知能と複雑系 108－3
（1997．5．26）

Ｗｅｂページからのユーザの興味の遺伝的アルゴリズムに基づく抽出

渥美雅保
創価大学工学部情報システム学科
matsumi@t.soka.ac.jp

概要

近年の World Wide Web(WWW) の指数的な成長に対して，Web 上の情報にインデクスを付け情報収集を支援するサービスが多く提供されているが，ユーザの興味にあった情報を効率的に見つけるという点からはほど遠い状態にある．ユーザの興味を情報収集に反映させるためには，ユーザの興味を表すユーザプロファイルを情報収集システムに蓄積し，それらを情報収集において活用する個人適応化機能を情報収集システムに持たせることが有用な方法と考えられる．本論文では，情報検索，情報フィルタリング，ブラウジングといったユーザの情報収集を個人適応化するために，ユーザの情報収集過程からプロファイルを遺伝的アルゴリズムに基づき抽出する方法を提案する．また，情報収集結果の評価尺度として，ランク正確度，ランク再現度という尺度を導入する．そして，WWW からの情報検索を対象とした実験により，本手法に基づくプロファイル抽出の有用性と限界を示す．

Extraction of User's Interests from Web Pages based on Genetic Algorithm

Masayasu Atsumi
Dept. of Information Systems Science, Faculty of Eng., Soka Univ.

Abstract

For the current exponential growth of World Wide Web (WWW), there have been a lot of information gathering tools to help people cope with the volume of information on the WWW. However, these tools are insufficient to find information attracting personal interest of a user with high precision and recall. In order to make information gathering reflect interests of a user, it is useful to make information gathering tools accumulate interests of each user and apply them at information gathering. This paper proposes a method based on genetic algorithm that extracts a user's interest from information gathering processes such as information retrieval, information filtering and browsing. We also introduce the rank precision and rank recall that measure effectiveness of information gathering. We present experimental results of interest extraction and extracted interest application in WWW information retrieval and discuss the effectiveness of our approach.

1 はじめに

World Wide Web(WWW)上の情報空間からユーザの興味にあった文書を収集するために，多くの情報収集ツールが提供されている．しかし，情報空間の規模の大きさと，それに加えて，ユーザが自らの興味を適切に表現することの困難さとから，情報収集の正確度と再現度は低く，真にユーザの欲する情報が効率的に収集できる状況にはなっていない．ユーザの興味を情報収集に反映させるためには，ユーザの情報収集過程からユーザの興味を表すプロファイルを抽出し，それらを情報収集で活用することが有用な方法と考えられる [2].

本稿では，情報検索，情報フィルタリング，ブラウジングといったユーザの情報収集を個人適応化するために，ユーザの情報収集過程からプロファイルを遺伝的アルゴリズム [5] に基づき抽出する方法を提案する．また，情報収集結果の評価尺度として，正確度と再現度 [3] を拡張したランク正確度とランク再現度という尺度を導入する．

以下，2章では遺伝的アルゴリズムに基づくプロファイル抽出方法，3章ではプロファイルの評価方法，4章では WWW からの情報検索を対象とした実験結果を示し本手法の有用性と限界を論ずる．

2 興味の抽出

2.1 興味のモデル

興味は，ある意図のもとでの内的な記憶と外的な刺激の両者により喚起される．文書収集においては，ユーザの意図により想起される索引語とそれを用いて収集された文書の内容の両方が興味を喚起する源となる．特に，後者が興味の具現化と潜在的な興味の活性化に役立つ．従って，ユーザの興味は，ユーザの文書閲読過程から抽出しなくてはならない．

情報収集の機械処理においては，文書内容を語の集合により近似的に表現できると仮定されている．この考えに従えば，人がある文書をみて喚起される興味はその文書中の語の集合により近似的に表現でき，逆に興味を減退させる要因もその文書中の語の集合により表現できる．これより，いくつかの文書を閲読して，興味を喚起させた語に正の重みを付け，興味を減退させた語に負の重みを付けた重み付の語集合は，ユーザの興味を近似するとみなしうる．

