E_007

Bloggerの嗜好を利用した協調フィルタリングと内容類似性によるWeb情報推薦システム

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

寺田 道生[†] Michio Terada　小原 恭介[†] Kyosuke Kohara　山田 剛一[†] Koichi Yamada　絹川 博之[†] Hiroshi Kinukawa　中川 裕志[‡] Hiroshi Nakagawa

1. はじめに

現在 Web 上で配信されるニュース（以下 Web 記事）の利用が高まっており、その需要に沿うように現在日本において一日あたり 1000 件以上 Web 記事が配信されている。

Web 記事に対する推薦手法として Grouplens[1]など、協調フィルタリングを利用したものがある。小原ら[2]は Blogger の嗜好を利用した協調フィルタリングを提案・実装している。Blogger を利用することで協調フィルタリングが抱える、コールドスタートや悪意あるユーザによる攻撃といった問題を解消している。

しかしながらリンク情報だけでは、Web 記事の内容情報が加味されない。そこで本研究では小原らの手法に加え、あらかじめ Web 記事の内容類似性を調べてトピック別に分類し、Blogger の嗜好を Web 記事単体ではなく、トピック単位に拡張することで、更なる推薦精度の向上を目指す。

2. 協調フィルタリングによる情報推薦

小原らの Blogger の嗜好を利用した協調フィルタリングによる Web 情報推薦システムの概要を説明する。

2.1 協調フィルタリング

協調フィルタリングとは、ユーザ A と関心が近いユーザ B が好む情報を、ユーザ A にも推薦する方法である。小原らは Blogger を協調フィルタリングにおける仮想ユーザとし、Web 記事の推薦システムを構築した。

2.2 協調フィルタリングへの Blogger の嗜好の適用

Blogger が書いた一つの記事（エントリ）に注目すると、Blogger が興味を持った Web 記事への感想や意見とともに、ニュースソースへのリンクがよく見受けられる。これはエントリを解析することで、その Blogger が過去にどんな Web 記事に興味を持ったのかという嗜好が分かることを示している。これを利用し、現在既に多く存在する Blogger[3]を、協調フィルタリングにおける仮想ユーザと見なす。これにより協調フィルタリングが抱える、推薦において多数のユーザを必要とするコールドスタート問題、推薦精度を落とすように行動するユーザによる信頼性低下の問題を解決することができる。

リンクをしたという行為自体を、Blogger の Web 記事への興味の表れととらえ、リンクの有無の２値を協調フィルタリングに用いる。

[†]東京電機大学大学院工学研究科
[‡]東京大学情報基盤センター

3. リンクの有無のみを用いた評価の問題

リンクの有無を、Blogger の Web 記事に対する評価として利用することを前章で述べたが、これには Web 記事の内容が考慮されていないため、同じトピックを扱う Web 記事に対して Blogger がリンクしていても、ニュースサイト間の違いによってアドレスが異なっていれば、そのリンクにおける興味対象は別々に扱われてしまうといった問題が発生する(図１)。

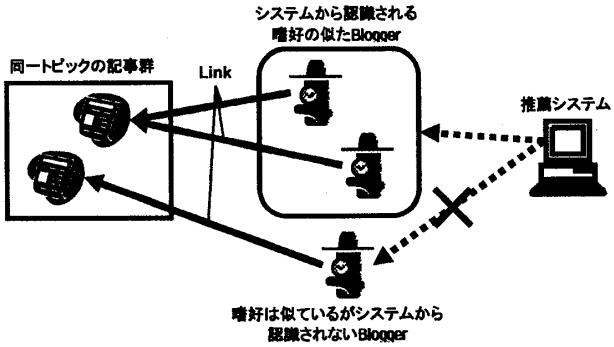

図１．興味対象が記事単位の場合の問題点

4. 内容類似性による Web 記事トピック分類

３章で述べた問題を解決するため、本研究では収集した Web 記事をあらかじめトピック別に分類することで、ユーザのリンク対象の範囲を記事単位からトピック単位へと拡張する。これにより同一トピックの複数の Web 記事を同一視することができ、ニュースサイト間の違いや、トピックの続報の Web 記事に対して対処することができる。この章では Web 記事をトピック別に分類する分類クラスタ作成の方式を説明する。

4.1 Web 記事のベクトル化

収集した Web 記事の語に TF・IDF 法による重み付けを行い、記事の内容を特徴づける語群を生成する。

(1) 語の抽出範囲

収集された Web 記事から語の抽出を行う。既にタイトルと本文に分けてデータベースに格納されているため、タイトルはそのまま利用し、本文に関しては 100 文字を超えてから、最初に出現した句点か改行までを抽出範囲とする。これは Web 記事に関わらずニュース記事というものが、記事の冒頭にその概要を簡潔にまとめるという特性に基づくものである。

(2) 特徴語群の生成

抽出した語に対して重み付けを行う。重み付けとして先述したように TF・IDF 法を用いる。DF 算出元として、あらかじめ(1)と同様の範囲で抽出した、特定期間の Web 記事の語群の DF 値を用意し、利用する。

