

情報検索結果から抽出した目的毎ページセットによる
Web ページ分類提示方式

4 T-5

和田 義毅 新井 克也
NTTソフトウェア研究所

1. はじめに

WWW 上の情報量の増加により、製品比較や論文調査のような、関連情報を網羅的に収集する検索が多く行われるようになってきている。このような、情報収集型の検索では、検索サービスが推薦するURLリスト（以下：ページセット）の品質が情報の収集効率に大きく影響する。

例えば、キーワード検索のように、不要な情報を多数含むページセットの場合には、効率の良い情報収集は困難である。検索者に効率よく情報収集させるためには、ディレクトリのような人間の手で目的別に分類した、高品質のページセットが必要となる。

しかし、ディレクトリの登録情報量はあまり多くないため、目的の情報が見付けられないことも多い。このため、多くの場合には、効率の悪いキーワード検索を用いて情報収集を行わなければならないのが現状である。

そこで、本研究では、多数の検索者の共同作業により情報を検索目的毎に分類することで、大量情報分類を実現することを目的とする。その実現手段として、検索目的毎に情報を分類することが情報検索の手段となる、新しい情報検索方式を提案する。

2. 提案方式

Web ページの分類に検索者の協力を得るためには、検索者に何らかのメリットを提供する必要がある。この点、協調フィルタリング^{[1][2]}は、参照情報に対する評価履歴を用いて、推薦情報の相互提供を実現しており、検索者に情報分類によるメリットを提供するための有望な手段といえる。

しかし、協調フィルタリングでは、個人の嗜好傾向の類似度に基づいて情報推薦を行っているため、評価履歴を個人毎に作成している。そのため、図1に示すように、評価履歴には、異なる検索目的での評価や、無目的なブラウジング中の評価などが混在している。本方式では、個人がある時点で持った検索目的の類似度に基づく推薦を行う必要があるため、検索目的別の評価履歴を作成する必要がある。

そこで、本研究では、評価履歴を検索目的別に管理する手段を新たに設け、検索目的毎のページセット(目的毎ページセット)を作成できる推薦情報の相互提供システムを提案する。これにより、本システムを利用する全ての検索者から目的毎ページセットを収集し、検索者に品質の良い目的毎ページセットを提供することが可能となる。

3. 評価履歴の目的別管理手段

本方式では、評価履歴を検索目的別に管理するための方法として、以下の3つの手段を用いた。

まず、第1の手段として、検索アプリケーションの終了や長時間の検索行動中断を検知し、検知時に検索目的変更の有無を確認するダイアログを表示した。検索目的の変更時には検索行動が中断されることが多いため、これにより、複数の目的での評価履歴が結合されてしまうことをある程度防止できる。

第2の手段としては、検索者が検索目的の変更を意思表示するための入力手段を用意した。これは、検索中断なしで検索目的を変更する場合でも、評価履歴を検索目的別に管理できるようにするためである。具体的には、新規検索、検索終了、検索中断、検索再開の意思表示のための入力手段を用意した。

第3の手段として、本方式では評価履歴を検索目的別に管理することが、適切な推薦を受けるための条件となっている。本方式では、評価履歴がキーワード検索のキーワードに相当し、推薦情報を決定するための検索キーである。他の検索目的での評価履歴を混入させた場合、検索精度が低下し、検索者自身にとって不利益になる。このため、検索者自身が評価履歴を検索目的別に管理することが期待できる。

4. 推薦情報の選択方法

本方式で用いた情報推薦方法は、検索目的別に作成した評価履歴の類似度に基づくものである。具体的には、検索中に参照したWeb ページに対して必

評価履歴	
情報ID	評価
http://wwwaaa.co.jp	必要
http://wwwbbb.co.jp	必要
http://wwwccc.com	不要
http://wwwddd.co.jp	必要
http://wwwkkk.com	不要
http://wwwlll.co.jp	必要
http://wwwmmm.com	不要
http://wwwxxx.ne.jp	必要
http://wwwyyy.or.jp	不要
http://wwwzzz.ac.jp	必要
http://www111.com	不要
http://www222.com	不要
....