Web文書を，ベクトル空間モデル [7] に従って，その文書に含まれる内容語からなるベクトルとして，

$$D_i = (f_{ij}) \tag{1}$$

と表す．ここで，f_{ij}は文書 D_iに含まれる内容語 t_j の重みを表す．重みはその語がその文書を特徴づける程度を表し，本稿では簡単に，文書中の語頻度を重みとして用いる．これより，Web 文書集合は，それらに含まれる異なる内容語の数を次元とする多次元文書空間を形成する．

ユーザの興味は，この文書空間において，単語に正または負の重みを付けたベクトルとして，

$$P = (c_j) \tag{2}$$

と表すことができる．ここで，c_jは興味 Pを表す単語 t_jの重みで，1と-1の間の実数をとる．ユーザの興味の表現 Pはプロファイルと呼ばれる．プロファイル Pには，興味を喚起した文書を特徴づける単語が正の重み付きで，興味を減退させた文書を特徴づける単語が負の重み付きで含まれる．このとき，プロファイル Pと文書 D_iとの間の類似性尺度

$$sim(P, D_i) = \frac{P \cdot D_i}{|P||D_i|} \tag{3}$$

は，Pに関する D_iの関連度を与える．

2.2 GA に基づくプロファイル抽出

ユーザプロファイルを Web 文書空間から抽出する問題を次のように定式化する．文書集合を $\{D_i\}$，ユーザが文書 D_iに対して判断した自らの興味との関連度を $r_i (0 \leq r_i \leq 1)$ とする．このとき，

$$gap(P, \{D_i\}) = \sum_i |r_i - S_{P,D_i}| \tag{4}$$

$$\text{where} \quad S_{P,D_i} = \begin{cases} sim(P, D_i) & \text{for } sim(P, D_i) \geq 0 \\ 0 & \text{for } sim(P, D_i) < 0 \end{cases}$$

を最小とするプロファイル Pを $\{D_i\}$ が形成する文書空間から求めよ．我々は，この問題をプロファイルを可変長の染色体として表現し，$gap(P, \{D_i\})$ をフィットネス関数として，フィットネス関数を最小化するプロファイル染色体を文書空間の遺伝的探索により求めることにより解く．

プロファイルを染色体として表現し，プロファイルを構成する単語とその重みの両方を同時に進化させるために，型を持った遺伝子を導入して，単語を型，重みを型の値として表現する．プロファイルは，この遺伝子を各遺伝子座に持ち，かつ各遺伝子座の型が染色体において唯一に保たれるという制約を満たす可変長の染色体として表現される．

Web文書空間の遺伝的探索は，プロファイル染色体の再生，交叉，突然変異によりなされる．交叉方法として，2点交叉を拡張した部分系列交換交叉と

呼ぶ方法を導入する．部分系列交換交叉は次のようになされる．交叉の親となる2つのプロファイル染色体の長さを l_1, l_2とする．$l_1 \neq l_2$のとき，各染色体について，先頭と末尾を含む遺伝子間の (l_1+1)点，(l_2+1)点からそれぞれ2点がランダムに選ばれる．$l_1 = l_2 = l$のとき，先頭と末尾を含む遺伝子間の $(l+1)$点からそれぞれ2点が，2点間の長さが同一という制約のもとでランダムに選ばれる．そして，2点間の遺伝子列を交換してできる染色体において各遺伝子座の型の唯一性が維持される場合に，2点間の遺伝子列が交換されて2つの子プロファイル染色体が生成される．また，突然変異としては，型の突然変異と値の突然変異を用いる．型の突然変異は，ランダムに選ばれた遺伝子座の遺伝子の型と値をランダムに選ばれた別の型と値に，染色体の各遺伝子座の型の唯一性を維持して置換する．これより，プロファイル染色体を構成する単語がその重みと共に置き換えられる．一方，値の突然変異は，ランダムに選ばれた遺伝子座の値をランダムに選ばれた別の値に置換する．これより，プロファイル染色体を構成するある単語の重みが置き換えられる．