タイトルに出現した語に関しては更に 2 倍の重み付けを行い、最終的に得られた重みの値の上位最大 50 語をその Web 記事の特徴語群とする。

4.2 分類クラスタの構成

Web 記事の同一トピックを表現する分類クラスタは、4.3 節で説明する類似度がしきい値以上の Web 記事群の特徴語群によって構成される。

同一トピックの各 Web 記事に、4.1 節の説明と同様の方法で語に重み付けを行う。そして全ての語を集約させて、重みの値の上位最大 50 語を当該クラスタの特徴語とする。分類クラスタを一つの Web 記事のように見なすため、クラスタに含まれる Web 記事数が増えるにしたがって、出現しやすい語に関しては TF が増加しやすい傾向にある。そのため分類クラスタの特徴語の重みを、クラスタに含まれる Web 記事数で正規化している。

4.3 Web 記事の分類

収集された Web 記事と分類クラスタとの類似度を計測し、分類クラスタの更新や新規生成を行う。

(1) 類似度の算出方法

Web 記事、分類クラスタとも、類似度を 4.1 節で説明した特徴語のベクトルで表現し、コサイン距離によって算出する。

(2) トピック別分類手法

Web 記事をトピック別に分類する手法として、逐次新規 Web 記事と既存の分類クラスタの類似度を測り、しきい値以上の場合当該クラスタに追加していく、1 パス法と呼ばれる手法を取る。

どのクラスタにも追加されなかった場合、その Web 記事を元に新たなクラスタを生成する。その際のクラスタの特徴語群は、新規 Web 記事の特徴語群のみとなる。

5. Web 情報推薦システム

2 章と 4 章の手法を組み合わせた Web 情報推薦システムを図 2 に示す。システムは次の構成となる。

(1) Web 記事・Blog エントリの収集

国内約 60 のニュースサイトに対しクローラを走らせ、数十分おきに Web 記事を取得する。Blog エントリはエントリ更新情報が集約される ping サーバ、各 Blog サービスのトップページ、既知の Blog を定期巡回し取得する。

(2) トピック分類

4 章で説明した手法を用いて、収集された Web 記事をトピック別に分類する。

(3) Blog エントリとトピックの対応付け

Web 記事をトピック別に分類する際に、どの Web 記事がどの分類クラスタに所属しているかを表す対応表を作成する。Blog エントリに含まれるリンクと対応表を照らし合わせ、Blog エントリとトピックを対応づける。

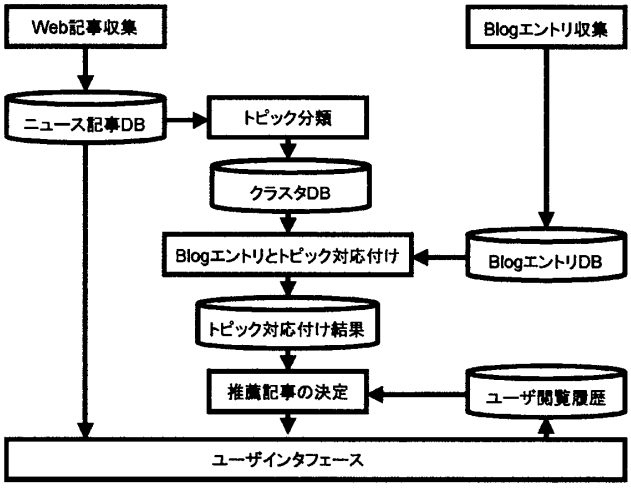

図 2．Web 情報推薦システム

これによりリンク情報をもとにした Blogger の嗜好が、従来の Web 記事単位から、トピック単位へと拡張される。

(4) 推薦記事の決定

(3)の対応付け結果とユーザの Web 記事閲覧結果を用いて協調フィルタリングを行い、推薦する記事を決定する。なお閲覧結果もトピック単位へと拡張し、ユーザにはトピック単位で推薦を行う。

(5) ユーザインタフェース

収集され時系列に表示された Web 記事の表示と、閲覧結果による推薦トピックの表示を行う。ユーザがどの記事のリンクをクリックしたかを保存しておき、その情報を協調フィルタリングに用いる。ユーザが評価値を決めることはなく、記事へのアクセスの有無のみを利用する。

6. おわりに

Web 記事をトピック別に分類することにより、推薦の対象範囲をトピック単位へと広げ、Web 記事の内容類似性を加味した協調フィルタリングによる Web 情報推薦システムを提案した。今後は Blogger の嗜好が記事単位の時に比べ、どの程度推薦精度が向上したのかを評価する予定である。

協調フィルタリングでは、誰も評価していない対象を推薦することができず、発行されたばかりの Web 記事に対応できない。最新の Web 記事に関しては、内容に基づくフィルタリング方法を行う方式を検討する。

参考文献

[1] 小原恭介，山田剛一，絹川博之，中川裕志：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，第 19 回人工知能学会全国大会，2C2-02C，北九州，2005.

[2] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. GroupLens: Applying collaborative filtering to usenet news. Communications of theACM, Vol.40,No3, pp. 76-87, 1997.

[3] 総務省：ブログ及び SNS の登録者数（平成 18 年 3 月末現在），http://www.soumu.go.jp/s-news/2006/060413_2.html