

ハイパーリンク情報を利用した協調提示システム

齋藤 章† 中西 泰人‡

† 東京大学大学院 工学系研究科

‡ 電気通信大学大学院 情報システム学研究科

WWW上では、その構成規模の拡大と共に情報検索の困難性が問題となっている。その問題を解決するために情報フィルタリングに関する研究が数多くなされており、その中の一つに複数のユーザが協調してフィルタリングする協調フィルタリングがある。本論文では、協調フィルタリングを行なう提示システムを実装し、さらにHTMLのハイパーリンクを利用することで、ユーザがユーザプロフィールを構築する際の負荷を従来のシステムよりさらに軽減することを目的としたシステムの構築を行なった。ユーザによる評価実験を行ない、リンク情報を利用した協調提示システムの有効性を検討した。

Collaborative Recommender System using HyperLink Information

Akira Saito† Yasuto Nakanishi‡

† Graduate School of Engineering, University of Tokyo

‡ Graduate School of Information Systems, University of ElectroCommunications

On WWW, it takes much efforts and skills for users to retrieve valuable information. There are many studies about Information Filtering to try to cope with this problem. Especially Collaborative Filtering is an effective approach. We propose one Recommender System using Collaborative Filtering. Our purpose is to collect more information and construct user profiles with less effort using Hyper Link Information on HTML files. We developed a prototype system and examined our proposal with users' experiments.

1 諸言

WWWの普及と共に、多種多様なマルチメディア情報を閲覧できるようになり、多くの有用な情報を得ることが可能になった。WWWは構成規模が非常に大きく、情報の数も質も時間の経過に伴い動的に変化するため、WWW上のどこに何があるかを把握することが困難になるという問題も同時に引き起こしている。WWWの規模が増大している現状を考慮すると、情報検索の困難性はいっそう深刻になりつつあると考えられる。

こうした問題の解決策としては検索エンジンが一般的に用いられている。検索エンジンは、大きく分けて<http://www.yahoo.co.jp>のようなディレクトリ型と<http://www.goo.ne.jp>のようなロボット型とがあるが、それぞれに一長一短がある。

ユーザの要求が曖昧な場合には、ロボット型の検索エンジンに比べてディレクトリ型の方が適してい

と思われるが、その反面、情報のリアルタイム性や情報の多様性、ユーザカスタマイズといった問題点を持っているため[1, 2]、ディレクトリ型のこうした問題点を解決する一手法として、情報フィルタリングに関する研究が盛んに行なわれている。情報フィルタリングにはシステムが内容を判断した上でユーザに提示を行なう研究や、ユーザ同士が協力して情報の絞り込みをする協調フィルタリングなどの研究がある。

システムが多種多様な情報の内容を理解し、ユーザの代わりに取捨選択することは非常に困難であるため、これを回避する情報フィルタリングのシステムとして、協調フィルタリングを用いた提示システム(本稿では以下協調提示システムと呼ぶ)に関する研究も多く行なわれている[2]。

協調提示システムは、類似した興味や関心を持つ人同士をシステムが判別し、興味の対象が似通ったユーザが過去に価値のあると評価した情報を、互い

に提示させるためのシステムである。この手法では、システムが情報の内容を理解して提示を行なうのではなく、提示する情報に対するユーザの評価値をもとにして提示を行なうため、以下のような短所をもつ [3]。

1. システム内で提示しあう情報は、あるユーザが過去に評価を行ない、システム内にあらかじめ取り込んでおく必要がある。
2. システムがユーザ同士の類似度を精度よく判別するためには、ユーザは多くの情報に対して評価値を設定する必要がある。

そこで本論文では、WWW における情報フィルタリングシステムを構築するにあたって、これらの点を改善することを目的として、HTML に記述されたハイパーリンクを利用した協調提示システムを提案する。

