論　文

情報フィルタリングにおける遺伝的アルゴリズムを用いたユーザプロファイルの作成手法

正員　柳本豪一（大阪府立大学）
正員　藤中　透（大阪府立大学）
正員　吉岡理文（大阪府立大学）
正員　大松　繁（大阪府立大学）

A Method Creating a User Profile for Information Filtering by Genetic Algorithm

Hidekazu Yanagimoto, Member, Toru Fujinaka, Member, Michifumi Yoshioka, Member, Sigeru Omatu, Member (Osaka Prefecture University)

This paper proposes a method creating a user profile for an information filtering system by genetic algorithm. Generally, user's interest is wide-ranged and unclear. In an information retrieval, it is difficult that a user expresses his query with suitable keywords. A proposed method extracts a user's perference from the documents which the user has read and evaluated. In addition, this method efficiently expores a search space for a user perference. This method uses simple genetic algorithm. We develop an information filtering system, and evaluate selection capability by this method. Finally, we compare this method with a relevance feedback in selection accuracy.

キーワード：情報検索，情報フィルタリング，遺伝的アルゴリズム

1. はじめに

近年，Webページに代表されるように大量に電子化されたドキュメントがネットワーク上で公開されている。インターネットとの接続サービスを提供するプロバイダの登場に伴い，インターネットを利用するユーザも増加してきている。このような状況において，ユーザが必要とする情報を適切に見つけ出すため，サーチエンジンや電子図書館のような検索サービスが登場してきている。しかしながら，ユーザが実際に検索を行ったとしても，検索結果が多すぎて必要な情報が見つからないという情報洪水[1]と呼ばれる現象が発生している。

本論文では，関連研究の調査や最新技術の動向調査などの長期的興味に着目し，上記の問題を解決するため，遺伝的アルゴリズムを用いたユーザの興味抽出手法を提案する。また，提案手法による評価実験を行い，本手法の有効性の確認を行う。

2. 従来手法

従来，ユーザの興味を自動的に抽出する手法としては，関連フィードバック[2]を利用したものが一般的であった。関連フィードバックでは，検索要求をユーザが興味を持つドキュメントに近づけ，興味のないドキュメントから遠ざけることで実現している。実際の処理としては，各ドキュメントがベクトルで表現されているので，ベクトルの和　差により操作を行うこととなる。このため，関連フィードバックでは，ユーザベクトルの変化が，評価されたドキュメントベクトルの向きに限定されてしまうので，ユーザの興味の探索範囲が狭いものとなってしまう。

探索範囲を広げるために，遺伝的アルゴリズムを利用する情報フィルタリングシステム[3][4]が提案されている。このシステムはエージェントシステムであり，遺伝的アルゴリズムを用いて新しいユーザプロファイルを作成し，そのユーザプロファイルを持つエージェントを評価することにより適切な情報の推薦を実現している。しかし，このシステムでは，システムが持っているユーザプロファイルを変化させるために，遺伝的アルゴリズムの交叉と突然変異を用いていただけである。このため，遺伝的アルゴリズムの適合度による遺伝子の選択など，最適値の探索能力を十分に活用していなかった。

また，メールなどの半構造化されたドキュメントを対象とした情報フィルタリングシステム[5]も存在している。この手法では，ドキュメントの構造情報をもとに，自動的にドキュメントの特徴情報を取得し，それをもとに情報の分

類を行っている。また，発信者とユーザの関係をあらかじめ記述しておくことで，人間関係に基づいた情報フィルタリングが可能となっている。しかし，本手法は情報の構造情報を用いているため，ウェブページで公開されているような構造化されていないデータを扱うことは困難である。

他にユーザの興味を抽出するシステムとしては，興味表現システム[6]がある。この手法では，ユーザの短期的な興味に着目し，ユーザが入力した検索要求から興味の抽出を行っている。しかし，本手法はユーザの興味への追従性に注目しているため，長期的な興味抽出にそのまま利用することが困難である。

3. 提案手法

従来の手法の問題点を解決するため，本論文では遺伝的アルゴリズムを用いて，少ない評価データであっても広範囲の興味領域を探索し，ユーザの興味を適切に抽出する手法を提案する。

