知能と複雑系 116－8
（1999．5．24）

階層化された知識の継承による情報フィルタリング

沼尾正行，横山　甲

東京工業大学 大学院情報理工学研究科 計算工学専攻
Email: numao@cs.titech.ac.jp

概要:計算機ネットワーク上の情報が増大するにともない，利用者が好む情報のみを取捨選択することが困難になりつつある．このため，利用者が好むであろう情報の分野を予想し，推薦するシステムが試作されているが，嗜好を獲得するまでに多くのフィードバックが必要なことが問題になる．本論文では，ディレクトリ型検索システムの参照履歴を階層的に辿ることにより，社会的に一般的な評価，およびユーザと似た嗜好を持った他者の評価を集める手法を提案する．これにより，幅広いフィードバックを大量に得ることが容易になる．この手法を情報推薦システム FRUIT として実装し，より少ないフィードバックで有効な推薦を行なえることを実証した．

Inheriting Hierachical Knowledge in an Information Filtering System

Masayuki Numao and Masaru Yokoyama

Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152 Japan

Abstract: As contents on the computer network increases, it becomes difficult to select favorite information. Although some experimental systems are constructed to anticipate and recommend a favorite field, they require much feedback. To collect socially general evaluation and evaluation of others who have the similar preference, we propose a technique that hierarchically traces reference history of directory type retrieval system. We implemented this technique as an information recommendation system FRUIT, and verified that it recommends effectively based on less feedback.

1 はじめに

WWW をドメインとし，利用者の嗜好を推測することで，利用者が「今欲しい」と思っているページを提示するような，情報推薦システムについて述べる．

情報推薦システムは，システムに対する利用履歴を基に利用者の興味の対象を推定し，その対象に適合した情報を提示するものであるが，有効なページ推薦を行なうまでに多くのフィードバックを必要とする．本研究では，既存の手法を統合し，より少ない量のフィードバックから有効な推薦を提示し得るシステムの構築を目指す．まず，階層型知識体系上で上位のより一般的なカテゴリに対する履歴で，利用者の参照履歴を補う．同時に，社会的かつ一般的な評価傾向，および利用者と似た嗜好を持つと判断される他者の評価傾向を参考にすることで，利用者に対するページの適合度を評価する．

このような手法を WWW ページ推薦システム FRUIT (FilteRing system Using Information Tree) として実装し，ページに対する好みの推定に用いられた各フィルタリング経路の有効性を相対的に検証した．さらに，実際の使用状況に近い環境において運用を行ない，システムがどのように使用されていくか，利用者が提示された推薦を受け入れるかどうかについても評価を行った．

2 階層型知識体系の情報フィルタリングへの適用

階層型知識体系は情報の属する概念を階層的に分類してあり，利用者は自分の求める情報を徐々に詳細にして木構造を階層的に辿ることで，多くの分類の中からでも比較的容易に求める情報を引き出すことが可能である．

本研究ではディレクトリ登録型検索サーバ [6] 内に保持されている知識体系をそのまま用いる．体系によるカテゴリがページの内容を表すと仮定し，カテゴリ以外の，ページ内容に関する解析は行なわない．従って体系内のページの分類方法の正確さが，そのままフィルタリング精度に影響する．

2.1 分類ごとの利用者の嗜好の違い

人の嗜好は，情報のカテゴリごとに異なっていることが考えられる (図1参照)．例えば，スポーツ

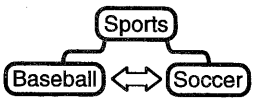

図 1: カテゴリ毎の嗜好の違い

に関して趣味を同じくする人同士が，音楽に関してもまた同じ趣味を持っているとは限らない．そこで，利用者が情報要求を出したカテゴリ毎に評価傾向を比較し，推薦者を選定することにした．

2.2 情報の階層構造の利用

利用者の参照履歴は，定義された階層全てに対して保持される．すなわち，*/Recreation/Sports/Basketball* に対する参照履歴は，カテゴリ */Recreation*, */Recreation/Sports*, */Recreation/Sports/Basketball* それぞれに保持される．したがって，上位のカテゴリには多くの参照履歴が蓄積されることになる．

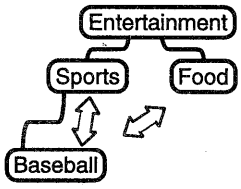

図 2: 階層構造を用いたフィルタリング

図2に示すようにより大まかな，上位に分類された分野に対する参照履歴と，より詳細な，下位に分類されたカテゴリに対する参照履歴とを用いることで，フィルタリング精度の向上をはかる．具体的には，以下のような手法を用いる．

