．i の重み付けされた特徴ベクトルは，

$$\begin{aligned}\vec{g}_i &= \vec{w}\vec{f}_i^T \\ &= \langle w_1 f_{i,1}, \ldots, w_m f_{i,m} \rangle\end{aligned}$$

である． cosine measure 関数 cm により，重み付けされた特徴ベクトルの集合 $\langle \vec{g}_1, \ldots, \vec{g}_n \rangle$ に対して i の関連度ベクトル

$$\begin{aligned}\vec{r}_i &= \langle r_{i,1}, \ldots, r_{i,n} \rangle \\ &= \langle \mathrm{cm}(\vec{g}_i, \vec{g}_1), \ldots, \mathrm{cm}(\vec{g}_i, \vec{g}_n) \rangle\end{aligned}$$

が求められる．関連度行列，すなわち関連度ベクトルの集合 $\langle \vec{r}_1, \ldots, \vec{r}_n \rangle$ を用いて，各ユーザのビュー上の座標が計算される．

ユーザはビューを見て，フィードバックの標的（関連のないユーザ）を示す．ここでユーザ i が自分と近すぎるユーザ j を選んだとする．このとき，各キーワード k について， k に対するそれぞれの重み付き特徴量 $g_{i,k}$ および $g_{j,k}$ の差が大きい場合，エージェントは重み w_k を大きくし，差が小さい場合は w_k を小さくする．すなわち，フィードバック関数を fb，特徴量の差の閾値を s とすると，

$$\mathrm{fb}(w_k) = \begin{cases} w_k + \delta & \text{if } |g_{i,k} - g_{j,k}| > s \\ w_k - \delta & \text{otherwise} \end{cases}$$

これにより，i と j で特徴量に差があるようなキーワードがより重視されるようになり，結果として j が i から遠ざかる．

4 協調フィードバック

前章で述べた学習過程は，エージェント間の協調を導入することにより改善することができる．具体的には，エージェント間で重み付けの値を交換することにより，ユーザからのフィードバックの回数を減らすことが可能である．

協調の手順は以下のとおりである．エージェント j による重み付けを

$$\vec{w}_j = \langle w_{j,1}, \ldots, w_{j,m} \rangle$$

とする． j によって重み付けされた i の特徴ベクトルは

$$\begin{aligned}\vec{g}_{j,i} &= \vec{w}_j \vec{f}_i^T \\ &= \langle w_{j,1} f_{i,1}, \ldots, w_{j,m} f_{i,m} \rangle\end{aligned}$$

であり，したがって j によって計算される i の特徴ベクトルは，

$$\begin{aligned}\vec{r}_{j,i} &= \langle r_{j,i,1}, \ldots, r_{j,i,n} \rangle \\ &= \langle \mathrm{cm}(\vec{g}_{j,i}, \vec{g}_{j,1}), \ldots, \mathrm{cm}(\vec{g}_{j,i}, \vec{g}_{j,n}) \rangle\end{aligned}$$

となる．エージェント j は特徴行列 $\langle \vec{r}_{j,1}, \ldots, \vec{r}_{j,n} \rangle$ を用いて j から見たビュー中での各エージェントの座標を計算する．

前章で述べたフィードバックが行われるたびに，エージェントはそれぞれの重み付け $\vec{w}_1, \ldots, \vec{w}_n$ を情報交換する．そして，エージェント j は，受け取った重み付けから新たな重み付けを生成する関数 wg を用いて，重み付け $w_{j,k}$ を $\mathrm{wg}(w_{1,k}, \ldots, w_{n,k})$ に更新する．

