D-031

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 情報推薦システムのためのクラスタリング手法の検討

A Study of Clustering Method for Web Contents Recommendation System based on Content Similarity and Collaborative Filtering by Using Bloggers' Interests

奥山 透† 寺田 道生† 小原 恭介† 山田 剛一† 絹川 博之† 中川 裕志‡
Toru Okuyama Michio Terada Kyosuke Kohara Koichi Yamada Hiroshi Kinukawa Hiroshi Nakagawa

1. はじめに

近年、Web 上で配信されるニュースの利用が増えている。ユーザの需要に合うように、様々なジャンルで膨大な数のニュースが配信されている。この増加するニュースの推薦手法として、小原[1]は Blogger の書いた記事内のリンク情報を利用した協調フィルタリングを提案・実装し、寺田[2]は小原のシステムに、記事内の内容情報を利用したクラスタリングを組み合わせることを提案している。このクラスタリングにおいては生成するクラスタの大きさが推薦精度に大きな影響を与えるため、本研究では Web 情報推薦システムに適したクラスタリング手法について検討した。

2. 協調フィルタリングによる情報推薦

小原の 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 が書いた記事（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。これを利用し、現在既に多く存在する[3]Blogger 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の信頼性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。

2.3 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 URL 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。

†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. 内容類似性による Web 記事トピック分類

2.3 節で述べた問題を解決するため、寺田は収集した 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した。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。

3.1 Web 記事のベクトル化

収集した Web 記事の語に重み付けを行い、記事の内容を特徴づける語群を生成する。はじめに、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。次に、抽出した語に対して重み付けを行う。重み付けには TF･IDF 法を用いる。DF 算出元の初期 Web 記事集合として、あらかじめ抽出範囲と同様の語の範囲でインデクシングされた 2004 年 10 月からの Web 記事 10 万件を利用する。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。

3.2 分類クラスタの構成

分類クラスタは類似度が閾値以上の Web 記事群の特徴語群によって構成する。Web 記事集合における同一トピックの Web 記事群に **3.1** 節の説明と同様の処理を行い、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。なお分類クラスタの特徴語群にはクラスタの大きさによる正規化を行っており、語の重みの値をクラスタに属する Web 記事数で割る。これにより同一トピックの Web 記事を多く含む分類クラスタにおいて、TF が増加し語の重みが高くなることを防いでいる。

3.3 Web 記事のクラスタリング

新たに収集された Web 記事と既にできている分類クラスタ（**3.2** 節）との類似度を計測し、分類クラスタの更新または新たなクラスタの生成を行う。類似度は **3.1** 節で説明した特徴語のベクトル利用し、コサイン距離によって算出する。

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、閾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。もし、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、**3.2** 節で説明したものを適用する。

3.4 Web 情報推薦システム

以上の手法を組み合わせた Web 情報推薦システムの構成を図 1 に示す。

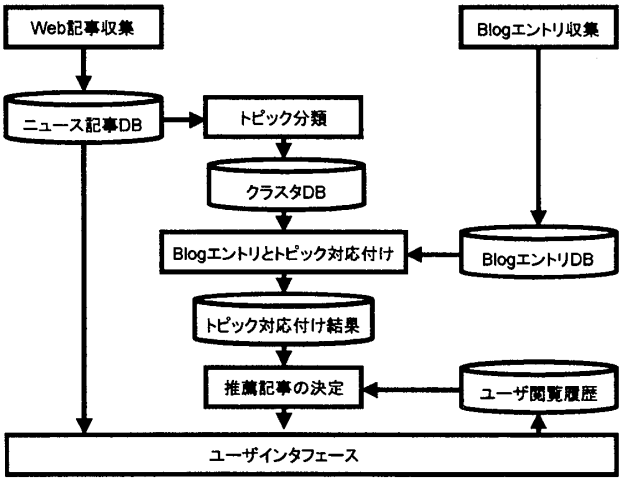

図 1. Web 情報推薦システムの概要

4. 分類クラスタのクラスタリング方式の改良

3 章の手法を用いたクラスタリング方式では、それぞれの Web 記事が適切なトピックの分類クラスタに追加されるとは限らない。そこで、Web 情報推薦システムにおいて、推薦精度を上げるために、生成された分類クラスタのクラスタリング方式を改良する手法を提案する。それを実現するための 2 つのアプローチを以下に示す。