- 上位のカテゴリの参照履歴の利用:
  参照履歴の少ないカテゴリに対して情報要求が行われた場合に行う．例えば，バスケットボールに関するページが見たい場合には，コンピューターに関して気が合う人からよりも，スポーツに関して気が合う人からの推薦の方が受け入れ易いであろうと考えられるので，そのような推薦者を選定する．
- 下位のカテゴリの参照履歴の利用:
  より細かな区分のカテゴリに対する参照履歴を比較することにより，ページに対する精度の高い評価予想を行う．

2.3 複数の分類先への履歴保持

Yahoo[6] では，他分野へのリンクとして同一ページを複数のカテゴリに分類している．本手法では，参照履歴を分類されている全てのカテゴリに保持する．例えば，「地域情報/世界の国と地域/アメリカ合衆国/趣味とスポーツ/スポーツ/バスケットボール/-NBA」から「趣味とスポーツ/スポーツ/バスケットボール/NBA」へのリンクが張られている．これを利用し，NBA に関してのページを多く参照している利用者は，バスケットボールに対して興味を持っていると共に，アメリカ合衆国に対しても興味を持っていると仮定する．これにより，推薦されるページが狭い範囲に限られてしまうことを防ぐことができる．

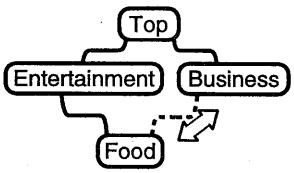

図 3: 他カテゴリへのリンクの例

3 フィルタリング手法の統合

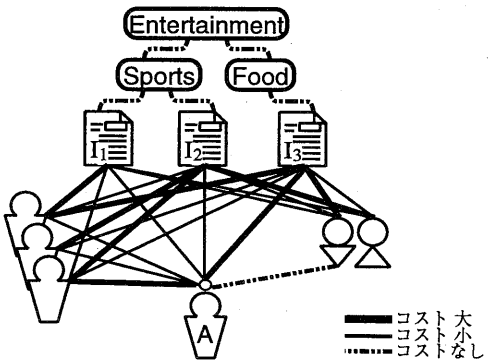

図 4: 3 種類のフィルタリング手法

利用者本人の参照履歴を用いるフィルタリングの欠点を補うため，本手法では図 4に示すように利用者のページに対する評価を予想する際に，利用者の属する社会的集団内の一般的な評価傾向を参考にする．同時に利用者本人と似通った参照履歴を持つ他者を推薦者として選定し，その人の評価傾向を参考にすることで，利用者からのフィードバック量を軽減する．

利用者 u がカテゴリ c_{req} のページの推薦を受ける場合を考えよう．c_{req} 中の各ページ i の適合度 V_i^u は，社会的評価に基づく適合度 S_i^u，推薦者に基づく適合度 O_i^u，および本人の履歴に基づく適合度 P_i^u から，次のように計算される．

$$V_i^u = S_i^u + O_i^u + P_i^u$$

この適合度 V_i^u が上位となるページをカテゴリ c_{req} 中から選んで，推薦することになる．以下，S_i^u，O_i^u，P_i^u の計算法について述べる．

3.1 社会的評価に基づくフィルタリング

利用者の属する社会的区分内での一般的な評価を，利用者個人の評価と似たものであると仮定すると，社会的評価に基づく適合度:

$$S_i^u = \sum_{s \in status(u)} E_i^s$$

ここで，$status(u)$：利用者 u の属する (複数の) 社会的区分，E_i^s：社会的区分 s に当てはまる利用者のページ i に対する評価の平均，である．すなわち，利用者の社会的区分を幾つかの観点からアンケートし，各区分毎の評価値の和を社会フィルタリング経由での評価値とする．

3.2 推薦者に基づくフィルタリング

利用者 u と利用者 o 間の距離 $D(u,o)$ を次のように定義する．

$$D(u,o) = \sum_{c \in c_{top} \sim c_{req}} |R(u,c) - R(o,c)|$$

ここで，$R(u,c)$：カテゴリ c に対する利用者 u の履歴数，$c_{top} \sim c_{req}$：情報要求の行なわれたカテゴリ c_{req} と階層最上位のカテゴリ c_{top} との間にあるすべてのカテゴリの集合，である．距離の近い利用者同士は似通った評価傾向を持つと仮定し，最も距離の近い数人を推薦者とする．