本手法では，以下に示す処理を実行することで，ユーザの興味を表すユーザプロファイルを作成する。

（1） 対象ドキュメントのベクトル表記
（2） ユーザによるドキュメントの評価
（3） 遺伝的アルゴリズムによるユーザプロファイルの作成

以下では，各段階のこれらの処理について，詳細な説明を行う。

〈3・1〉 対象ドキュメントのベクトル表記 検索に利用されるすべてのドキュメントは，ベクトルで表現することとし，ベクトルの各要素の値はtf*idf[7][8]を用いて計算を行う。tf*idfはドキュメントにおけるキーワードの出現頻度と，そのキーワードが出現するドキュメントの総数をもとに，各キーワードの重要度を求める手法として，よく用いられるものである。本手法で用いる重みの計算式としては，参考文献(7)で提案されているものを用いる。

$$w_j^i = tf_j^i \cdot \log \frac{N}{df_j} \quad \cdots\cdots (1)$$

w_j^i：ドキュメント D_iでのキーワード T_jの重み
tf_j^i：ドキュメント D_iでのキーワード T_jの出現頻度
df_j：キーワード T_jを含むドキュメント数
N：全ドキュメント数

式(1)は，特定のドキュメント内でのみ数多く出現するキーワードに対して大きな重みとなる式である。この計算式によりドキュメントの内容を的確に示すキーワードに対して，大きな重みを割り当てることが可能である。

ドキュメントを表すベクトルの次元は，全ドキュメントに含まれるキーワード数となるので，個々のドキュメントベクトルには，そのドキュメントに含まれれないキーワードに対応する要素も含まれる。この要素に対応する値は，式(1)のtf_j^iが0となるので，w_j^iが0となる。また，すべてのドキュメントに含まれるキーワードの場合，$\log(N/df_j)$が0となるので，同様にw_j^iが0となる。

〈3・2〉 ユーザによるドキュメントの評価 ドキュメントは，ユーザがそれを閲覧し，そのドキュメントに興味があるかないかの2段階で評価する。この評価をもとにユーザの興味の抽出を行い，ユーザプロファイルの作成を行う。つまり，興味があると答えたドキュメントのベクトルに対して類似度が高く，また，興味がないと答えたドキュメントに対して，類似度が低くなるようなユーザプロファイルを作成する。本手法では，類似度として，式(2)に示すようなドキュメントベクトルとユーザプロファイルの内積を利用する。

$$Sim_i = prof \quad D_i \quad \cdots\cdots (2)$$

ここで，Sim_iはドキュメントiとユーザプロファイルの類似度，$prof$はユーザプロファイルベクトル，D_iはドキュメントiのドキュメントベクトル，・は内積を表す。

この類似度は，遺伝的アルゴリズムの適合度関数や，ユーザプロファイルによるドキュメントの選択における評価値として利用する。

〈3・3〉 遺伝的アルゴリズムによるユーザプロファイルの作成 本手法では，ユーザの興味を表すユーザプロファイルを遺伝的アルゴリズムにより作成する。この遺伝的アルゴリズムには，バイナリコードで遺伝子を表現するSimpleGA[9]を利用する。

探索する遺伝子は，ユーザプロファイルそのものとする。したがって，ユーザプロファイルはバイナリコードとなり，ユーザが興味のあるキーワードに対しては1，興味のないキーワードに対しては0を設定する。

遺伝的アルゴリズムにおける遺伝的操作としては，選択は重み付きルーレット方式，交叉は一点交叉，突然変異はランダムに遺伝子の値を1から0，0から1に変更するものを利用する。各遺伝的操作を指定した世代まで繰り返すことによってユーザプロファイルを作成する。

4. 実験と考察

本論文では，二種類の実験を行う。まず，本論文で提案する遺伝的アルゴリズムを用いたユーザの興味抽出手法によりユーザプロファイルを作成する。つぎに，その性能を評価する実験を行う。以下では，その実験結果を示し，その考察を行う。

〈4・1〉 実験環境 本手法の性能の評価実験を行うために，検索対象となるドキュメントとして，電子図書館に関する92のドキュメント[10]を用いる。

各ドキュメントに対して形態素解析を行い，キーワードの抽出を行う。本実験で用いた形態素解析ツールは，茶筅version 2.0[11]である。茶筅が各キーワードに割り当てた品詞のうち，「名詞-一般」，「名詞-固有名詞」，「名詞-サ変

接続」,「名詞-非自立-一般」の品詞を持つキーワードのみを，ドキュメントベクトルおよびユーザプロファイルに利用するキーワードとする。これは，形容詞や副詞などの通常検索ではあまり利用されないキーワードを省き，ドキュメントベクトルやユーザプロファイルの大きさを小さくするためである。全ドキュメントから得られたキーワード数は，6,186個であった。