重み付け生成関数 wg の実装の方法はいくつか考えられる．我々は二通りの方法でこれを実装した．一つは受け取った重み付けの平均を計算する．

$$\mathrm{wg}(w_{i,k}) = \frac{\sum_{j=1}^{n} w_{j,k}}{n}$$

この場合，すべてのエージェントの重み付けは等しくなる．

別の方法は，受け取った重み付けを加味するときに，関連度を考慮する．つまり，エージェント i は，j から受け取った重みと i と j の間の関連度の値を乗算する．関数 wg は，

$$\mathrm{wg}(w_{i,k}) = \sum_{j=1}^{n} r_{i,i,j} w_{j,k}$$

となる．

5 評価

本章では，これまでの章で述べた手法の評価結果を示す．我々は，国際会議の参加者の支援について評価を行った．第 2 回マルチエージェントシステム国際会議（ICMAS'96）の発表者のうち 40 名の URL を用い，第 2章で述べた手法により 277 個のキーワードを選択した．

5.1 フィードバック手法の評価

フィードバックの性能を調べる手法としては，情報検索における precision-recall[3] を採用した．precision-recall とは，一度の検索で得るデータの個数を変化させて，精度の変わり方を測る手法である．precision は得たデータのうち正しいデータ（実際に関連のあるユーザ）の割合，recall はすべての正しいデータのうち得ることができたデータの割合である．一般に，recall が高くなると precision は低くなる．

ここでは，ユーザは自分と興味を同じくしない人のうちもっとも近くに置かれているものを選んでフィードバックを与えると仮定し，あるユーザが別のユーザと興味を同じくするかどうかは正解行列によって規定されるものとする．本実験で用いる正解行列は 40 個の行と列を持ち，各要素は真理値（関連する／しない）を持つ．この行列は，著者らの知識と直観にもとづいて作成された．

図 2は，第 3章で述べたフィードバック手法による precision-recall の変化である．左右の軸は recall，上下は precision，前後はユーザインプットの回数を表す．グラフから，フィードバックを重ねる毎に precision が向上しているのがわかる．

本実験では，フィードバックを行っても precision は比較的低いままになっている．この原因の一つとして，WWW ページが，頻繁に更新されていないなどの理由で，現在のユーザの特徴を十分に表していないことが考えられる．いくつかの重要なキーワードが完全に抜け落ちているような状況では，高い precision を得られる重み付けを学習することは困難である．仮にシステムが十分なキーワードを取得できる状況にあったときに，本論文で提案した手法が正しく動作するかどうかを見るために，我々は以下のような実験を行った．まず，あるランダムに作成された重み付け（目標となる重み付け）によって重み付けされた特徴ベクトルを用いて関連度を計算する．これと完全に整合する理想的な正解行列を作成し，これを用いてフィードバックを行う．

理想的な正解行列と著者らの知識と直感にもとづいて作成された正解行列を用いて行われた実験結果を，図 3に示す．idealの線は理想的な正解行列を用いて得られた結果，normalの線は著者らによって作られた正解行列を用いて得られた結果であり，recall は 60% である．

著者らによって作られた正解行列を用いた場合の precision は約 55% から 65% であり，理想的な正解行列を用いた場合は約 80% から 95% になる．理想的な状況では，提案された手法が有効であることがわかる．より正しい特徴ベクトルを求めるのに十分なキーワードを取得するためには，なんらかのさらなる工夫が必要である．これには，ユーザから不足分を補完するようなキーワードの入力を要求したり，特徴ベクトルそのものを動的に変化させるような関連フィードバック（relevance feedback）[9] を用いるなどが挙げられる．

5.2 協調手法の評価

図 4は，図 2と同じ条件下での，単独フィードバックと協調フィードバックの比較の結果を示している．左右の軸はすべてのユーザからの入力の回数を表し，上下の軸は recall が 60% のときの precision の平均を表す．グラフ中で，singleの線が協調しないフィードバックの結果を，relevanceが関連度を考