→ 検索目的 A (必要, 必要, 不要, 必要)

← 無目的なブラウジング (不要, 必要, 不要)

→ 検索目的 B (必要, 不要, 必要, 不要)

図1 検索目的別に管理されていない評価履歴

* Web Page Classification to Purposeful PageSet Using Information Retrieval Results
 † Yoshinori Wada, Katsuya Arai
 ‡ NTT Software Laboratories

要/不要の評価を検索者に行わせ、この評価履歴と、システムに蓄積されている他の検索者が作成した評価履歴の類似度をスコア化し、スコアが最大のものを選択する方法を取った。スコアは、同じ評価があればスコアを+1し、異なる評価があれば-1する単純な方法を使用した。推薦するページセットは、選択された評価履歴の中から必要と評価されているページを抽出することで作成した。

5. 性能評価

5.1 評価項目

本方式は、検索者から収集した評価履歴を元に推薦情報を作成するため、ユーザー数が少ない場合には推薦情報が作成できないケースが発生する。このため、まず、確実に情報推薦を行うために必要なユーザー数の概算値を調べておく必要がある。

また、本方式では検索目的毎に評価履歴を新規作成するため、推薦を受けるためにはある程度の評価履歴を作成する必要がある。この評価履歴の大きさについても検証を行う。

5.2 評価モデル

本方式が機能するために必要なユーザー数を試算するために、Yahoo![®]を評価モデルとして利用する。Yahoo!は2万5千のカテゴリに50万のサイトを登録⁴⁾しており、ディレクトリとしては最大の規模を持つ。Yahoo!の最下層のカテゴリから平均的な大きさの一つを選択し、このカテゴリ内で推薦を実現するための必要ユーザー数を求めることで、Yahoo!規模の分類を本方式を実現するための必要ユーザー数を試算する。また、この中で、推薦を受けるために

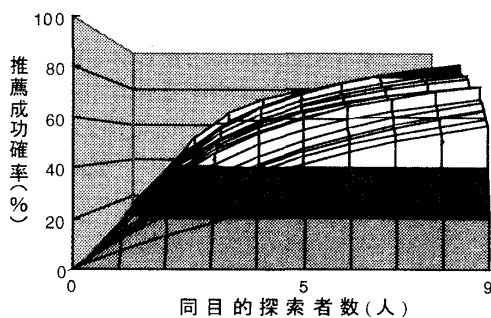


図2 同目的探索者数と推薦成功確率

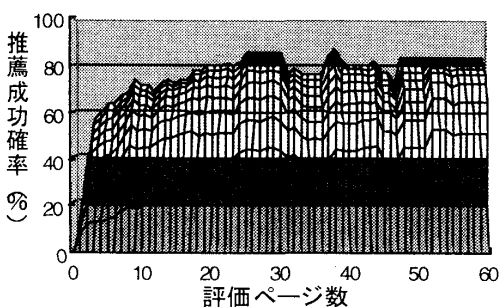


図3 評価ページ数と推薦成功確率

作成しなければならない評価履歴の大きさについても検証する。

5.3 評価結果

Yahoo!の平均的な大きさのカテゴリとして風邪(influenza)のカテゴリを選択し、これを分割した3つの検索課題を設定した。ここで、3つの検索課題を設定したのは、Yahoo!をモデル化した場合、各階層で平均3分岐するためである。実験では、風邪の予防・治療に関して①食事療法、②生活習慣、③民間療法に関するものの検索という課題を用いた。

この課題で各10人分ずつ評価履歴を作成し、同目的の検索者が過去に何人いれば推薦が受けられるかを検証した。この時の、同目的の検索者数、及び、評価履歴の大きさで推薦成功確率の関係を図2、図3に示す。図2のように、同目的の検索者5人以上では、推薦確率の上昇がほぼ飽和状態になり、80%程度の推薦確率が得られた。また、図3より10~20ページの評価履歴を作成すれば、ほぼ上限に近い推薦確率が得られることが分かる。これはキーワード検索の検索結果1~2枚分の大きさであり、十分許容できる量であると考えられる。

また、Yahoo!規模での必要ユーザー数は、1課題当たり5人必要な場合、1カテゴリ当たり15人必要であり、これをYahoo!のカテゴリ数倍すると、37万5千人となる。しかし、これは、一人のユーザーが一個の評価履歴しか作成しない場合の数である。仮に、一人のユーザーが月に平均1個の評価履歴を作成し、評価履歴は1年間有効であるとした場合、必要ユーザー数は31250人となる。

このことから、Yahoo!規模の情報分類を行うために必要なユーザー数は、数万人程度であろうと推定される。

6. まとめ

今回の実験では、本方式で80%程度の推薦成功確率が得られることと、Yahoo!規模の情報分類を行うためには数万人のユーザーが必要であるという結果が得られた。今後は、検索目的の数を増やした場合にも、本方式による推薦が機能するかどうかを検証し、本方式の実現性を検証する予定である。

参考文献

1. Resnick, P, Iacovou, N, Suchak, M, Bergstrom, P, Reidl, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. CSCW94 Proc. Pp.175-186(1994)
2. Shardanand, U. and Pattie Maes Social Information Filtering: Algorithm for Automating "Word of Mouth", CHI95 Conference Proc pp.210-217(1995)
3. <http://howto.yahoo.com>
4. <http://howto.yahoo.com/chapters/7/1.html>