E-004

トピッククラスタを利用した協調フィルタリングに基づく Web 情報推薦システム

Web Content Recommendation System based on Collaborative Filtering and Topic Clustering

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

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

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

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，様々なジャンルの膨大な数のニュース（以下 Web 記事）が配信されている．

これらの推薦手法として，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し，寺田[2]は小原のシステムに Web 記事のトピック分類を組み合わせることを提案している．トピック分類は，内容類似性によるクラスタリングによって実現を目指している．本研究では，トピック分類と協調フィルタリングを利用した情報推薦システムについて検討する．

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（Blog 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 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 つの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問題点

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．

3．Web 記事のトピック分類

3.1 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．このトピック分類で行う処理を以下に示す．

(1) Web 記事のベクトル化

初めに，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しているため，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または改行までを抽出範囲とする．次に，抽出した語に対して TF・IDF 法で重み付けを行い，タイトルに関しては更に 2 倍の重み付けを行う．最後に，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．

(2) トピッククラスタの構成

トピッククラスタは，類似度が閾値以上の Web 記事集合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(1)と同様の処理を行い，重みの値の上位最大 50 語をそのクラスタの特徴語群とする．

(3) Web 記事のクラスタリング

クラスタリングには 1 パス法を利用し，新たに収集された Web 記事と既存のトピッククラスタ(2)との類似度をそれぞれの特徴語群のコサイン距離により計測し，閾値以上ならば Web 記事をクラスタに追加する．

このとき，複数のクラスタに追加される場合がある．どのクラスタにも追加されなかった場合は，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(2)で説明したものを適用する．

3.2 トピック分類の問題点

1 つの Web 記事が複数のクラスタに追加された場合，それら複数のクラスタのいくつかを 1 つに結合することが必要な場合がある．また，トピック分類や結合により大きくなり過ぎたクラスタは，分割する必要がある．

4. トピッククラスタの結合と分割

4.1 トピッククラスタの結合

(1) 対象となるクラスタ群の決定

収集されたばかりの新しい Web 記事が複数のトピッククラスタに分類された場合，それらのクラスタは類似したトピックである可能性が高いと言えるので，それらを結合の候補とする．

(2) クラスタ間の類似度計算

まず，候補となったクラスタ群より，2 つのクラスタの組み合わせをすべて求める．次に，クラスタ同士の特徴語群のコサイン距離により，類似度を計測する．このとき，閾値以上の全ての組み合わせを結合の対象とするのではなく，その中で類似度が最も高いものを対象とする．この類似度が，閾値以上の場合はクラスタの結合を行い，閾値に満たない場合は，ここで処理を終了する．

(3) 特徴語群の生成

結合されたクラスタ内の Web 記事集合に 3.1 節(1)と同様の処理を行い，重みの値の上位最大 50 語をそのクラスタの特徴語群とする．この(3)までの処理の例を図 1 に示す．

(4) 繰り返し

結合すると，対象となるクラスタ群が変化するので，クラスタがまだ複数ある場合には，(2) から再度行う．

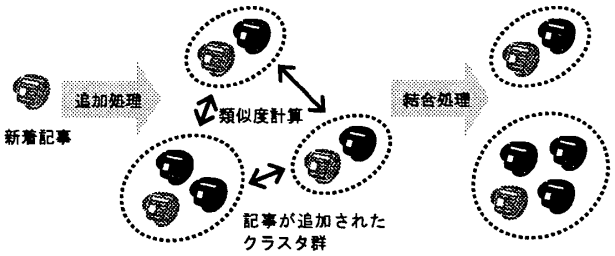

図 1　トピッククラスタの結合例

4.2 トピッククラスタの分割

トピッククラスタの分割は，分割対象のクラスタに属する記事群を新着記事群と同様に扱い，3.1 節(3)の方法でクラスタ群を生成することにより実現している．ただし，本項での閾値は，3.1 節(3)での閾値より高くしている．

(1) 対象となるクラスタの決定

生成されたクラスタ群の中で，Web 記事数が多く，クラスタの生成日から最終更新日までの期間が長いものを分割の対象とする．

(2) クラスタ内の記事の再クラスタリング

まず，分割対象のクラスタである Web 記事の集合を記事単位に分割し，その内の 1 つの記事から新規クラスタを生成する．次に，残りの記事の内の 1 記事と，既に生成されたクラスタ群との類似度を求め，閾値以上ならその記事をクラスタに追加する．閾値に満たない場合は，その記事を要素とする新規クラスタを生成する．この(2)までの処理の例を図 2 に示す．

(3) 特徴語群の生成

(2)で生成されたクラスタ内の Web 記事集合に 3.1 節(1)と同様の処理を行い，重みの値の上位最大 50 語をそのクラスタの特徴語群とする．

(4) 繰り返し

元となったクラスタ内の残りの記事について，同様の処理を(2)から行う．

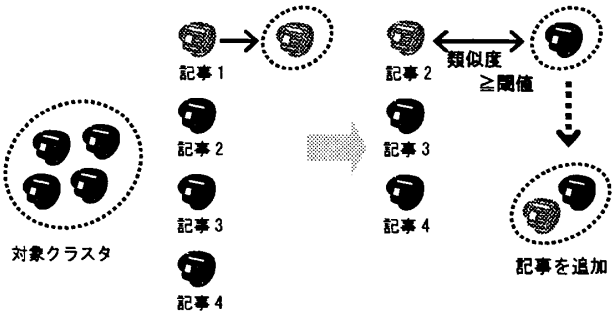

図 2　トピッククラスタの分割例

5. Web 情報推薦システム

2 章，3 章，4 章の手法を組み合わせた Web 情報推薦システムの処理の流れを図 3 に示す．

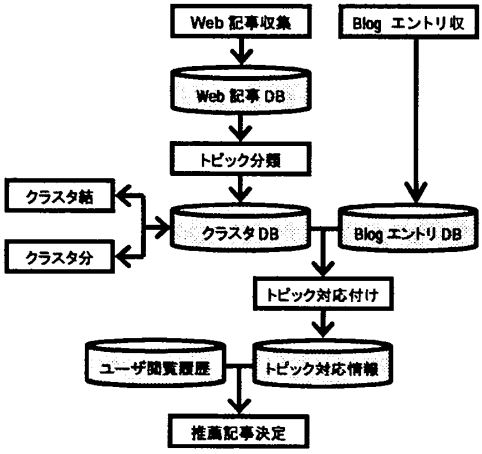

図 3　Web 情報推薦システムの流れ

6. おわりに

クラスタの分割と結合の処理を組み込むことにより，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ができる．結果的に，情報推薦システムの推薦精度の向上を実現することができると考えられる．今後は，推薦精度を測る評価実験を行う．

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