選定された推薦者の集合 $rcndrs$ を用いて，ページ i の適合度は次のように計算される．

$$O_i^u = \sum_{r \in rcndrs} P_i^r$$

3.3 本人の履歴に基づくフィルタリング

利用者本人の参照頻度の高いページと同じカテゴリに属するページは，利用者の興味を引くものであることが予想される．そこで，ページの属するカテゴリを多く参照している場合に高い評価を与える．すなわち，本人の履歴に基づく適合度:

$$P_i^u = \alpha(c(i))R(u, c(i)).$$

ここで，$c(i)$：ページiの属するカテゴリ，$R(u,c)$：カテゴリcに対する利用者uの履歴数，$\alpha(c)$：カテゴリcの深さによって与えられる補正パラメータ，である．

4 実験

利用者本人の参照履歴のみを用いるよりも，本手法が有効な推薦を行えるかどうかについて，相対的な評価を行った．推薦リストに差を生じさせるため，システムが提示するカテゴリ数は3個に絞った．被験者の嗜好を明確に反映させるため，趣味性が高いと思われる/Entertainment 及び /Recreation 以下のページに対象を限定した．

被験者は，自分の評価をシステムに全く返していない状態から実験を始める．推薦を受けた場合も受けなかった場合もシステムに対して参照履歴が蓄積されるように実験を進め，40履歴を残すまで実験を続ける．推薦として提示されたカテゴリ全てを総合的に評価し，どのシステムが面白い推薦を行なったかについて，1～3位までの順位をつけてもらった(同率の場合は，平均が2になるように調整した)．

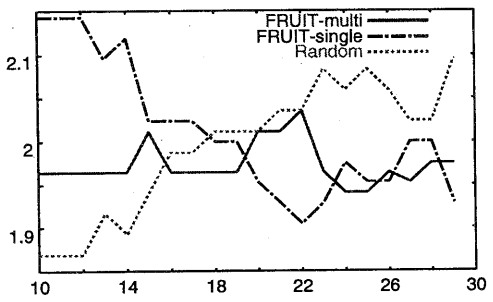

図 5: 比較実験結果

実験結果を図5に示す．FRUIT-multi は他者の履歴を用いた場合，FRUIT-single は本人の履歴のみを用いた場合，Random は評価を乱数を用いて決めた場合である．グラフの横軸はシステムに返した参照履歴数であり，縦軸は順位である．全体の傾向をつかむため，前後10個の平均をとってグラフをプロットした．1～3位までの相対評価のため，1が最も良い値となり，右下がりのグラフが得られれば学習効果が得られたことになる．

- FRUIT-multi は，全般的に平均の2を下回っており良い値を残している．やや，右下がりのグラフが得られた．
- FRUIT-single は始めは最も評価が低いが，20回ほどの履歴の学習で他システムと同等以上の評価を与えられている．
- Random で推薦を提示するシステムの評価がほぼ単調に低下しているのは，他システムの評価が相対的に上がったためである．
- 初期状態で Random が最も評価が高いのは，システム実装上評価の定まらないカテゴリに関しては，上位のカテゴリを推薦する手法を取っており，URL を含まないカテゴリを多く推薦してしまうためである．
- 20履歴までの FRUIT-multi と FRUIT-single の評価の差 (multi> single) が社会的一般的評価値，他者からの推薦を用いることの利得であり，以上の結果から，FRUIT がより少ないフィードバックで利用者の嗜好を獲得可能であることを示せた．

さらに，なるべく多くの履歴を集めることを目標に以下の実験を行った．利用者はほぼ同時期にシステムを使い始め，履歴を残すことでシステムに嗜好を学習させる．フィードバックを返していくごとに，推薦の有効性が増すかどうか，利用者がシステムの提示した推薦を受けとるかどうかを基準に絶対的に評価した．

推薦するカテゴリ数は10件として，約1ページ分のリストが提示される．候補となるカテゴリ数は最大350程で，絞り込み率はかなり高い．1998年12月25日の公開から約1カ月で，65人の利用者と約2,400件の利用履歴を得た．利用者には「ディレクトリ検索に困ったら【推薦】要求する」旨を伝えているのみで，回数，頻度は各自に任せた．

上の階層には多くの履歴が蓄積され，適合度計算時に推薦されやすくなる．この偏りを修正するた

めに，分類カテゴリ c の階層の深さに応じてパラメータ $\alpha(c)$ を変化させた[1]．履歴をまったく返していない利用者同士で距離が0になってしまうため，特別な値を距離として設定している．利用者の負担を軽減するため，推薦されたページを見た場合を1，見ない場合を0として「クリック率」だけで評価を行った．