ユーザが明示的に評価をおこなった情報だけをシステムに取り込むのではなく、その情報からリンクが張られている情報に対してもシステムがユーザの評価値を用いて自動的に評価を行なうことで、より多くの情報に関する評価を収集する。これにより、ユーザの作業負担を軽減しながらも、提示しあうための情報とユーザ同士の類似度の判定のための情報をより多く収集することが可能になると考えられる。

2 本システムの概要

2.1 本システムの特徴

WWW の大きな特徴としてリンク情報がある。これは、情報源同士の関連付けをしたものといえるため、リンク元の URL とリンク先の URL とは、内容に関して何らかの関係があるはずであり、評価値に関してもある程度同じことがいえると思われる。

一般的な協調提示システムのように、ユーザが実際に評価を行なった URL だけを用いてユーザプロフィールを作成し、それらをもとに情報を提示し合うだけでは、先に述べた協調提示システムの短所を回避することができない。そこで本論文では、システムが HTML ファイル内でリンクの張られている情報についての評価を、ユーザの評価値をもとに推定し、それらをユーザプロフィールに自動的に加えることで、より多くの情報とそれに関する評価をシステムに取り込むことを提案する。

協調提示システムでは、ユーザは自分の興味や関心を表したユーザプロフィールをシステム上に作成

する。多数のユーザがプロフィールを作成すれば、システムの上には、一種のコミュニティが形成される。システムは、様々な URL への評価値の集まりとしてユーザプロフィールを収集し、それらをもとにユーザの類似度を判定する。

新規のユーザに情報の提示を行なう時には、システムはこのコミュニティの中で興味や関心が類似しているユーザをユーザプロフィールを下に選択し、選択されたユーザが高く評価した情報を新規のユーザに提示する。この手法は、多くのユーザが登録されていることが前提であり、コミュニティが十分大きければ、各ユーザは、ある程度自分と同じ関心を持つ他のユーザを見つけられるという仮定に基づいている。

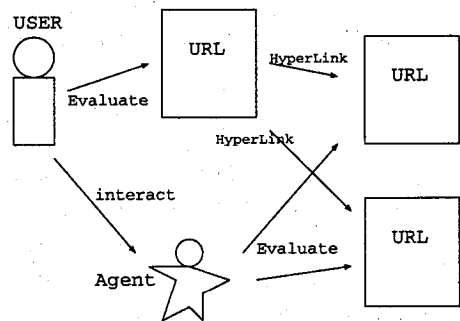


図 1: 本システムの特徴

2.2 本システムの提示手法

2.2.1 評価値の収集

本論文では、WWW の情報検索を目的としたため、ユーザプロフィールとしては URL に対する数値的な評価とし、評価は 5 段階評価 (評価が最大なものを 5 点とし最小のものを 1 点とする) とした。各ユーザは、WWW をブラウジングしながら、評価したい URL に対して 5 段階で評価を行なう。その評価値と URL がサーバに転送され、サーバでは以下の手法でユーザプロフィールを構築する。

1. URL へ対する評価をまとめた表 (以下 URL 評価表と呼ぶ) に、新たに評価を行なった URL とそれに対する評価値を追加する。
2. 評価された URL の内容を形態素解析し、単語を抽出する。
3. 2 で得られた単語の中から、事前に選択された評価の対象とする単語 (以下キーワードとす

る) 以外を不要語として取り除く。

4. ユーザの興味や関心を各キーワードに対する点数として表現した表(以下キーワード評価表と呼ぶ)を、新たに評価された URL に対する評価値とこれまでのキーワード評価表におけるデータを用いて、加重平均により学習する。

キーワード評価表の学習方法の例を挙げると以下のようなになる。ユーザが、ある URL を 5 点と評価した。この URL には、「サッカー」というキーワードが 7 個含まれており、またこのユーザのキーワード評価表の「サッカー」という項目は、21 回の学習の結果、重みが 4.2 点になっていたとする。このとき、この重みは、

$$\frac{4.2 \times 21 + 7 \times 5}{21 + 7} = 4.4$$

となり、4.4 へ変更される。この URL にその他のキーワードが存在すれば、同様の手法で重みが増え変わる。

このため各ユーザは、ユーザプロフィールとして、各 URL に対する評価値の表と各 HTML ファイルから抽出されたキーワードに対する評価値の表の 2 つをもつことになる。さらに本論文では、ユーザが評価を行なったファイルからリンクの張られている情報についても、評価値を推定した上でシステム内に自動的に取り込む。