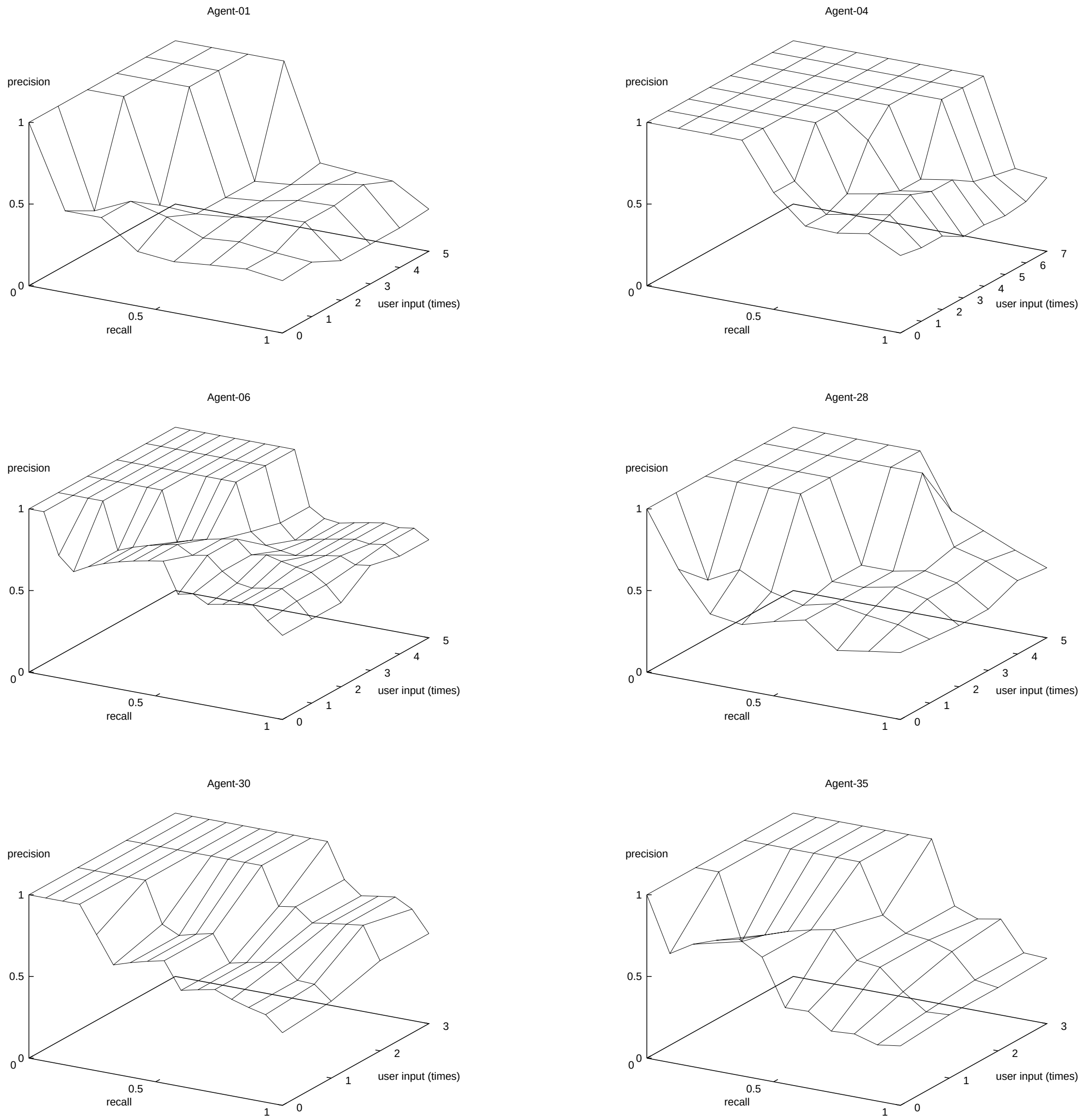

図 2: フィードバックによる precision-recall の変化

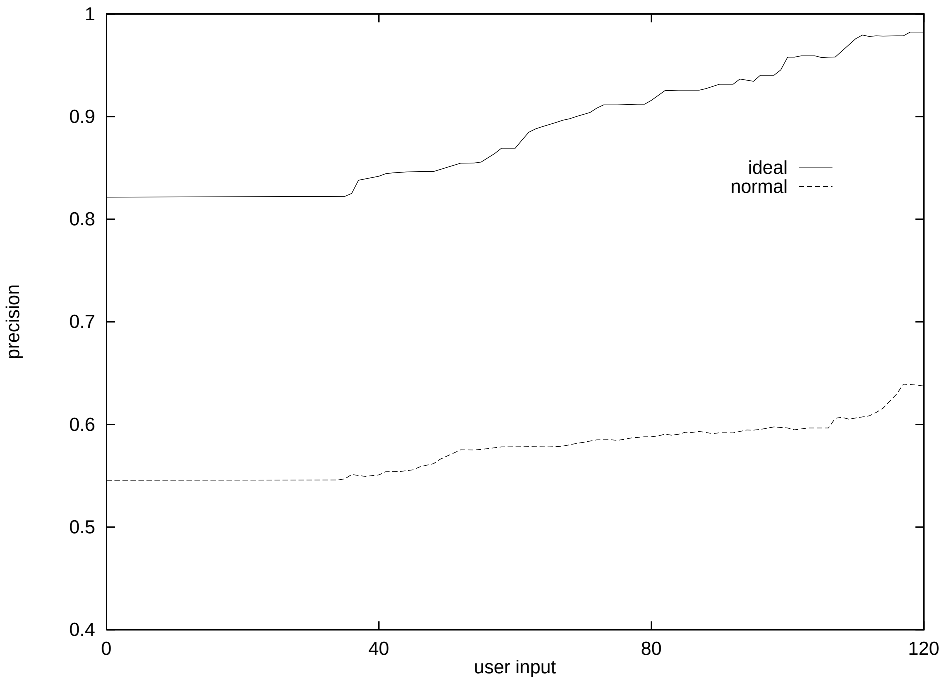

図 3: 理想的な正解行列を用いた結果

慮した協調フィードバックの結果を，averageが平均を用いる協調フィードバックの結果を表す．

60% の recall と 59% の precision を得るために必要なユーザインプットの回数が，エージェント間の協調を導入することによって，約 85 回から 50 回に（一ユーザあたり 2.1 回から 1.3 回に）減少している．協調の方法（平均または関連度考慮）による違いはあまり影響しない．この理由の一つとして，ユーザの数が少ないためユーザの観点に偏りがなかったことが考えられる．望ましい重み付けがユーザ毎に異なるときには関連度の考慮が重要になると思われる．

6　おわりに

本論文では，国際会議の参加者やボランティア活動の志願者のような集団を対象に，潜在的なコミュニティ（興味の近い人々）を可視化することにより，新たなコミュニティの形成を支援するシステムについて述べた．本システムは，各ユーザ間の興味の近さ（関連度）を求める．そしてユーザ間の距離がそのユーザ間の関連度を反映するように，各ユーザを二次元平面上に配置する．

本システムでは，ユーザが自らのプロファイルを入力する手間を軽減するため，またプロファイルを公開することにともなうプライバシーの問題を避けるため，各ユーザが WWW 上に公開している情報をもとに各ユーザのプロファイルを自動的に作成する．また，各ユーザに対するエージェントが存在し，ユーザに特化したビューを提供する．ビューを用いたメッセージの同報通信も可能である．

我々は，ユーザからのフィードバックを用いて，各エージェントがユーザのプロファイルを表すキーワード間の適切な重み付けを学習する手法を開発した．さらに，各エージェントが学習結果を交換することによって，学習過程を高速化できることが示された．ICMAS'96 の発表者のデータを用いた評価実

