クリックストリームに対する相関ルールを用いた情報推薦システムの実装

2001MT113 吉岡 宏晃
指導教員　河野 浩之

1　はじめに

検索エンジンに格納されている Web ページは 42 億ページ以上に達する．しかし，Web 検索エンジンによって得られた大量の検索結果の中から，個々の利用者が欲しい情報を取得することは極めて困難である．そのため，利用者にとって有用な情報だけを取得する手法の提案が望まれている [1].
本研究では, Web サイトの訪問者がクリックにより Web ページを渡り歩いた軌跡 (クリックストリーム) を分析し，その情報を相関ルールによってグループ化することで似た関心をもつ訪問者に最適なページを推薦するシステムを実装する．

2　相関ルール

アイテム集合とデータベースの中から，$X \Rightarrow Y$ というルールを抽出するための技術である [2]．相関ルールでは，ルールを抽出する尺度として支持度 (support) と確信度 (confidence) を用いる．トランザクションが X を含む確率を支持度といい，X を含むトランザクションが Y も含む条件付確率を確信度という．ここで，アイテム集合が多い場合，全ての相関ルールを調べて重要なものを選ぶという方法は，潜在的な相関ルールの数があまりに多いため効率が悪い．より効率的に価値のある相関ルールをデータベースから抽出するために，Agrawal らが提案した Apriori アルゴリズムを用いる．Apriori アルゴリズムとは，ユーザが支持度と確信度のしきい値を与え，それら以上の相関ルールを重要な相関ルールとして抽出する手法である．本研究では，相関ルールの価値を左右する支持度と確信度の最適区間を検証する．

3　CSAR システムの構成

本研究の推薦システムを CSAR[*1]システムとよぶことにする．CSAR システムは Squid，Apriori，Apache を用い，**図 1** のように構成する．次の番号は**図 1** と対応する．

1. ユーザはブラウザで Squid に Web ページ情報を要求
2. Squid はその Web ページのキャッシュがあればユーザに提供
3. もしキャッシュがなければ Internet から Web ページ情報を取得
4. Squid は取得した情報をキャッシュとして蓄積

[*1] Click Stream Association Rule.

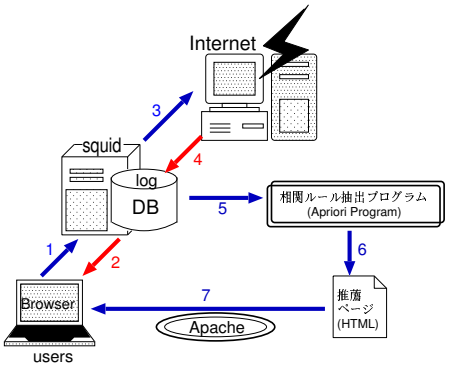

図 1　CSAR システムの構成図

5. アクセスログを加工して Apriori Program に適用
6. 5. で抽出した相関ルールを用いて推薦リストを生成し，推薦ページ (HTML) を作成
7. Apache を介して推薦ページをユーザに提供

4　CSAR システムの実装

研究室に Squid を構築し，研究生らの協力のもと，アクセス嗜好情報を蓄積する．蓄積されたアクセスログの中から IP アドレスと URL を抽出し，推薦対象外の URL(gif，css 形式など) や，一般的に知られているサイト (yahoo!，google，Nanzan) を Perl[3] を用いて除いた．そのデータを，相関ルールを生成する Apriori Program[4] にかけ，生成されたルールから推薦リストを作成した．相関ルール $C \leftarrow BE$ において「B および E を閲覧しているユーザの中で，まだ C を閲覧していないユーザに C を推薦する」といった具合である．推薦対象の研究生にデータを推薦するため，HTML と Perl-CGI を用いて推薦 Web ページを作成し，**図 2** のように Web 上に掲載した．

推薦ページの価値を左右する支持度と確信度において，確信度を以下の 3 パターンに分け，最適区間を検証する．

- パターン 1…70.0%〜90.0%
- パターン 2…50.0%〜70.0%
- パターン 3…60.0%〜80.0%

次に支持度を，推薦するページの数に着目し検証した (**図 3**)．3 パターンを比較した結果，支持度 20% の時の各ユーザに対する推薦ページの平均は約 2〜3 ページであり，推薦ページ数としては適当である．