5. 評価された URL から、絶対パスでリンクされている URL を判別し、ダウンロードする。得られた内容を形態素解析し、キーワードを抽出する。
6. これまでに作成されているキーワード評価表を用いて、リンク先の URL に対してシステムが評価を行ない、このリンク先 URL も URL 評価表に追加する。

評価値の推定の例としては、以下のようなになる。推定する URL にキーワードである「野球」が 1 個と「サッカー」が 3 個あり、このユーザのキーワード評価表における「野球」と「サッカー」の重みが、それぞれ 2.0 と 4.0 とする。この場合には、この URL に対する評価値を、

$$\frac{2.0 \times 1 + 4.0 \times 3}{1 + 3} = 3.5$$

として推定し、この評価値と URL を URL 評価表に追加する。

2.2.2 ユーザの類似度の判定

ユーザに対してシステムが新たな情報の提示を行なう場合には、URL 評価表を用いてシステムが興味や関心の類似度が高いと判別したユーザが過去に高く評価した URL を提示する。ユーザの類似度の推定は各種手法についての検証が報告されているが [4]、本論文では以下のように評価関数を決定した。

まず両者(ユーザ A とユーザ B とする)の URL 評価表の中から、共通して評価している URL (以下要素と呼ぶ)のものを抜きだし、それらの両者の評価を比較する。要素 l に対するユーザ A、ユーザ B の評価をそれぞれ、 $X_{l,A}$ 、 $X_{l,B}$ とし、両者の差を変数 $x_l = (X_{l,A} - X_{l,B})$ とする。この 2 ユーザ間の類似度 Z は

$$Z_l \stackrel{\text{def}}{=} \text{sig}(2 - |x_l|) \times (2 - |x_l|)^2 \quad (1)$$

を要素 l による類似度とした上で、

$$Z \stackrel{\text{def}}{=} \sum_{l=1}^m Z_l \quad (2)$$

となる評価関数で決定した。ただし m は要素の総数とする。システム内で類似度 Z が最大となる他のユーザを選択し、その選択されたユーザがしきい値以上の評価を行なった URL を新規のユーザに提示する。本論文ではしきい値を 4.0 とした。

2.3 本システムの実装

システムは 1 対 N のサーバクライアント型をとり、各ユーザはクライアントから評価の入力を行なう。

サーバは 1 台の Sun Ultra1 と 1 台の DOS/V の上に Java を用いて実装した。システムがリンク先を評価する際に利用するキーワードは、前述した代表的なディレクトリ型検索エンジンの一つである yahoo のカテゴリーページを形態素解析して得られた 313 個の名詞を用いた。また形態素解析には、奈良先端科学技術大学院松本研究室の茶釜 [6] を組み込み利用した。クライアントは Visual C++ を用いて実装し、ネットワークに接続している Windows マシン上で動作する。ユーザは、クライアントを用いて WWW 上をブラウジング可能であり、評価したい URL に対して評価値を入力したり、サーバに対して新たな情報の提示を要求することができる。

サーバからの情報提示は、クライアントのビュー部に表示される。

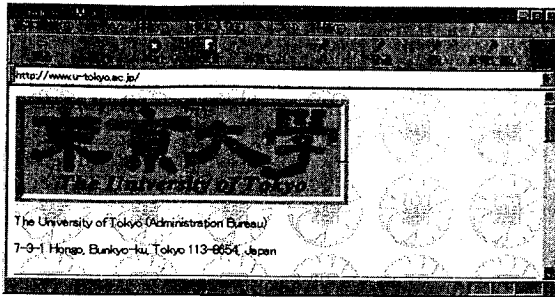


図 2: クライアント画面

3 実験と評価

3.1 実験方法

本論文で提案する手法の有効性を検証するため、リンク情報を利用しない協調提示システム (以下システム 1) とリンク情報を利用する協調提示システム (以下システム 2) について、ユーザによる実験を行ない、両者の比較を行なった。

