個人の感性モデルに基づく対話型遺伝的アルゴリズムを用いた推薦システムの提案

宮地 正大
Masahiro MIYAJI

1　はじめに

情報化技術の発展により，Web 上に膨大な量の情報が溢れている．これらの大規模なデータから情報を閲覧するユーザが，求めるコンテンツを的確に得ることは難しい．そのため，一部のオンラインショッピングサイトやニュースサイトなどではユーザ個別に推薦内容を変化させるパーソナライズされた推薦システムが組み込まれている [1, 2]．個人の感性をモデル化することで利用者個別の嗜好に合わせた推薦が可能であると考えられており，様々な研究が行われている [3]．

本研究では，対話型遺伝的アルゴリズムの要素を取り入れることでユーザの感性に近い推薦を行う手法を提案する．対象問題として，専門性の高い用語が頻出するレポート（テキストデータ）群で構成される Web 記事である IS Report および，オンラインショッピングサイト (楽天市場[*1]) の商品情報を用いた．

2　推薦システム

2.1　協調フィルタリング

Amazon[2] や Google[1] などで用いられており，多数のユーザの中から行動履歴の類似したユーザを抽出することで，そのユーザの参照したコンテンツを推薦し合う方式である．他のユーザとの好みの類似性を基本としたアルゴリズムであるため，ユーザが未知の商品が推薦される場合がある．また，推薦・予測にコンテンツ自体の先験情報が必要ないため手軽に導入が可能であり，多分野のコンテンツが入り交じるシステムでの利用が可能である．しかし，システムの条件としてすべてのコンテンツをユーザが評価する必要があるため，ユーザが多数いることが必要となる．そのため，誰も評価していないコンテンツは推薦される可能性が低くなり，推薦されるコンテンツが集中する問題点もある．

2.2　内容ベースフィルタリング

ユーザの行動履歴とコンテンツに含まれるメタ情報（著者・出版社・内容など）を特徴ベクトルとしてマッチングさせる手法である．推薦システムがユーザ評価を必要としないため，全コンテンツを公平に推薦対象とすることができ，小規模なシステムへの導入が可能である．

2.3　推薦システムのパーソナライズ

ユーザ個人の感性をモデリングする手法として前述の内容ベースフィルタリングが有用であると考えられる [4]．ユーザの嗜好に応じてパーソナライズされた推薦を行うためには，ユーザの嗜好を表す感性モデルを何らかの方法で学習・予測する必要がある．代表的な手法としては，確率推論に基づく手法であるベイジアンネットワークや隠れマルコフモデル，ユーザの求める要素（感性）のパラメータを推定・最適化する手法である対話型遺伝的アルゴリズムなどが挙げられる [5, 6]．

[*1] http://www.rakuten.com

3　対話型遺伝的アルゴリズム

3.1　概要

対話型遺伝的アルゴリズム (iGA: interactive Genetic Algorithm) は，多点探索の最適化アルゴリズムである GA をベースとした対話型最適化手法である．人間の感性のモデルを設計変数空間のランドスケープとして捉え，その空間における最良点，もしくは最良域を探索する．iGA を実装したシステムは，ユーザに対して，多数の候補解を提示し，ユーザは感性や好みに基づいてそれらを評価し，その評価値を用いてシステムは遺伝的操作を適用する．これらの操作を繰り返すことで，集団全体をユーザの好むものへと変化させる．iGA は感性による評価を必要とするアプリケーションに利用されている．

3.2　アルゴリズムの設計

iGA では，他の最適化問題と同様に，最適化の対象とする候補解を設計変数として表現する．例えば，服飾デザイン支援システムであれば，デザインする服の形状や色，装飾などが設計変数として定義され，各解は，その設計変数のベクトルによって構成される．最適化を行う遺伝的操作のフェーズでは，さらにこの設計変数を 01 のビット列や遺伝子の型と実数値などの染色体に修正して用いる．

Fig. 1 に遺伝的操作の流れを示す．まず，試行の最初に，染色体を多数含む母集団を初期化する．そして，この染色体一つ一つに対してユーザによる評価を行う．評価値の高い染色体を，親個体として選択し，これらに交叉として情報を組み替える操作を与えることで，より高い評価が期待される子個体を生成する．また，探索途中に局所解に陥ることを防ぐために確率的に突然変異を行う．これらの選択，交叉，突然変異を一連の流れを 1 世代の操作とする．何世代か繰り返すことで，徐々に評価の高い集団へと進化させていく．

