2C2-02

Bloggerの嗜好を利用した協調フィルタリングによるWeb情報推薦システム

A Web Contents Recommendation System based on Collaborative Filtering by Using Bloggers' Interests

小原 恭介[*1] Kohara Kyosuke　山田 剛一[*1] Yamada Koichi　絹川 博之[*1] Kinukawa Hiroshi　中川 裕志[*2] Nakagawa Hiroshi

[*1] 東京電機大学大学院 工学研究科 Graduate School of Engineering, Tokyo Denki University

[*2] 東京大学 情報基盤センター Information Technology Center, The University of Tokyo

With the growth of internet, we can get tremendous amount of contents on the web. In this situation, it is difficult to get information that we are interested in. So, we need a system that is able to recommend desired information. In this paper, we propose a recommendation system based on a new collaborative filtering by using bloggers' interests which we can get by analyzing the reference from blog entries to web contents. By the proposed filtering, we can solve the following two problems of the traditional collaborative filtering: (1) the "cold-start" problem; (2) the problem which preserves system reliability from attack of malicious users.

1. はじめに

現在、Web 上では様々な情報が日々公開され、その数は半年ごとに2倍に増加していると言われている。このような状況の中で、ユーザにとって興味や価値のある情報を追い続けていくのは困難である。そこで、これらの情報を推薦してくれる情報推薦システムが実現できれば利用価値は高い。

推薦システムを実現するには、ユーザにとって未知の情報をユーザの嗜好を考慮してフィルタリングする必要がある。このフィルタリング方法としては、協調フィルタリングが有効であることが知られているが、コールドスタートと呼ばれる問題や、悪意あるユーザの攻撃といった問題が存在する。

本稿では、これらの問題を解決するために、Blog 記事を収集し解析することによって得られる、Blogger の嗜好を利用したフィルタリング方式を提案し、それを用いた Web 記事推薦システムの概要について述べる。

2. 情報フィルタリング方式

推薦システムの核となる情報フィルタリング方式は、キーワードや単語出現頻度などの情報を用いた内容に基づくフィルタリング方式と、情報に対する他のユーザの評価に基づく協調フィルタリング方式に大別される。

2.1 内容に基づくフィルタリング方式

内容に基づくフィルタリング方式は、過去にユーザが好んだアイテムの内容を分析することでユーザの嗜好傾向を示したユーザプロファイルを生成し、そのユーザプロファイルに類似した情報を推薦する方式である。ユーザプロファイルはユーザの関心を表現するキーワードベクトルからなり、このベクトルと、各情報内容を表すキーワードベクトルとの類似度を算出することで情報を選別する。この方式の問題点は、ユーザが過去に高い評価をした情報に類似した情報ばかりが推薦されてしまうため、ユーザの嗜好変化に追随することが困難なことである。よって、ユーザに新たな嗜好の発見を促す作用が働かない閉鎖的なシステムになってしまう。また、キーワードなどで表せない図表などが持つ価値を推薦に織り込むことが難しいという問題もある。

2.2 協調フィルタリング方式

協調フィルタリング方式は、ユーザ A と関心が類似した別のユーザ B が高く評価した情報を、ユーザ A にも薦めるというものである。具体的にはユーザの各情報に対する評価値を元に、ユーザ間の類似度を計算し、類似度の高いユーザの評価情報を優先的に参照し推薦を行う。この方式では内容に基づくフィルタリングのように推薦内容が収束してしまう問題もなく、また、情報自体がどのような特性をもっているかに言及しないため、映画や音楽や Web ページなどといったどのような情報にも適用できる利点がある。

しかし、内容に基づくフィルタリング方式が1ユーザで行えるのに対して、協調フィルタリング方式では複数のユーザの情報を使う必要があり、そのため以下の問題が発生する。

(1) コールドスタート問題

協調フィルタリング方式を用いた推薦システムで有効な推薦を行うには、数多くのユーザの参加が不可欠である。しかしながら、推薦システムの初期段階では、ユーザ数の少なさから類似ユーザが発見できず、有効な推薦が期待できないという問題[Schein 2002]が発生する。

(2) ユーザの信頼性の問題

ユーザ間の類似度を求める際に、各情報に対する評価の類似度のみを用い、ユーザの信頼度は考慮されていない。このため、悪意あるユーザが推薦システムを利用した場合、その推薦アルゴリズムを逆手にとり、システムの推薦精度を落とすように行動することが可能であるという問題がある。[Massa 2003]

3. Blogger の嗜好を用いた方式の提案

2.2 節で述べた問題を解決するために、本研究では、各 Blog の作成者である Blogger を協調フィルタリングにおける仮想ユーザとして用いる方式を提案する。

3.1 Blog の特徴

Blog とは、ネットで見つけた面白い Web サイトや、ニュース記事へリンクを張り、自分の意見などを記述した記事が時系列

連絡先: 小原恭介 〈 kohara@cll.im.dendai.ac.jp 〉
東京電機大学大学院工学研究科情報メディア学専攻
〒101-8457 東京都千代田区神田錦町 2-2
Tel: 03-5280-3631