3.1.1 初期データの構築

協調提示システムは、内含する情報を各ユーザ間で提示しあうため、コミュニティを形成するための初期データを作成する必要がある。

本論文では、以下のようにして両システムの初期データを作成した。被験者 10 人は、WWW 上の様々な情報をブラウジングし、4 点以上の評価とした URL だけ 20 個を登録する。この 10 人の被験者間では、情報のやりとりはなく、他のユーザがどの URL を評価したのかはわからない。これにより 10 人のユーザプロフィールを作成し、このユーザプロフィールを両方のシステムの初期データとした。

この合計 200 個の初期データを解析したところ、各ユーザ間で重複して評価されている URL があった。この場合には 1 つの URL であるとみなして、初期データとして含まれている URL 数を求めたところ 140 個であった。

3.1.2 両システムの比較実験方法

被験者 20 人は以下の条件にしたがって、両提示システムを使い、実験を行なった。

- 初期データを作成した被験者以外の新規に参加する被験者は、提示を受けるまえに 4 点以上の評価となる URL を 5 つ選び、その 5 つを両方の提示システムで登録する。
- 提示を受ける回数および評価をする回数は、両方の提示システムで同じとする。
- 被験者一人当たりの提示の回数と評価の回数は、それぞれ少なくとも 20 回以上とする。
- 両方の提示システムで、同一の URL が提示された場合、同一の点をつける。

クライアントのインターフェイスは、両システムとも同じとした。また被験者は、実験前においては両システムの違いを知らされていないため、どちらがどのシステムかは判別できない。

3.2 実験結果と評価

協調提示システムは、誰かが新たな URL に対して評価を行わない限り、その URL は他のユーザに提示されることがないという短所を持つ。そこでまず第一に、本システムで用いた手法が、各ユーザの負荷を軽減しつつ、より多くの新規情報をシステム内に取り込むことができるかどうかを検討した。

20 人が合計 622 回 URL に対して評価を行なった実験結果から、新規の URL がどれだけシステム内に取り込まれているかを比較した。システム 1 では 41 個の新規の URL が取り込まれたのに対し、システム 2 では、約 6 倍の 235 個の新規の URL が取り込まれた。

各ユーザが行なう URL への評価の回数は同じにするという条件で比較実験を行なったため、ユーザに対する負荷は両システムで同じである。本論文で提案する提示システムでは、ユーザが明示的に評価を行なわないリンク先の情報が自動的にシステム内に取り込まれるため、ユーザの負荷を軽減しつつシステム内の情報量を拡大することができると思われる。

協調提示システムが、各ユーザ間における類似度の判別を精度良く行なうためには、各ユーザは数多くの情報に対して評価を行なわなければならない。そのため第二に、各ユーザの負荷を増加させることなく、システムが、より多くの情報に対する評価を収集できるかどうかを検討した。

本論文で行なった実験では、ユーザの負荷は実際

に URL に対して評価を行なった回数であると考えられる。そこで、(URL 評価表における要素の数)/(URL に対するユーザの評価回数)とする変数を定義し(この変数を y とする)、これらと比較した。この変数 y は、ユーザが実際に評価を行なった URL と URL 評価表に含まれる要素の割合を示すため、この変数の値が大きいほど、より少ない負荷でより多くの情報に対する評価をシステムが収集できていることになる。

システム 1 においては、リンク情報を用いていないため、ユーザの評価回数とこのユーザの URL 評価表の要素数は等しくなるため、 y は 1 となる。一方システム 2 では、システムが各ユーザに代わりリンク先の URL の評価値を推定し、URL 評価表の要素の数を大きくする手法をとっているため、各ユーザとも 1 より大きくなる。システム 2 についての実験結果を被験者と変数 y とのヒストグラムで求めたところ図 3 のようになった。