2.3 情報収集に GA を用いた関連研究

Yangら [9] は，情報検索において，検索質問の各項の重みをユーザの関心にあった検索ができるように最適化する検索質問改良に遺伝的アルゴリズムを用いている．**Arita**ら [1] は，文献からキーワードを抽出するために遺伝的アルゴリズムを用いている．**Sheth**[8] は，情報フィルタリングにおいて，ユーザのプロファイルを関連フィードバック法により修正・評価した結果から，新しいプロファイルを生成するために遺伝的アルゴリズムを用いている．**Ferguson**[4] は，文書集合からのプロファイル抽出において，ある文書に含まれる単語の同義語集合の固定長列の染色体によりプロファイルを表現し，正確度と再現度を最大にするプロファイルの探索に遺伝的アルゴリズムを用いている．本研究では，文書集合から式 (4) を最小にする任意長の重み付き単語列をプロファイルとして遺伝的アルゴリズムにより抽出する．

3 プロファイルの評価

3.1 ランク正確度とランク再現度

情報収集結果の評価尺度としては，一般に，正確度と再現度が用いられる．これらは，収集された文書集合をユーザの関心の有無の2値で評価する場合には有用であるが，Web検索エンジンの検索結果のランク付き出力のように，出力順序とユーザの関連度判断の順序の一致度を評価するのには十分でない．ある文書集合 $\{D_i\}$ に対して，D_iに対するユーザの関連度判断 r_iの順位と，プロファイル Pに関する D_iの関連度 $sim(P, D_i)$ の順位の一致の程度は，**Kendall** の順位相関係数 [6] により測定できる．本稿では，この一致の程度をより詳細に評価するための尺度として，正確度と再現度をそれぞれ拡張したランク正確度とランク再現度という尺度を導入する．

ランク正確度は，計算された関連度順に提示される n 個の文書のうち，ユーザの興味からみて m 位までの文書が占める割合で，文書集合のサイズを N とするとき，提示文書数 $n(1 \leq n \leq N)$ とユーザの閲読要求順位 $m(1 \leq m \leq N)$ に対して，

$$RP_{n,m} = \frac{n \text{ 文書中 } m \text{ 位以内の文書数}}{n} \quad (5)$$

により与えられる．これは，ユーザが自らの興味に関して m位までの文書を見たいという閲読要求を持って n個の関連度順に提示される文書を見る場合の正確度を与える．ランク正確度は，ユーザの閲読要求順位の増加に関して単調増加する性質を持つ．任意の提示文書数 n に対して，ランク正確度が$\alpha(0 \leq \alpha \leq 1)$となる閲読要求順位 mを求めれば，n文書を見ることにより順位正確度αで閲読可能な文書の順位を知ることができる．

ランク再現度は，ユーザの興味からみて m 位までの文書のうち，計算された関連度順に提示される n 個の文書中に含まれるものの割合で，文書集合のサイズを Nとするとき，ユーザの閲読要求順位 $m(1 \leq m \leq N)$ と提示文書数 $n(1 \leq n \leq N)$ に対して，

$$RR_{n,m} = \frac{n \text{ 文書中 } m \text{ 位以内の文書数}}{m} \quad (6)$$

により与えられる．これは，ユーザが自らの興味に関して m 位までの文書を見たいという閲読要求を持って n個の関連度順に提示される文書を見る場合の再現度を与える．ランク再現度は，提示文書数の増加に関して単調増加する性質を持つ．任意の閲読要求順位 m に対して，ランク再現度が$\beta(0 \leq \beta \leq 1)$となる提示文書数 n を求めれば，閲読要求順位 mまでの文書をランク再現度βで見るために必要な文書数を知ることができる．

ランク正確度，及びランク再現度において，$n = m$の場合の尺度を特に，ランク一致度と呼ぶ．ランク一致度は，文書集合のサイズを Nとするとき，任意の $n(1 \leq n \leq N)$ に対して，