各ドキュメントのドキュメントベクトルは，6,186次元のベクトルで表されており，各要素にはtf*idfにより求められた重みが設定されている。このドキュメントベクトルは要素に0が多く含まれるベクトルであるため，ユーザプロファイルを作成する際に，そのまま利用することは好ましくない。なぜなら，ドキュメントベクトルの要素が0となっている項目は，適合度の算出に反映を及ぼさないため，遺伝子の選択にも影響を及ぼさない。また，その要素に対応するキーワードは，評価したドキュメントに含まれていないキーワードであるため，ユーザの評価が行われているとは言えない。それゆえ，このような要素の値を本手法で決定すると，ユーザの興味を正確に表さず，推薦精度を低下させる原因となってしまう。したがって，ユーザプロファイルの推定を行う際には，前処理として評価した全ドキュメントのドキュメントベクトルが0でない要素だけを取り出して評価データとした。ユーザプロファイルの推定においても同じ制限を設け，前処理で切り捨てられたキーワードに対する重みには0を割り当てることとしている。遺伝的アルゴリズムの各種パラメータは，表1に示すものとする。

本論文では2種類の実験を行い，遺伝的アルゴリズムにより作成されたユーザプロファイルの性能を検討する。1つ目の実験(実験1)は，遺伝的アルゴリズムにより求められたユーザプロファイルが，ユーザの興味に応じてドキュメントを分離しているか否かについて評価する。2つ目の実験(実験2)は，得られたユーザプロファイルで，実際にドキュメントの評価を行い，ユーザの興味にあったドキュメントを選別できるか否かについて評価する。

〈4・2〉 実験1 本実験では，作成されたユーザプロファイルがユーザの興味に応じて，ドキュメントを分離できるか否かを評価する。特に，ユーザが評価したドキュメントより作成されたユーザプロファイルが，推定に利用したドキュメントを，ユーザの評価に応じて分離できているか否かについて考察する。

ユーザプロファイルの探索を行うために，4篇のドキュメントを利用する。このドキュメントには，ユーザが興味を持つドキュメント1篇と興味がないドキュメント3篇が含まれている。このドキュメントからユーザプロファイルを作成し，作成されたユーザプロファイルが4篇のドキュメントを，ユーザの評価に応じて類似度の点で分離できるか否か検討する。4篇のドキュメント情報を表2に示す

表1 遺伝的アルゴリズムのパラメータ
Table 1. Parameters in the simulation.

Population	200
Generation	5000
Crossover	0.5
Mutation	0.0001

表2 評価ドキュメント
Table 2. Document-feature lists.

document	number of keywords	user's interest
document 1	341	×
document 2	311	×
document 3	307	×
document 4	302	○
total	858	

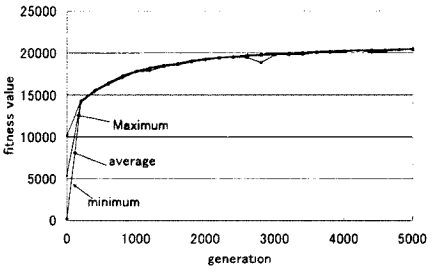


図1 適合度の変化
Fig. 1. Fitness value.

各ドキュメントは，全ドキュメントから得られたキーワード数である6,186個の要素を持つベクトルで表されている。したがって，ユーザプロファイルは6,186次元のベクトルで表されることとなる。しかし，4篇のドキュメントでは，上記のキーワードのうち858個のキーワードしか利用されていない。このため，前節で説明したように，遺伝子は858次元のベクトルとし，遺伝的アルゴリズムによる推定の後，残りの要素を0とすることでユーザプロファイルを作成する。

新しく生成された遺伝子の選択を行うため，遺伝的アルゴリズムで用いる適合度関数を式(3)に示す。

$$Fit = 3 * Sim_4 - 1 * (Sim_1 + Sim_2 + Sim_3) \quad (3)$$

Fitは遺伝子に対する適合度を示し，Sim_iはドキュメントiとユーザプロファイルの類似度を表す。類似度は式(2)を用いて計算される。

適合度関数に含まれる係数は，興味のあるドキュメント数と興味ないのドキュメント数の比を利用している。

図1に世代に伴う適合度の推移，図2〜図5に各ドキュメントの類似度を示す。

図1から分かるように，世代が進むごとに最大値，最小値ともに適合度が増加している。これは，遺伝的アルゴリズムによりユーザの興味に応じたユーザプロファイルが作成され，高い類似度が求められていることを示している。