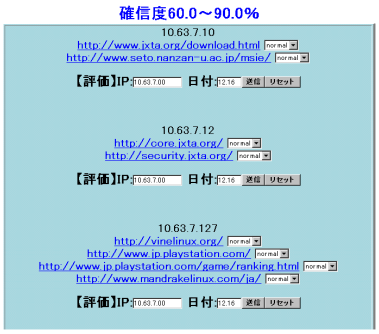

図 2　Web に掲載された推薦リスト

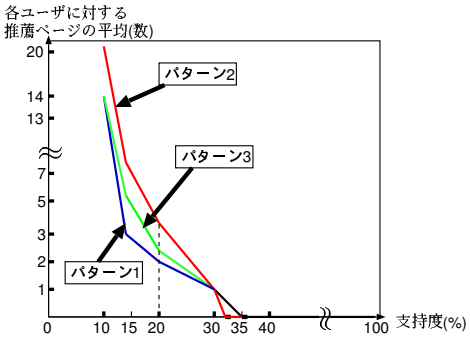

図 3　支持度に対する推薦ページ数の平均

5　推薦ページの評価

CSAR システムにおいて支持度を 20% に設定し，各 3 パターンの確信度の推薦リストを研究生 11 人に評価した結果を**図 4** に示す．ページの評価は，推薦した各 URL に，図 4 のように 7 段階の評価指数を設け，研究生一人一人の指数の平均を求め，さらに 3 パターンごと総平均を求めた．

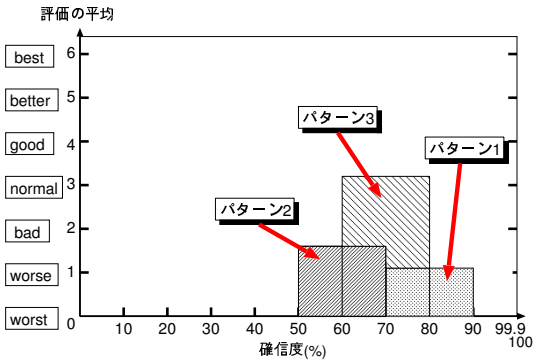

図 4　確信度に対する推薦ページの評価

全体の評価こそ良くなかったものの，3 パターンの確信度ごとでそれぞれ評価にばらつきが出た．確信度の大きい区間のパターン 1 では，推薦されるページが少なく，ページの質も研究生にとってあまり面白くないものであったため，評価は著しく悪かった．また，確信度の小さい区間のパターン 2 では，推薦されるページが，研究生にとって意外性のあるページもしくは興味のないページであり，これは研究生の嗜好によって左右されるので，極端に評価が良かったり悪かったりするページが見られた．最後に，確信度がパターン 1 と 2 の中間に位置するパターン 3 においては，他の 2 つのパターンに比べて評価が良かった．推薦されるページは最も多く，まったく興味のないページや，極端に面白みに欠けるページもあまり見られなかった．つまり，確信度が 60%～80% の区間のルールが適切であると言える．

6　おわりに

本研究では，研究生のアクセスログから相関ルールを用いて，研究生により興味のある Web ページを推薦する CSAR システムを実装した．そして，相関ルールにおいて，推薦ページの価値を決める支持度と確信度の最適区間を検証した．短期間，小規模な研究であったため，研究生の嗜好情報が少なく，また人数も少なかったため，良い推薦ページを推薦できる結果には至らなかった．しかし，より多くのユーザとその嗜好情報が収集できれば，ユーザにより興味のある Web ページを推薦できるようになると思われる．

今後の課題として，DHCP サーバからの IP アドレスの動的割り振りを静的にし，ユーザが IP アドレス一つで特定できるようにする必要がある．また，より多くの嗜好情報を収集するため，各ユーザの Cookie のキャッシュを用いる方法も有力であると考える．最後に，ユーザの嗜好情報を扱うということは，各個人のプライバシーに関わるため，厳密なデータ管理を必要とすることにも配慮する必要がある．

謝辞

本研究に対し御指導下さった河野浩之教授はもちろんのこと，アクセスログの蓄積や推薦ページの評価に御協力頂いた研究生に心より感謝致します．

参考文献

[1] 寺尾隆雄：“Web 上の情報推薦システム”，情報処理学会誌，Vol.44 No.7，pp.696-701(2003.7).
[2] 福田剛士，森本康彦，徳山豪：“データサイエンス・シリーズ 3 データマイニング”，共立出版株式会社 2001.
[3] 内田保雄，富田満：“Perl 基礎講座”，オーム社 2003.
[4] Christian Borgelt's：Apriori-Association Rule Induction，http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html (accessed 2004-11-10)．