4.1 分類クラスタの結合

3 章の手法で Web 記事を分類するには、その記事と全ての分類クラスタの間で類似度を測る。類似度がしきい値以上であるとき、その記事はクラスタに追加される。追加後、記事が追加されたクラスタが複数存在する場合、それらのクラスタ同士の類似度を測る。閾値以上の場合、それらのクラスタは同一のトピックについて述べていると考えられるので、結合することができる。これを図 2 に示す。

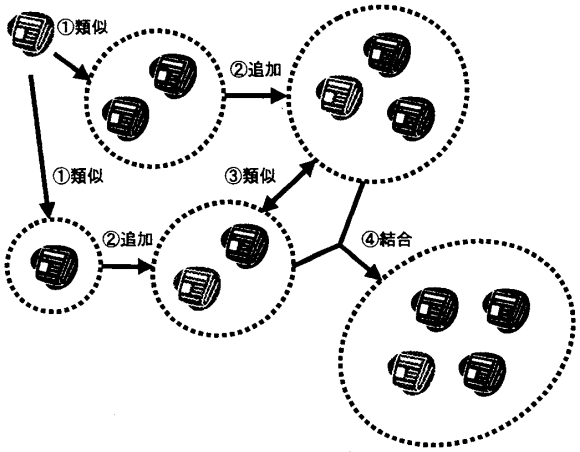

図 2. 分類クラスタの結合

4.2 分類クラスタの分割

(1) 分割するべき分類クラスタの決定

生成された分類クラスタ群において、Web 記事を特に多く含んでいるクラスタを分割の対象とする。単に記事数が多いだけでなく、クラスタが生成されてから最後に記事が追加されるまでの期間が長いものを対象とする。これは、時間の経過によりトピックが移り変わっている可能性が高いからである。

(2) 分類クラスタの分割

分割するべきクラスタ内の全ての Web 記事内容の類似度をコサイン距離によって計測し、閾値以上の Web 記事同士でクラスタを新たに生成する。この際、階層型クラスタリング手法で行う。階層化したクラスタの利用法として、協調フィルタリングの際にクラスタ内の対象とする記事が少ない場合に 1 つ上の階層のクラスタで処理を行うことや、推薦記事をユーザに提供する際にそれと関連する記事も推薦することが挙げられる。クラスタの階層化の例を図 3 に示す。この場合、a、b、c の 3 つのクラスタに分割される。

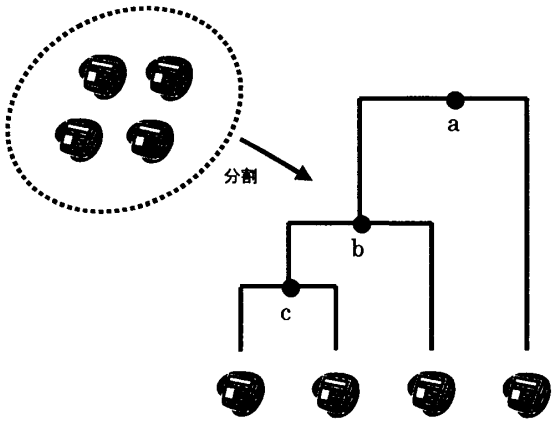

図 3. 分類クラスタの階層化

5. おわりに

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 情報推薦システムのためのクラスタリング方式の改良について検討した。この手法により、Web 記事をトピック別に分類する際、適切な大きさの分類クラスタを生成することで、情報推薦システムの推薦精度向上が実現できる。

今後は、実際にこの手法を採用した Web 情報推薦システムの評価・実験を行うとともに、Blogger の嗜好がどの程度反映されているか、また、ユーザによりよい情報を提供するためには、どのように推薦を行うべきなのかといった調査・検討を進めていく予定である。

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] 寺田道生，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と内容類似性による Web 情報推薦システム，FIT2006，E-007，2006

[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