各ドキュメントとユーザプロファイルの類似度の変化を

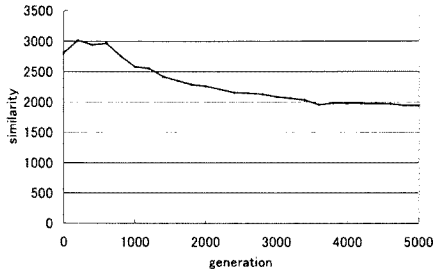


図 2　ドキュメント 1 の類似度
Fig. 2.　Similarity measure for document 1.

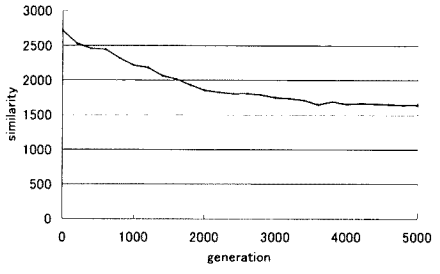


図 3　ドキュメント 2 の類似度
Fig. 3.　Similarity measure for document 2.

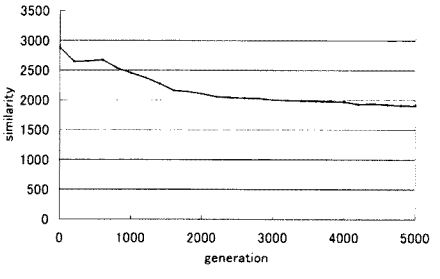


図 4　ドキュメント 3 の類似度
Fig. 4.　Similarity measure for document 3.

見ると，世代が進むにつれて，興味のあるドキュメントの類似度は増加し，興味のないドキュメントの類似度は減少している。したがって，遺伝的アルゴリズムにより作成されたユーザプロファイルは，ユーザの興味を適切に抽出し，ドキュメントを興味に応じて分離していると考えられる。

〈4・3〉 実 験 2　本実験では，遺伝的アルゴリズムにより作成されたユーザプロファイルの推薦精度について評価を行う。

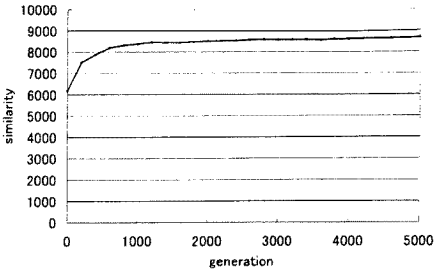


図 5　ドキュメント 4 の類似度
Fig. 5.　Similarity measure for document 4.

本実験では，実験 1 で用いた 4 篇のドキュメントに新たに 8 篇のドキュメントを加えた合計 12 篇のドキュメントから，ユーザプロファイルを作成する。さらに，実験 1 で作成したユーザプロファイルと，本実験で作成したユーザプロファイルの推薦精度を比較する。また，関連フィードバックを用いて作成したユーザプロファイルによる推薦も同様の条件で行う。

92 篇のドキュメントはあらかじめ興味があるか否かの評価を行っており，表 3に示すような評価結果となっている。

追加したドキュメントに含まれるキーワード数とユーザの興味について，表 4に示す。実験 1 で用いたドキュメントと新たに追加したドキュメントに含まれるキーワードは 1,935 個であり，これは表 2と表 4のキーワードの総和から，重複分を省いたものである。このキーワードの個数を遺伝子の長さとする。

遺伝的アルゴリズムで用いる適合度関数は式 (4) とする。適合度関数の係数については，実験 1 と同じ規則により決定している。

$$Fit = 2 * \sum_{i \in ID} Sim_i - 1 * \sum_{j \in ND} Sim_j \quad \cdots\cdots\cdots (4)$$

ここで，IDは興味のあるドキュメントのドキュメント番号の集合，NDは興味のないドキュメントのドキュメント番号の集合を表している。

それ以外の条件については，実験 1 と同様に遺伝的アルゴリズムを実行し，ユーザプロファイルを作成する。

比較として，ユーザの検索要求を抽出する手法である関連フィードバックを用いてユーザプロファイルの作成を行う。関連フィードバックの計算方法としては，式 (5) を用いる。

表 3　ドキュメントの評価
Table 3.　User Preference.

interested documents	25
uninterested documents	67

表4 評価ドキュメント
Table 4. Evaluated document.