この結果から、本論文が提案したシステムでは、

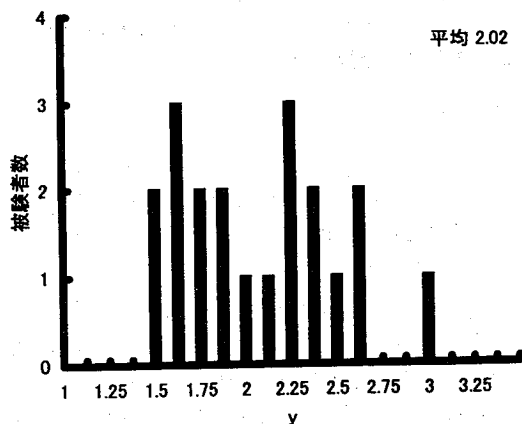


図 3: 各被験者の要素数と負荷の関係

実際に評価が行なわれた要素の約 2 倍の要素が URL 表に加えられていることがわかる。そのため、ユーザの負荷を軽減しつつも、より多くの情報に対するユーザの評価値を収集できていると考えられる。

本システムでは、リンク元に対するユーザの評価値とキーワードに対する評価表を用いてリンク先の評価値を推定する。そこで第三に、このリンク先の評価値の推定が妥当であるかどうかを検討した。この検討のため、ユーザが実際にリンク先に対して入力した評価値とシステムの推定した評価値との比較を行なった。

以下の式 (3) のように変数 α を定義し、システム

による評価値の推定の妥当性を検討した。実際の評価値と推定された評価値に近いほど、各ユーザにおける α を集計して得られる平均と標準偏差は、0 に近づく。そのため、これらの値が小さい程、システムによる評価値の推定としては優れていることになる。

$$\alpha \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n |a_i - b_i|}{n} \quad (3)$$

ただし、システムが評価値を推定した URL の数を n 個とし、URL i に対する、ユーザの評価値を a_i 、システムの推定値を b_i とする。実験結果から、各被験者の α を計測し、ヒストグラムを求めたところ、図 4 のようになった。

比較のために、キーワード評価表を利用せず、

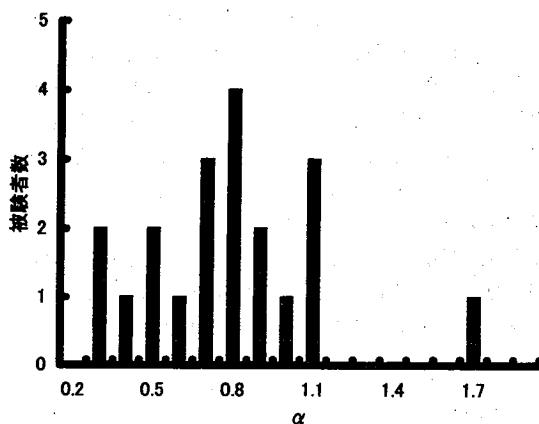


図 4: システムの予測と実際のユーザの評価との差の平均

リンク元とリンク先は同じ評価値になると仮定して α の平均を求めたところ 1.22 となり、また標準偏差の平均は 0.88 となった。それに対し、システム 2 における α の平均は 0.77、標準偏差の平均は 0.58 となった。

この結果から、ユーザの入力した評価値だけを用いてリンク先の評価値を推定するよりも、キーワード評価表を併せて用いる方法の方が、より精度よく評価値を推定していると考えられる。

協調提示システムにおいては、ユーザに新たに提示された情報に対する評価値が高ければ、効果的な提示を行なっているといえる。そこで、第四に、システムが新たに提示した情報に対するユーザの評価の比較を行なった。