4　提案推薦システム

4.1　提案システム概要

本研究では前章で述べた iGA に基づく記事推薦システムを提案する．iGA により，ユーザの感性パラメータを予測することで，ユーザの感性に近い推薦を行うことを

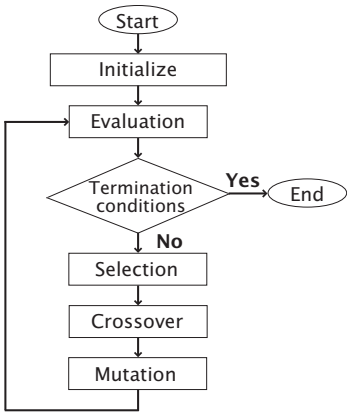

Fig.1　遺伝的操作の流れ

目標としている．提案手法の流れを以下に示す．

1. 推薦対象のコンテンツを特徴ベクトルとなる単語列に分解
2. ユーザの利用履歴から感性パラメータ候補を生成
3. 感性パラメータ候補に類似するコンテンツの提示
4. 手順 2,3 を繰り返す

コンテンツ評価フェーズでは，コンテンツのタイトルのみがユーザに推薦コンテンツとして提示されるため，提示されたすべてのコンテンツの詳細情報を閲覧せずにユーザに評価してもらうことは難しい．そのため，提案手法では提示されたコンテンツタイトルの中からユーザが次に遷移・閲覧したコンテンツを，遺伝的操作における評価とする．コンテンツの特徴ベクトルの定義には先行研究で多く用いられている手法と同様に，コンテンツに付加されている説明文に出現する単語を特徴ベクトルとし，重み付けは TF・IDF 法を用いる [7, 4]．本研究においては，それらの組み合わせ及びパラメータをユーザの感性パラメータの候補とする．しかし出現単語数が膨大な数になることから，解探索に悪影響を及ぼすことが予想される．そのため，膨大な量の設計変数を機械的に扱いやすい形に変形することで，解探索の性能を向上させる研究が行われている．その手法として，設計変数を主成分分析により，別の主成分へと写像することで次元数を削減する手法 [5, 6] や，初期個体の生成時に予め SVM によるユーザ嗜好の学習を行わせることで個体の収束を早める手法 [8] などが考案されている．

本研究では，設計変数間に重みとは異なる関連度を定義することで，別次元同士の設計変数での遺伝的操作を可能とする手法を用いる．それにより，各コンテンツが全種類の設計変数を保持する必要がなくなる．単語間の関係性を文書内での単語の共起確率から作成し，それらの組み合わせ及びパラメータを最適化対象とする．

4.2　特徴単語の重みの決定

コンテンツの特徴ベクトル項目として抽出した単語の重みを決定する必要がある．その手法として，TF・IDF 法を用いる．TF・IDF 法とは，単語の出現頻度 (Term Frequency: TF) 及び逆文書頻度 (Inverse Document Frequency: IDF) を用いた文書内での単語の重要度を表す指標であり，それぞれ式 (1)～ 式 (3) として表される．

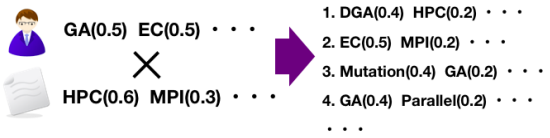

Fig.2　感性パラメータの生成

$$tf(i,j) = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

$$idf(i) = 1 + \log_2 \frac{|D|}{|\{d : d \ni t_i\}|} \tag{2}$$

$$tfidf = tf \times idf \tag{3}$$

$n_{i,j}$ は単語 i の文書 j における出現回数，$|D|$ は総ドキュメント数，$|\{d : d \ni t_i\}|$ は単語 i を含むドキュメント数である．同一ドキュメント中で頻出な単語は TF 値が高くなり，多くのドキュメントで用いられている単語は IDF 値が低くなる．

4.3　単語の共起確確率に基づく関係ネットワーク

TF・IDF 法によって求めた単語同士の関係を定量的に表す必要がある．本研究では，コンテンツデータから出現単語の共起確率を用いて作成する．同一文書内で頻出の単語の組み合わせは，関係性が高いという仮定に基づいているが，一般的な用語に収束してしまう可能性が高いため，様々な工夫が行われている [9]．本研究では，問題を簡略化するため，単純な共起確率からネットワークを作成する Luhn らの研究に基づいて行う [10]．

4.4　感性パラメータからの推薦コンテンツの決定

本研究における感性パラメータは任意の数の単語及び重みで構成される．ユーザがコンテンツへのアクセスを行った際に，ユーザの感性パラメータと閲覧コンテンツのもつ特徴パラメータから新たに感性パラメータの候補となる複数の単語の組み合わせを生成する．Fig. 2 に候補パラメータ生成の例を示す．