$$RA_n = \frac{n \text{ 文書中 } n \text{ 位以内の文書数}}{n} \quad (7)$$

により与えられる．ランク一致度は，計算された関

連度順に提示される n 個の文書がユーザの興味からみて n 位までの文書である割合を与える．

3.2 プロファイルの適応

プロファイルは蓄積されてそれ以降の情報収集に用いられる．文書集合 $\{D_i\}$ から抽出されたプロファイル P がユーザの興味を反映している程度は，

$$A(P) = \frac{N - gap(P, \{D_i\})}{N}, \quad 0 \leq A(P) \leq 1 \quad (8)$$

と表すことができる．これを，P の初期適合度という．

プロファイルの情報収集における利用では，いくつかのプロファイルを適合度に応じて適切に組み合わせるとともに，それらの適合度をフィードバックにより更新し，プロファイルをユーザの興味に適応させることが必要である．プロファイルの組合せは，次のいずれかの方法によりなされる．第一の方法は，各プロファイルに関して各文書との関連度を計算し，各文書について計算された最大の関連度をその文書の関連度とする．この方法を最大結合法と呼ぶ．第二の方法は，プロファイルを次のように結合する．組合わせるプロファイル集合を $\{P_k\}$，その数を K とする．また，プロファイル P_k に含まれる単語 t_j の重みを c_{kj} とする．このとき，いずれかのプロファイルに含まれる単語 t_j とその重み $\sum_{k=1}^{K} c_{kj}$ からなるベクトルを結合プロファイルとする．そして，この結合プロファイルを用いて各文書の関連度を求める．この方法を加法的結合法と呼ぶ．プロファイル P の適合度は，フィードバックを F_d とするとき，

$$A(P) = \begin{cases} 1 & \text{for} A(P) + \gamma \times F_d > 1 \\ A(P) + \gamma \times F_d & \\ & \text{for} 0 \leq A(P) + \gamma \times F_d \leq 1 \\ 0 & \text{for} A(P) + \gamma \times F_d < 0 \end{cases} \quad (9)$$

により更新される．ここで，γ はフィードバックへの適合度の感度を表す．フィードバック F_d の選択には，Kendall の順位相関の有意性検定の結果を用いることができる．これより，プロファイル P が文書集合と有意な正相関を示した場合には F_d を正の値とすることにより P の適合度を上げ，そうでない場合には F_d を負の値とすることにより P の適合度を下げる．

4 実験結果

4.1 プロファイルの抽出実験結果

Web 文書集合からのプロファイル抽出は，文書を特徴づける文書ベクトルの抽出と，それらが定める文書空間からのプロファイル抽出によりなされる．プロファイルとしては，ユーザによりある一定の興味のもとで $0, 0.3, 0.7, 1$ の 4 段階の関連度評価を与えられた N 個の文書集合 $\{D_i\}$ から，フィットネス関数

$$F(P) = N - gap(P, \{D_i\}) \quad (10)$$

を全人口全世代を通して最大，即ち $gap(P, \{D_i\})$ を最小としたものを抽出する．

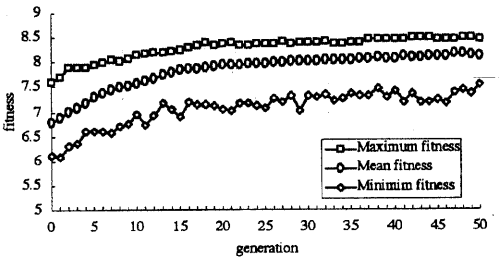
まず，プロファイル抽出への遺伝的アルゴリズムの適用可能性に関する実験結果を示す．プロファイルの抽出用文書集合としては，表 1の各検索質問に対して Alta Vista を用いて収集した 10 個の文書からなる文書集合を用いた．プロファイルの遺伝子表

No.	*Query*	*No.*	*Query*
1	agent	4	programming
2	agent+learning	5	genetic+programming
3	agent+network	6	java+programming

表 1 Queries for retrieval of web documents

現パラメータは次のとおりである．プロファイル染色体の初期長は 1 から 16 までのランダムな値をとる．プロファイルを構成する語の重みは，1 と −1 の間の 0.1 刻みの離散値をとり，重みが負の値をとる率は 0.25 とする．また，遺伝的操作パラメータは，人口数 100，型の突然変異率 0.01，値の突然変異率 0.025，交叉率 0.6，交叉規則は部分系列交換交叉とした．

図 1にプロファイル染色体のフィットネス値の世代推移を示す．表 2に，kendall の順位相関係数の有意



Fitness curve of profiles, averaged over 3 runs,
for documents retrieved by the query "agent+network".