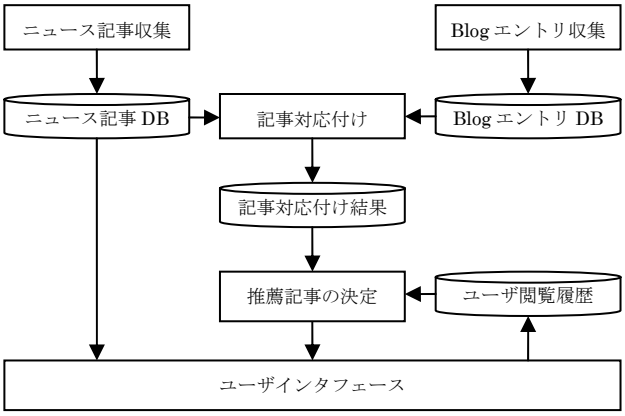

図 1. Web 記事推薦システム概要

図 2. システムの GUI

に配置されている Web サイトとされている。しかし、厳密な定義はなく、個人の日記的な Web サイトの総称となっている。HTML の記述が必要ない weblog ツールの普及や、Blog サービスプロバイダの登場により、誰でも手軽に開設・更新できることから近年急速に広まった。調査[Technorati 2005]によると日本の全インターネットユーザ 6,500 万人のうち Blog 作成経験者は 255 万人にも達している。

3.2 協調フィルタリングへの利用

ある Blog の1つの記事(エントリ)に着目すると、そのエントリ中で参照している Web サイトやニュース記事等は、その Blog の作成者(Blogger)が興味を持った情報の対象であり、また、そのエントリ内の本文はその対象に対する評価を表すものである。よって、Blog を大規模に収集し解析すれば各情報に対して多数の Blogger の評価が得られる。このことから、Blogger を協調フィルタリングにおける仮想的なユーザとして利用できると考えられ、この仮想ユーザ数の多さを考慮するとコールドスタート問題は発生しない。ここで、リンク先の情報に対する評価値は、エントリの本文から算出できれば良いが、技術的に困難である。このため、リンクしたという行動自体を Blogger の興味の表れと捕らえ、リンクの有無の2値を協調フィルタリングに用いる。

ユーザの信頼性問題に関しては、Blogger が直接システムのユーザでない点を考えれば安全であると考えられる。また、各 Blogger の信頼度をトラックバック数やコメント数等の Blog 固有の特徴を利用して計ることで、推薦の期待値に反映させることも可能である。

4. Web 記事推薦システム

前節で説明した提案方式を用いた推薦システムの概要を図 1 に示す。推薦対象は、収集の容易さと Blog エントリとの対応付けのしやすさを考慮して Web 上のニュース記事とした。システムは次のパートから構成される。

(1) ニュース記事の収集

国内の約60のニュースサイトに対して数十分おきにニュース記事を巡回取得するクローラーを作成した。ここでは、各記事の URL、タイトル、本文、作成時刻の抽出を行う。

(2) Blog エントリの収集

Blog エントリの収集は、エントリの更新が通知される update ping サーバに対する巡回を随時行い、各 Blog サービスプロバイダのトップページに対する巡回と既知の Blog に対する巡回を定期的に行う。メタデータや本文の抽出は可能な限り RSS によって取得を行うが、取得できないものにたいしては HTML を解析し抽出を行う。

(3) Blog エントリとニュース記事の対応付け

Blog エントリとニュース記事との対応付けは、現段階では単純に Blog エントリ内にニュースソースへのリンクの有無の2値で行っている。しかし、これではリンクをはらずにニュース記事を一部引用するのみの場合は対応が把握できない。また、同内容の記事でも違うニュースソースへのリンクを別に扱ってしまうため、推薦精度に悪影響をおよぼしている可能性がある。

これらは今後、文章の類似度を用いるなどして、同一に扱えるように改良したい。

(4) 推薦記事の決定

ここでは、Blog エントリとニュース記事の対応付け結果とユーザ閲覧履歴を用いて協調フィルタリングを行い、ユーザに推薦する記事の決定を行う。

(5) ユーザインタフェース

システムのインタフェースでは、収集されたニュース記事から時系列による表示と、推薦記事の表示が行われる(図 2)。ユーザがクリックした記事はシステムによって記録され、協調フィルタリングに利用される。ユーザは各記事に対する評価値を入力することはなく、記事をクリックしたかどうかの2値を用いる。

5. おわりに

提案方式を利用したシステムにより、ユーザの嗜好にあったニュース記事の推薦が可能となった。今後、システムの推薦精度の評価を行う予定である。また、提案方式の推薦精度向上という観点では4章で述べた対応付け方法の強化を行いたい。

協調フィルタリングには、誰も評価していないものは決して推薦されないという特徴があるため、本システムにおいても公開された直後の記事は推薦されにくい点が課題となる。これを改善するため、内容に基づくフィルタリングとのハイブリッド方式の検討も視野にいれている。

参考文献

[Schein 2002] A. Schein, A. Popescul, L. Ungar, D. Pennock: Methods and metrics for cold-start recommendations, 25th Annual ACM SIGIR Conference, 2002, pp. 253-260

[Massa 2003] Paolo Massa: Trust-aware Decentralized Recommender Systems, Phd Proposal,2003, University of Trento, http://sra.itc.it/people/massa/massa03trustaware.pdf

[Technorati 2005] Technorati: http://www.masahikosatoh.com