document	number of keywords	user's interest
document 5	201	×
document 6	230	×
document 7	310	×
document 8	676	×
document 9	152	○
document 10	280	×
document 11	431	○
document 12	316	○
total	1,581	

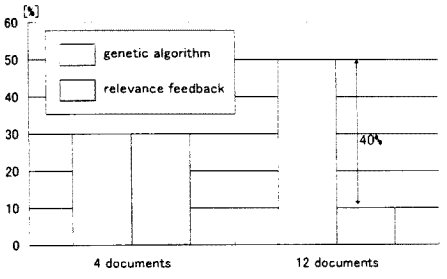


図6 適合率の比較
Fig. 6. Comparison of the precision.

$$prof = 0.7 * \sum_{D_i \in R} D_i - 0.3 * \sum_{D_j \in N} D_j \quad \cdots\cdots (5)$$

ここで，Rはユーザが興味があるとしたドキュメントの集合，Nはユーザが興味がないとしたドキュメントの集合を表している。

2つの手法より得られたユーザプロファイルを用いて，検索対象とした全ドキュメントとの類似度を計算し，ユーザプロファイルの作成に用いたドキュメントを除いた上位10篇のドキュメントを選び出した。その10篇のドキュメントのうち，ユーザの興味と一致するドキュメントが含まれる割合を適合率とし，その適合率の観点から性能の比較を行った。以上の実験より得られた結果を図6に示す。

図6では，評価ドキュメントが4篇の場合には，関連フィードバックと，本手法の結果にほとんど差が見られない。しかし，12篇の場合，提案手法が関連フィードバックと比較して約40%の適合率の向上が見られる。これは評価ドキュメントが増加することによって，ユーザの興味が絞り込まれず，興味領域が広がったためと思われる。このため，関連フィードバック法ではユーザの興味を絞り込めず，中間的なユーザプロファイルを作成したため，適合率が減少したと考えられる。遺伝的アルゴリズムを用いたユーザプロファイルの作成手法では，評価ドキュメントの増加に伴い，適合率が増加している。これは，遺伝的アルゴリズムが持っている広範囲の探索能力が有効に作用したため，ユーザの興味を適切に抽出できたことを示している。

表5 収束世代数
Table 5. Generation of the convergence.

Evaluted document	Generagion
4 documents	2532
13 documents	3472

〈4·4〉 考　察　本論文では，遺伝的アルゴリズムを用いたユーザプロファイル作成手法の提案を行った。実験1により，本手法でユーザの評価に対応したユーザプロファイルを作成することができ，ドキュメントを評価することが可能であることが確認できた。

また，実験2より，広い興味領域を対象とする場合には，遺伝的アルゴリズムによる広範囲で高速な探索と言う特徴を利用して，適切な興味抽出が行えることを確認できた。したがって，ユーザの興味の範囲が広い一般的な検索の利用状況においては，本手法が有効であると考えられる。

以上の2つの実験により，プロファイルを作成する際の評価ドキュメントの個数については，評価ドキュメントが多い方が精度の良いプロファイルを作成できることが確認できた。しかし，実際の使用状況を考えると，あらかじめユーザが大量のドキュメントを評価することは大きな負担となる。このため，従来手法と同程度の性能を出る4篇以上の評価ドキュメント数が好ましいと思われる。

各評価実験において，5,000世代における適合度の95%の値に達した世代数を表5に示す。この結果より，本手法を用いてユーザプロファイルを作成する際に必要となる世代数は，3,000世代であると思われる。これは評価データを4篇，13篇にした場合に，ともに3,000世代前後でドキュメントの分離が行われているためである。評価ドキュメントが増加するにつれて，収束する世代数は増加するが，実際の使用においてユーザが評価できるドキュメント数の制限を考えると，上記の世代数は妥当である。

本論文では，0,1の2値の値によってユーザプロファイルを表しているため，本プロファイルの動作はドキュメントからのキーワード選択として動作する。このため，従来のキーワード選択手法であるtf*idfと関連フィードバックより選択されたキーワード数について検討を行う。表6は，4篇のドキュメントを評価ドキュメントにおいて，各手法から選択されたキーワード数である。tf*idfではユーザが興味があると評価したドキュメントに含まれるキーワード数，関連フィードバックでは式(5)より求められたものから，正の値を持つキーワード数とした。関連フィードバックでは，実数値ベクトルとして表されるため，正の値を持

表6 抽出キーワード数
Table 6. The number of keywords.