システム 1 が新たに提示した情報に対するユーザの評価値と、システム 2 が新たに提示した情報に対

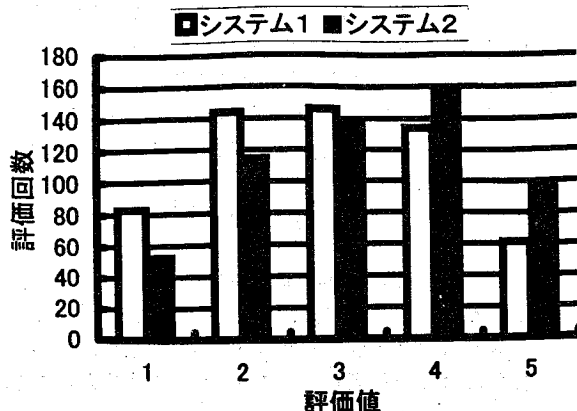


図 5: システムの提示に対する評価値の総計

するユーザの評価値を比較したところ、図5のようになった。両システムが行なった提示に対する評価の平均をそれぞれ求めると、システム1が、2.90に対し、システム2では、3.24となり、評価値の平均ではシステム2がシステム1を上回った。これは、システム1では、各ユーザプロフィールにおいてURL評価表の大きさが小さいために類似度の判別が精度良くできない、システム全体に取り込んだ情報の量が少ないため評価の高い情報を提示できない、といったことが原因であると思われる。

4 考察

本論文で提案したHTMLファイル内のリンク情報を用いた協調提示システムでは、ユーザが提示しあうための情報を自動的により多く取り込むため、各ユーザの負荷を軽減しつつより多くの新規情報を取り込むことができる。また、リンク先の情報を取り込むにあたって、リンク先の評価値をリンク元の評価値だけを用いて推定するのではなく、学習したキーワードに対する評価を併せ用いれば、自動的に取り込んだ情報の評価値をより精度良く推定できるようになることがわかった。情報の自動的な取り込みとその情報に対する妥当な評価値の推定により、ユーザの負荷を軽減しつつも各ユーザのプロフィールとシステム内の情報の量を拡大できるために、より効果的な協調提示システムを構築できると考えられる。

現在のシステムでは、リンク先に対する評価値を推定する場合、あらかじめ決められたキーワードだけを用いている。そのため今後の課題としてユーザの評価値をより良く推定するために、事前に決定さ

れた静的なキーワードだけを用いるのではなく、動的にWWW上からキーワードを抽出すること、また、各キーワード間の類別性なども[5]考慮した手法などを使う必要があると考えられる。

5 結言

本論文では、リンク情報を利用し、ユーザが明示的に評価をしていないリンク先の情報に対しても自動的に評価値を推定した上でシステムに取り込む協調提示システムを提案した。ユーザによる比較実験を行ない、システムの有効性を検討した。その結果、リンク情報を利用しない場合に比べて、1) ユーザが提示しあうための情報および情報に対するユーザの評価をより多くシステム内に自動的に取り込むことができる。2) ユーザの入力からキーワードに対する評価を学習した上で、リンク先の評価を行なう事は有効である。3) リンク先の情報の取得およびキーワードに対する評価の学習により、システムがユーザに新たに提示する情報に対する評価が向上する、ということが得られ、本論文で提案したシステムの有効性を確認した。

参考文献

- [1] 神場, 坂上, 古関: プッシュ型とプル型を統合したパーソナライズ情報配信システムの提案と実装, 情報処理学会論文誌, Vol.39, No.5, pp.1514-1522 (1998)
- [2] 平岩, 神田: Info-Plaza 他人の知識を利用した情報フィルタリングシステム, Japan World Wide Web Conference '95 (1995)
<<http://www.iaj.or.jp/w3conf-janan/proceedings>>
- [3] Balabanovic, M. and Shoham, Y: Fab Content-Based Collaborative Recommendation, CACM, Vol.40, No.3, pp.66-72 (1997)
- [4] Shardanand, U. and Maes, P: Social information filtering Algorithms for automating 'word of mouth', CHI'95 Conference Proceedings, pp.210-217 (1995)
- [5] 笠原, 松沢, 石川: 国語辞書を利用した日常語の類別性判別, 情報処理学会論文誌, Vol.38, No.7, pp.1272-1282 (1997)
- [6] <http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html>