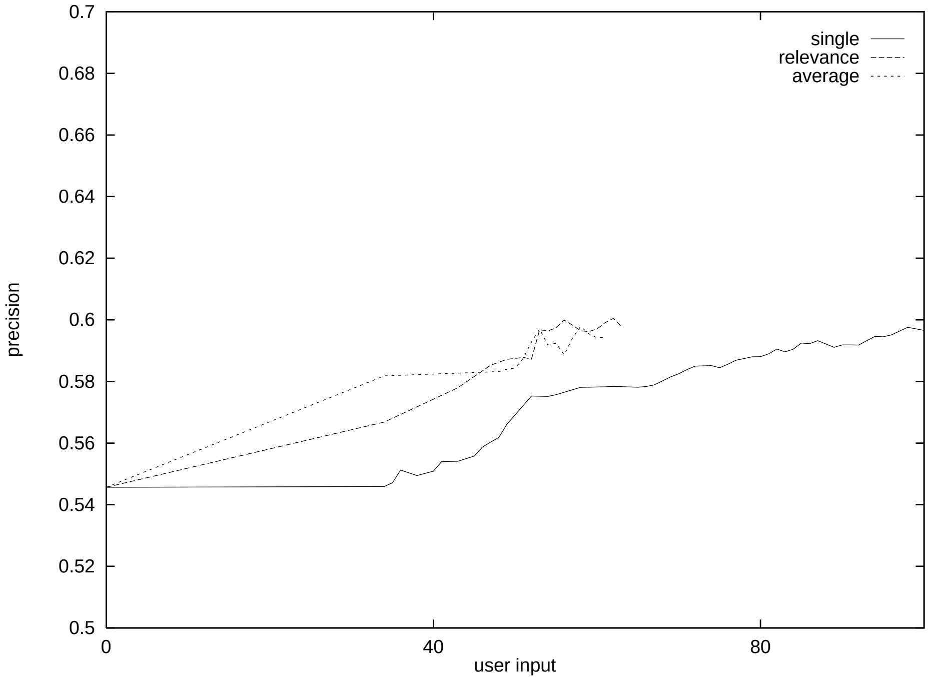

図 4: 単独／協調フィードバックの比較

験では，エージェント間の協調を導入することにより，60% の recall と 59% の precision を得るために必要なユーザからのフィードバックの回数を，約 85 回から 50 回に減らすことができた．

今後の課題としては，precision と recall を向上させる手法を開発することや，望ましい重み付けがユーザ毎に異なる場合の対処法について調べることが挙げられる．本システムを実際の場面で試験することも必要である．

参考文献

[1] Bert R. Boyce, Charles T. Meadow, and Donald H. Kraft. *Measurement in Information Science.* Library and Information Science. Academic Press, 1994.

[2] Matthew Chalmers and Paul Chitson. Bead: Explorations in information visualization. In Nicholas Belkin, Peter Ingwersen, and Annelise Mark Pejtersen, editors, *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR Forum, pp. 330–337. ACM Press, 1992.

[3] Valery I. Frants, Jacob Shaprio, and Vladimir G. Voiskunskii. *Automated Information Retrieval: Theory and Methods.* Library and Information Science. Academic Press, 1997.

[4] Toru Ishida. Towards CommunityWare. In *Proceedings of the Second International Conference and Exhibition on The Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM 97)*, pp. 7–21. The Practical Application Co. Ltd., 1997.

[5] 石田亨, 西村俊和. 広域情報ネットワークによるコミュニティ支援. 情報処理, Vol. 38, No. 1, pp. 48–53, 1997.

[6] Gerald Kowalski. *Information Retrieval Systems: Theory and Implementation.* Kluwer Academic Press, 1997.

[7] Toshikazu Nishimura, Hirofumi Yamaki, Takaaki Komura, Nobuyasu Itoh, Tadahiro Gotoh, and Toru Ishida. Community Viewer: Visualizing community formation on personal digital assistants. In *Proceedings of the IJCAI-97 Workshop on Social Interaction and Communityware*, pp. 25–30, 1997.

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

[9] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, Vol. 41, pp. 288–297, 1990.