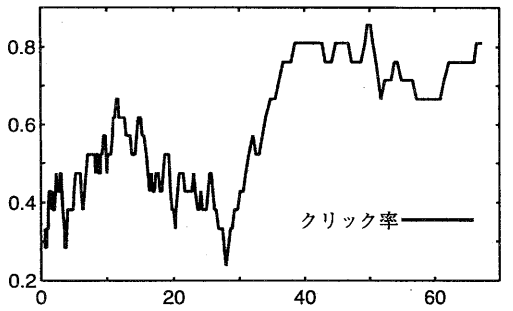

図 6: 実験結果

システムに対して50参照履歴以上を残した15人の利用者が推薦を受け入れたかどうかを，図6に示した．どの参照回数で推薦を要求しているかは利用者ごとに異なっている．縦軸は見た(=1)見ない(=0)を表し，最も良い値は1である．横軸は参照履歴数を表している．図の見易さのため，前後10履歴での評価を平均してプロットしてある．

- 利用者全体的に対して，システムは嗜好予測を徐々に正しく行なっている．
- 15人の被験者それぞれに関してもクリック率はほぼ同様の傾向を示した．
- 30履歴程でクリック率は落ち込むが，被験者の「飽き」が影響してることが考えられ，さらに解析が必要である．
- 各利用者とも平均して40履歴程を蓄積するまでに，早いペースで学習が進んでいる．
- 40履歴以上には7割5分程度の確率でシステムの推薦を受け入れている．システムの提示するページが，4回中3回は参照するに足る品質を備えていると言うことは，実用的なレベルでのフィルタリング性能を達成していると言える．
- 60履歴以上評価を返している利用者に関してプロットを続けると，データが少なくなるためにグラフの振動は大きくなるが，クリック率の上限は大体0.8程度になっている．

[1] 上位カテゴリばかりが推薦リストに載らないように調整した．単純に同層のカテゴリ数で調節できないのは[14, 15]による．

5 おわりに

本手法を他システムと比較し，表1にまとめる．従来の推薦システムは，利用者からの情報要求を想定しておらず，システムの予測する利用者の嗜好モデルはごく単純であった．たとえば，Fab[21]における成功例としてある利用者にインド料理の作り方を提供するエージェントが挙げられているが，個人の興味の全てがインド料理の作り方に向いている訳ではない．本手法は欲しい情報のカテゴリを利用者が明示できるという点で，インタラクティブな情報獲得支援を行なうことが可能であり，より実用性の高いフィルタリング手法であると言える．

参考文献

[1] 沼尾正行. Global Intelligence — 地球規模の「社会の心」における学習. マルチエージェントと協調計算ワークショップ (MACC'95), Vol. 5, , 1995.
[2] 丸谷 健介. 学習するネットワークを用いた情報のフィルタリングに関する研究. 東京工業大学理工学研究科計算工学専攻修士論文, 1997.
[3] 横山 甲. 動的なネットワークにおけるコストを用いた学習. 東京工業大学 工学部 情報工学科 卒業論文, 1997.
[4] Gregory Edwards. New Software Makes Eyetracking Viable:You Can Control Computers With Your Eyes. CSUN Alphabetical Listing of Presentations March 1998.
[5] 清水 勇喜. WWW におけるユーザの興味対象と閲覧時間の関係の調査. 情報処理学会（第57回）論文集,1998.
[6] ヤフー. Yahoo! Japan. http://www.yahoo.co.jp/.
[7] NTT. NTT DIRECTORY. http://navi.ntt.co.jp/.
[8] 村本 達也, 鷲崎 誠司. 階層型知識体系を用いたWWW情報の自動カテゴリ推定方法. 情報処理学会（第57回）論文集, 1998.
[9] 中川こころ, 高田 喜朗, 関 浩之. 可変なカテゴリ構造を用いたWWW検索支援方法. 情報処理学会（第57回）論文集, 1998.
[10] Duda, R., and Hart, P. Pattern Classification and Scene Analysis. John Wiley & Sons,New York 1973.
[11] Salton, G. Developments in automatic text retrieval. Science 253,974-979,1991.
[12] Daniel Billsus, and Michael Pazzani. Learning Probabilistic User Models. Workshop Notes of "Machine Learning for User Modeling", International Conference on User Modeling,6th,Chia Laguna,Sardinia,1997.
[13] Michael Pazzani, and Daniel Billsus. Learning