Fig. 2 では現在のユーザの感性パラメータが [GA(0.5), EC(0.5)] であるとき，[HPC(0.6), MPI(0.3)] の特徴ベクトルを持つコンテンツを閲覧した状況を表している．GA と HPC, EC と MPI という単語の組み合わせから新たに別の単語を生成し，それらを感性パラメータの候補としている．

レポートへの初回アクセス時には，コンテンツが持つ特徴パラメータを重みの合計が 1 になるよう，正規化した状態で入力される．Fig. 3 に交叉処理，Fig. 4 に突然変異処理の例を示す．Fig. 3 の親単語 A・B の交叉では単語の関係ネットワーク上での最短経路上からルーレット選択により選択を行う．この際の重みは A・B のノード間で線形となるように与える．また，パラメータ候補

から推薦コンテンツへは，すべての出現単語で構成されるベクトル空間上のユークリッド距離で最も類似しているコンテンツを提示する．

5　実験

5.1　実験目的および実験環境

本システムによってユーザの履歴から嗜好を学習し，類似するキーワードを主題とする記事が推薦結果に現れることを明らかにする．実験は IS Report システムの公開レポートデータを対象に行った．IS Report システムは同志社大学医療情報研究室が管理する研究レポート公開システムである[*2]．公開されているレポートとして研究に関する基礎知識や文献調査，研究報告などの内容であり専門性の高い用語が多数含まれている．

5.2　実験内容

本システムによるユーザのパラメータ推定を加えた推薦レポートと，パラメータ推定を加えないレポートの持つ主キーワードのみを特徴量として用いる手法を比較する．本実験では提案手法において，類似するキーワードを主題とする記事が推薦結果に現れることを明らかにするため，概念語ネットワーク導入以外の要素を極力排除した条件で行う．感性パラメータとなる子個体生成時の突然変異率は 0，学習に用いる感性パラメータ数は 1，推薦レポート数は 5 とした．なお，学習させるパラメータ数が 1 のため，感性パラメータの重みが 1 に正規化され，単語の選択確率は同一となる．実験には「The Compact Genetic Algorithm (主キーワード: GA)」レポートを用い，「PC クラスタ管理システム Sun Grid Engine の概要 (主キーワード: ジョブ)」を学習済みのレポートとして用いた．

5.3　実験結果

Table 1 に学習を行わない手法，Table 2 に提案手法においての「The Compact Genetic Algorithm」レポート閲覧時の推薦レポート及び，レポートが主とするキーワードの例を示す．提案手法においては「PC クラスタ管理システム Sun Grid Engine の概要」レポートを学習済みデータとして与えており，提案手法によって推定された感性パラメータも記載する．Table 1 に示される学習システムを組み込んでいない推薦結果では，本来の「The Compact Genetic Algorithm」レポートの主キーワードである [GA] のみをキーワードに全レポートから類似度の高い順に並べた結果である．そのため，遺伝的アルゴリズム (GA: Genetic Algorithm) に関連深いレポートが推薦結果として現れている．対して，Table 2 に示される学習システムを組み込んだ推薦結果では，学習済みレポートの主キーワードである [ジョブ] と閲覧中のレポートの主キーワードである [GA] について提案手法によって感性モデルとして新たなキーワードを生成した上で，そのパラメータに類似するレポートを推薦結果として表示している．

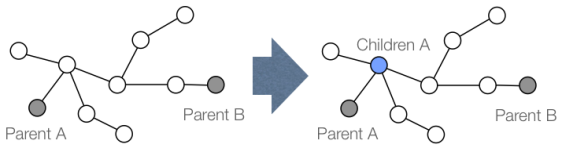

Fig.3　交叉の例

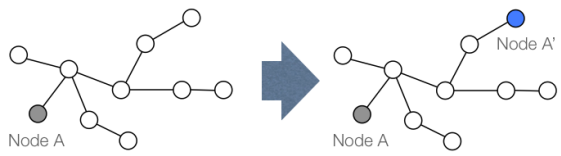

Fig.4　突然変異の例

Table1　学習なし推薦レポート

推薦レポートタイトル	主キーワード
PSA/AT(GA) の (Ala)10 への適用	AT
自作 SGA と ga2k との解探索の性能比較	ga2k
SGA の作成と動作確認	GA
MGG と sGA（simple GA）の性能比較	MGG
環境分散遺伝的アルゴリズムの検証	分散 GA

Table2　学習有り推薦レポート

感性モデル	推薦レポートタイトル	主キーワード
GA	PSA/AN(GA) における遺伝的操作	AN
パラメータ	PTH における力場パラメータの検討	パラメータ
ジョブ	大規模ジョブ問題におけるコーディング	ジョブ
実行	cron によるプログラムの自動実行	実行
パラメータ	ランキングにおける角度パラメータ	個体