図 1 Fitness curve

性検定の結果を示す．表 2では，遺伝的アルゴリズムによるプロファイル抽出方法を，ユーザによる関

Query No.		1	2	3	4	5	6
Profile by GA	training set	pos[.05]	pos[.05]	pos[.025]	pos[.025]	pos[.1]	pos[.025]
	non-training set 1	pos[.075]	pos[.075]	ind[+]	ind[+]	ind[+]	pos[.025]
	non-training set 2	pos[.05]	pos[.1]	pos[.05]	ind[+]	ind[-]	pos[.075]
Profile by non-GA	training set	ind[+]	pos[.025]	pos[.025]	ind[+]	ind[+]	ind[+]
	non-training set 1	ind[+]	pos[.075]	ind[-]	ind[+]	ind[-]	pos[.025]
	non-training set 2	ind[+]	ind[+]	pos[.025]	ind[+]	ind[-]	ind[+]

Profile by GA **expresses a profile extracted based on our genetic algorithm, on the other hand,** ***profile by non-GA*** **expresses a profile extracted from a document whose relevance to a user's interest is 1.0.** ***Training set*** **is a set of documents used to extract profiles, and** ***non-training set*** **is a set of documents that are not used for profile extraction. The item *pos*[p] means there exists a significant positive correlation by one-sided test of probability *p*. The item *ind* [+] means there exists a positive correlation but it is judged independent by statistical test. The item *ind* [-] means there exists a negative correlation.**

表 2 Significance test of ranking by profiles

連度評価が 1.0 の文書から抽出した内容語からプロファイルを構成する方法と比較している．訓練用文書集合とはプロファイル抽出に用いた文書集合，非訓練用文書集合とは同一の興味のもので収集された10個の文書からなる別の文書集合をいう．遺伝的アルゴリズムにより抽出されたプロファイルの関連度順位付けが，訓練用文書集合に関して非常によいことがわかる．これは，プロファイルがユーザの興味を忠実に捉えていることを示している．一方，非訓練用文書集合に関しては，関連度順位付けがうまくいかない場合がある．これは，プロファイルが，訓練用文書を特徴づける特殊な語を含むことに起因する．

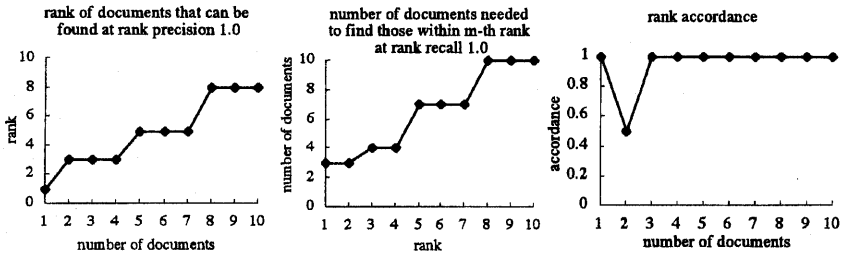
図 2に，ユーザがある個数の文書を順々に見る場合にランク正確度 1.0 で閲読可能な文書の順位，ある順位までの文書をランク再現度 1.0 で見つけるために見ることが必要な文書数，及びランク一致度を示す．図 2より，例えば，2個の文書を見ることにより3位までの文書を見つけることができることが，3位までの文書を見つけるためには4個の文書を見る必要があることがわかる．このように，これら尺度により，プロファイルの関連度順位付け性能をより具体的に確かめることができる．

次に，プロファイル抽出性能に種々のパラメータが及ぼす影響に関する実験結果を示す．第一に，文書空間の次元数は遺伝的アルゴリズムによるプロファイル抽出性能に大きな影響を及ぼした．各文書から頻度が高い内容語のみを抽出し，文書空間の次元数を低く抑えることが効率的な抽出には有効である．第二に，抽出されたプロファイル染色体長は，その抽出に用いられた文書集合に依存して様々に異なった．我々の可変長プロファイル染色体が適切なプロファイル長を求めるのに効果を発揮した．第三に，遺伝的操作のパラメータに関して，突然変異率をあげるとランダム探索に近づきプロファイルの抽出がうまくできないこと，交叉率は実験で用いた 0.6 の前後 0.3 から 0.9 の範囲で大きな差は認められないこと，部分系列交換交叉が一点交叉や2点交叉と比べてやや高いフィットネスを達成することが確かめられた．