	社会	推薦	内容	情報要求	備考
本研究	○	○	△	○	階層的知識体系
WiseWire[20]	○	×	○	△	商用情報配信サービス
Fab[21]	×	○	○	×	ハイブリッドシステム
Phoaks[25]	○	×	△	×	電子ニュース&WWW
ReferralWeb[23]	○	×	×	×	Referral Chain を仮定
Siteseer[24]	×	○	×	△	ブックマークエージェント
Mori97[33]	×	○	×	△	ブックマークエージェント
GroupLens[28][29]	×	○	×	×	電子ニュース
Ringo[30]	×	○	×	×	音楽推薦
WebWatcher[31]	×	×	○	×	HTML 構造, 強化学習
LASER[32]	×	×	○	×	HTML 構造,TFIDF
Mizoguti96[26]	×	×	○	×	帰納論理プログラミング
NewT[27]	×	×	○	×	GA
SYSKILL&WEBERT[22]	×	×	○	×	BayesianClassifier[10]

表 1: 既存システムとの比較

and Revising User Profiles: The Identification of Interesting Web Sites. Machine Learning 27,313-331,1997.

[14] 幸 嘉平太, 元田 敏裕, 川崎 隆二. ジャンル検索特性に基づくディレクトリ構築法の一考察. 情報処理学会（第 57 回）論文集, 1998.

[15] 幸 嘉平太. ディレクトリの利用履歴に基づく様々なドメインレベルでの嗜好特性について. 情報処理学会（第 56 回）論文集,1998.

[16] 横山 甲. 情報の社会的評価と個人的評価を同時に獲得するフィルタリングシステム. 人工知能基礎論研究会（第 30 回）論文集,1997.

[17] 本橋 健. 情報提供サービスにおける他者グループの嗜好情報提供方式. 情報処理学会（第 57 回）論文集,1998.

[18] 佐藤 直之, 橘高 博之, 鈴木 英明. 動的な興味変化を利用したコミュニティ構築方法の検討. 情報処理学会（第 57 回）論文集,1998.

[19] Paul Resnick, and Hal R. Varian. Recommender Systems. Communication Of The ACM, pp.56-58, March Vol.40 No.3 1997.

[20] WiseWire. http://www.wisewire.com/.

[21] Marko Balabanovic, and Yoav Shoham. Content-Based,Collaborative Recommendation. Communication Of The ACM,pp.66-72,March Vol.40 No.3 1997.

[22] Mark A., Daniel B., Scotto G., Seth H., Gordon K., Dong K., Ray K., Charles L., Alexius L., Jack M., Kazuo O., Michael P., Douglas S., and Brian S., Paul Y. Learning Probabilistic User Profiles. AAAI,pp.47-56,Summer 1997. http://www.ics.uci.edu/ pazzani/Agents.html.

[23] Henry Kautz, Bart Selman, and Mehul Shah. Combining Social Networks and Collaborative Filtering. Communication Of The ACM,pp.63-65,March Vol.40 No.3 1997.

[24] James Rucker, and Marcos J.Polanco. Personalized Navigation for the Web. Communication Of The ACM,pp.73-75,March Vol.40 No.3 1997.

[25] Loren Terveen, Will Hill, Brian Amento, David McDonald, and Josh Creter. A System for Sharing Recommendations. Communication Of The ACM,pp.59-62,March Vol.40 No.3 1997.

[26] 溝口文雄, 大和田勇人. 帰納学習に基づく情報フィルタリング. 人工知能学会全国大会（第 10 回）論文集, 1996.

[27] Pattie Maes. Agents that reduce work and information. CACM,Vol.37,No.7,pp.30-40,1994.

[28] Paul R., Neophytos I., Mitesh S., Petet B., and John R. GroupLens:An Open Architecture for Collaborative Filtering of Netnews. CSCW, Chapel Hill, NC, ACM Press, pp.175–186, 1994.

[29] Joseph A., Bradley N., David Maltz, Jonathan L. ,Lee R., and John Riedl. Applying Collaborative Filtering to Usenet News. Communication Of The ACM,pp.77-87,March Vol.40 No.3 1997.

[30] Upendra Shardanand, and Pattie Maes. Social information filtering: Algorithms for Automating "World of Mouth". CHI, Denver, CO, ACM, pp.210–217, 1995.

[31] Thorsten Joachims, Dayne Freitag, and Tom Mitchell. Web Watcher: A Tour Guide for the World Wide Web. IJCAI , pp.770–775, 1997.

[32] Justin Boyan, Dayne Freitag, and Thorsten Joachims. A Machine Learning Architecture for Optimizing Web Search Engines. ... , pp.1–7, 1996.

[33] 森 幹彦, 山田 誠二. ブックマークエージェントによる WWW の URL 情報の共有. JSAI Proceedings, pp.486-487, 1997.

[34] 山田 泰資, 小林 哲則. プロキシを利用した個人データベースの自動生成とその検索への応用. 情報処理学会（第 57 回）論文集,1998.