5.4　考察

感性パラメータの学習システムを組み込むことによって，推薦結果に違いが見られたが，その要因となっている感性パラメータの生成について考察する．本手法による学習とは，現在閲覧中のレポートのキーワードと，過去に推定された感性パラメータとなるキーワードを概念語ネットワーク上で最短経路となる単語の一つに，次世代の感性パラメータを遷移することを指す．本実験で感性パラメータ生成の際に用いた，[GA] と [ジョブ] の単語ネットワーク上での最短経路は [GA]-[パラメータ]-[値]-[実行]-[ジョブ] で構成されており，それぞれの選択確率はすべて同一である．そのため，今までの閲覧レポートの特徴を引き継いだ上で，その単語同士の概念上で間にあたる [パラメータ] や [実行] といったキーワードが用い

[*2] http://www.is.doshisha.ac.jp/isreport

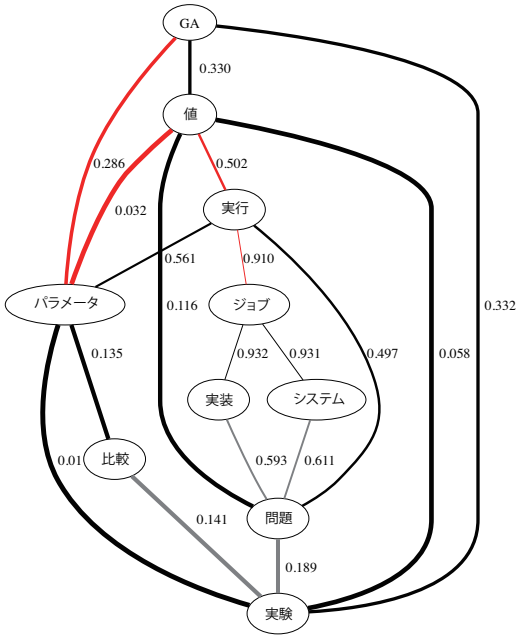

Fig.5　単語ネットワークの例

られた推薦が行われている．しかし，本実験では学習単語数を主キーワード 1 つに制限していたため，極端な推薦結果が得られたと言える．そのため，複数の単語を学習させるとこにより，複雑な特徴を受け継いだ推薦がされると期待される．

6　まとめ

本報告では，ユーザ個人の嗜好を学習することで推薦コンテンツを変化させる手法について提案した．コンテンツに含まれる特徴ベクトルと，それぞれの特徴ベクトル間の関連度を用いることで，ユーザのアクセス履歴から，ユーザの持つ感性モデルを対話型遺伝的アルゴリズムを用いることで推定する．そして推定した感性モデルに適合するコンテンツを提示する推薦システムを作成した．今までは，IS Report というテキストデータを対象に実験を行なっていたが，楽天市場オンラインショッピング公開データを対象とした際に，本手法において問題となる項目について検討した．今後は，提案手法の有用性を検討するため，被験者実験を行う．

参考文献

1) A. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pp. 271–280, 2007.
2) G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*, Vol. 7, No. 1, pp. 76–80, 2003.
3) Marko Balabanović. Exploring versus exploiting when learning user models for text recommendation. *User Modeling and User-Adapted Interaction*, Vol. 8, No. 1, pp. 71–102, 1998.
4) M. Pazzani and Daniel Billsus. Content-based recommendation systems. *The adaptive web*, pp. 325–341, 2007.
5) M. Tanaka, T. Hiroyasu, M. Miki, and H. Yokouchi. Extraction of design variables using collaborative filtering for interactive genetic algorithms. *2009 IEEE International Conference on Fuzzy Systems Proceedings*, August 2009.
6) M. Tanaka, T. Hiroyasu, M. Miki, Y. Sasaki, and M. Yoshimi. Automatic generation method to derive for the design variable spaces for interactive genetic algorithms. In *2010 IEEE World Congress on Computational Intelligence (WCCI 2010)*, July 2010.
7) G Salton, A Wong, and C S Yang. A vector space model for automatic indexing. *Commun. ACM*, Vol. 18, No. 11, pp. 613–620, November 1975.
8) A. Amamiya, M. Miki, and T. Hiroyasu. Interactive Genetic Algorithm using Initial Individuals Produced by Support Vector Machine. *The Science and Engineering Review of Doshisha University*, Vol. 50, No. 1, pp. 34–45, 2009. [In Japanese].
9) Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida. Keygraph : Automatic indexing by segmenting and unifing co-occurrence graphs. *The transactions of the Institute of Electronics, Information and Communication Engineers*, Vol. 82, No. 2, pp. 391–400, 1999.
10) H.P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal*, pp. 309–317, 1957.