4.2 プロファイルの適応実験結果

プロファイルの組合せ的適用とフィードバックによる適応実験の結果を示す．プロファイル集合としては，表 1の各検索質問に対して収集した 5個の文書からなる文書集合からそれぞれ抽出したプロファイルの集合を用いた．プロファイルの遺伝子表現と遺伝的操作のパラメータは，人口数が 50 であることを除いて 4.1で用いたものと同じものを用いた．

表 3に，2つのプロファイルを結合してサイズ 10 の新たな文書集合に適用した結果を示す．これら文書集合も同じ質問を用いて収集されたものである．プロファイルを適切に組み合わせることにより順位付け性能が向上していることがわかる．プロファイルの適応では，適合度の高い2つのプロファイルを選んで組合わせて適用し，プロファイルが有意な正相関を示したか否かにより，(9)に従ってそれらの適合度を更新する．実験の結果，適応を繰り返すことにより，ユーザの興味と有意な正相関を示す結合プロファイルが選択されるようになった．これらより，新たな文書集合に対して既存のプロファイルを組合わせて適用し，新たな文書集合から新たにプロファイルを抽出すると共に，既存のプロファイルの適合度を更新していけば，プロファイルをユーザの興味に適応させうることが確かめられた．



Ranking performance of a profile extracted from documents retrieved by the query "agent+network".

図 2 Ranking performance of a profile extracted based on GA

Comparison	Equal to Low	Intermediate	Equal to High	Higher
Maximum combination	17.3%	28.3%	45.7%	8.7%
Additive combination	10.8%	37.0%	23.9%	28.3%
Total	14.1%	32.6%	34.8%	18.5%

This table shows the comparison of ranking performance between combinatorial application of two profiles and separate applications of them. The label ***Equal to Low*** means performance of a combinatorial profile is equal to the lower performance of two profiles in the combination.The label ***Intermediate*** means performance of a combinatorial profile is intermediate between those of the two. The label ***Equal to High*** means performance of a combinatorial profile is equal to the higher performance of the two. The label ***Higher*** means performance of a combinatorial profile is higher than those of the two. The number of combinations is 46.

表 3 Effects of combinatorial profile applications

5 むすび

本稿では，WWWからの情報収集を個人適応化するために，ユーザが関連度評価をしたWeb文書集合から遺伝的アルゴリズムに基づきユーザのプロファイルを抽出する方法を提案した．また，情報収集結果の評価尺度としてランク正確度，ランク再現度を提案した．そして，WWWから収集した文書を用いて，ユーザのプロファイル抽出への本方法の適用可能性と種々のパラメータの抽出性能への影響を実験的に明らかにした．また，プロファイルの情報収集における組合わせ的適用の効果と，適用結果のフィードバックによるプロファイルのユーザの興味への適応可能性について確かめた．

参考文献

[1] Arita, S., Nishimura, K., Shimazu, H.: Keyword Extraction by Keyword-Fitness Optimization. Journal of Japanese Society for Artificial Intelligence, **11**, 551-556 (1995)

[2] Edwards, P., Bayer, D., Green, C.L., Payne, T.R.: Experience with Learning Agents which Manage Internet-Based Information. Machine Learning in Information Access, Papers from the 1996 AAAI Spring Symposium Technical Report SS-96-05, 31-40, AAAIPress (1996)

[3] Ellis, D.: New Horizons in Information Retrieval. Library Association Publishing Ltd.(1990)

[4] Ferguson, S.: BEAGLE:A Genetic Algorithm for Information Filter Profile Creation. University of Alabama, MH585 (1995)

[5] Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Publishing Company, INC. (1989)

[6] Kendall, M.G.: Rank Correlation Methods. Charles Griffin (1970)

[7] Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Communications of the ACM, **18**(11), 613-620(1975)

[8] Sheth, B.D.: A Learning Approach to Personalized Information Filtering. Master's Thesis, Department of Electrical Engineering and Computer Science, MIT (1994)

[9] Yang, J., Korfhage, R.R.: Query Optimization in Information Retrieval Using Genetic Algorithms. Proceedings of the Fifth International Conference on Genetic Algorithms, 603-611(1993)