Method	Keywords
tf*idf	302
GA	297
relevance feedback	254

つキーワードは興味を表すキーワードであると判断できるためである。表6より，本手法は関連フィードバックより絞り込まれていないようにみえる。しかし，数少ないデータからユーザの興味を狭く推測することはローカルミニマムに陥る可能性がある。本手法では，遺伝的アルゴリズムの探索と収束のバランスによりこの問題を回避していると考えられる。

5. おわりに

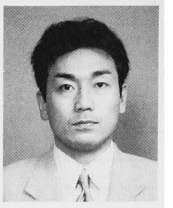
本論文では，遺伝的アルゴリズムを用いたユーザの興味抽出手法を提案し，実際のドキュメントを用いた評価実験を行い，その有効性を確認した。実験結果より，本手法はユーザの興味を抽出したユーザプロファイルを作成することが可能であることが分かった。また，ユーザプロファイルによって適切にドキュメントの選別できることも分かった。

今後は，実数値遺伝的アルゴリズムを用いたユーザプロファイル作成手法を開発することにより，単語間の重みを考慮したユーザプロファイルを利用することで推薦精度向上を目指す。また，多人数による評価実験を行うことで，異なった個人の興味抽出の精度の検討を行う予定である。

（平成12年11月27日受付，同13年3月7日再受付）

文　献

（1） 森田　昌宏，速水　治夫：“情報フィルタリングシステム-情報洪水への処方箋-” 情報処理，Vol.37，No.8，pp.751-758(1996)
（2） Salton, G. and Buckley, C.: “Improving Retrieval Performance by Relevance Feedback”, Readings in Information Retrieval, pp.355–364 (1997)
（3） Baclace, P.E.: “Personal Information Intake Filtering”, In Bellcore Workshop on High Performance Information Filtering, pp.1-15 (1991)
（4） Baclace, P.E.: “Competitive Agents for Information Filtering”, ACM, Vol.35, No.12, p.50 (1992)
（5） Malon, T.W., Grant, K.R., Lai, K.-Y., Rao, R. and Rosenblitt, D.: “Semistructured Messages are Surprisingly Useful for Computer-Supported Coordination”, ACM Transactions on Office Information Systems, Vol.5, No.2, pp.115-131 (1987)
（6） 砂山 渡，野村 勇治，大澤 幸生，谷内田 正彦：“Webページ検索におけるユーザの興味表現支援システム”，電子情報通信学会論文誌 D-I, Vol. J82-D-I, No.12, pp.1394-1402 (1999)
（7） 長尾真編：自然言語処理（1996）岩波書店
（8） Salton, G. and Buckley, C.: “Term-Weighting Approaches in Automatic Text Retrieval”, Readings in Information Retrieval, pp.323-328 (1997)
（9） David, E. Goldberg: “Genetic Algorithm in Search, Optimization, and Machine Learning”, Addison-Wesley Pub. Co. (1989)
（10） http://www.dl.ulis.ac.jp/
（11） http://cl.aist-nara.ac.jp/lab/nlt/chasen/



柳本豪一（正員） 1972年生。1996年大阪府立大学大学院工学研究科博士前期課程修了。同年日本電気株式会社入社。2000年より大阪府立大学工学部助手となり現在に至る。情報検索，進化型計算手法に関する研究に従事。情報処理学会，計測自動制御学会会員。工学修士。

藤中　透（正員） 1956年生。1984年京都大学大学院工学研究科博士後期課程研究指導認定退学。1986年同大学工学部助手となり，1990年より大阪府立大学工学部講師。最適制御およびインテリジェント制御関係の理論と応用に関する研究に従事。工学博士。システム制御情報学会，計測自動制御学会会員。

吉岡理文（正員） 1968年12月10日生．1996年3月東京大学大学院工学研究科博士課程修了　同年4月大阪府立大学工学部助手，1998年11月同講師となり現在に至る．画像処理等の研究に従事．工学博士，情報処理学会会員，計測自動制御学会会員．

大松　繁（正員） 1974年大阪府立大学大学院博士課程修了。同年徳島大学工学部情報工学科助手。1988年同知能情報工学科教授。1995年大阪府立大学工学部情報工学科教授となり，現在に至る。1991年電気学会論文賞，1995年計測自動制御学会論文賞，1996年市村賞受賞。ニューラルネットワークの研究に従事。IEEE Trans. on Neural Networkの Associate Editor。システム制御学会，計測自動制御学会，IEEE会員。工学博士。