D-027

小規模サイトにおける情報推薦を目的としたデータ統合手法
Data Integration Method for Recommendation on Small Website

清水 伸晋† 奥野 拓‡
Nobuyuki Shimizu Taku Okuno

1. はじめに

情報過多に対応するために，情報推薦システムが注目されている．これは利用者の好むコンテンツを予測して利用者に提示するシステムである．実際 Amazon.co.jp[1]における「おすすめ商品」のように， Web 上での利用が増加している．しかし，情報推薦システムを導入できているのは，Amazon.co.jp やアスクル[2]のような大規模サイトがほとんどであり，小規模サイトでは情報推薦を利用しにくいという問題がある．

小規模サイトを運用する際，以下の 2 つの問題が考えられる．

1. 運用の予算が少ないこと
2. サイトの利用者が少ないこと

Web サイトの運用では，商品紹介ページやブログ記事の掲載といったコンテンツの追加やバックアップなど様々な作業が必要である．ただでさえ予算の少ない小規模サイトにとって，情報推薦システムの導入，維持のコストが大きいことが導入の障害となっている．また，サイトの利用者が少ないことも問題となる．有効な推薦，つまり推薦したコンテンツを利用者がよく受け入れるような推薦を行うためには，利用者のふるまい情報を大量に蓄積する必要がある．ここでいうふるまい情報とは，利用者の閲覧記録や購入履歴といったデータを指す．サイトの利用者が少ないと，このふるまい情報が不足し有効な推薦とならない．

こうした問題をふまえ，本研究では小規模サイトにおいて，情報推薦の有効な利用を実現することを目的とする．

2. 協調フィルタリングの問題点

情報推薦の代表的な手法として，協調フィルタリングがある．協調フィルタリングは，利用者の既知のコンテンツを推薦しにくいために，利用者の満足度が高いという点で他の手法よりも優れている．したがって本研究では，この協調フィルタリングに着目した．

協調フィルタリングでは小規模サイトのようにふるまい情報が不足すると，コンテンツが推薦対象とならないという問題や類似利用者を探せないという問題が発生する．そのため一般に小規模サイトでは協調フィルタリングを利用することが難しい．この問題への対処として，複数のサイト間で情報を共有して不足を補い合うということが考えられる．

しかし情報を共有すると別の問題が発生する．コンテンツの情報，利用者のふるまい情報を共有することでそれぞれのサイトの持つ特徴が損なわれてしまう．小規模サイトには扱うコンテンツが特定の種類に特化しているなど運用方針やユーザ層に特徴がある．この特徴は広いジャンルのコンテンツを扱い，ユーザ層の広い大規模サイトにはない強みである．したがって情報を共有する場合は，それぞれのサイトの特徴を維持して情報を共有し，推薦が行われるべきである．

本研究ではデータの不足する小規模サイト同士がデータを共有する際に，データ共有によって各サイトの特徴が損なわれてしまうという問題を解決する．問題解決のアプローチとして，コンテンツ管理者がコンテンツに付加する情報を利用する．コンテンツ管理者とはショッピングサイトにおいてアイテムの紹介ページを作成するような役割を担う人員を意味する．

3. データの統合

ふるまい情報をサイト間で統合するということは，それぞれのサイトで扱うコンテンツを関連づけて，関連づけられたコンテンツ同士のふるまいデータを共有するということである．ここでコンテンツを関連づける手法として以下の 2 つが考えられる．

1. 同一コンテンツの関連づけ
2. 類似コンテンツの関連づけ

同一のコンテンツであると識別できる属性があれば同一コンテンツ同士の関連づけは可能である．例えば，書籍における ISBN はこの一意に識別できる属性である．しかし，この手法は適用できるコンテンツが限られるという問題がある．一方，類似コンテンツを関連づけるという手法がある．これはコンテンツに付加された属性からコンテンツ間の類似度を計算して，類似度が一定値以上である組み合わせを関連づけるというものである．属性を付加するための管理コストが必要であるが，どのような情報を付加するのかによってサイトの特徴を表現できる．したがって，特徴を維持してふるまい情報を統合するためには，類似コンテンツを関連づける手法が適している．

類似コンテンツを決定する具体的な手法として，それぞれのコンテンツに数種類の属性を付加し，それら類似度を計算し関連づけるコンテンツを決定するという手法がある．また単一のサイト内のみでの運用を想定したものであるが，類似コンテンツを関連づけてふるまい情報を補うことを試みた先行研究もある．Hu[3]らは，映画に対して説明文を付加し，その説明文の類似度によって関連づける映画を決定するという手法を提案している．

これらの手法を特徴の現れやすさと運用コストという 2 つの観点で比較する．説明文を用いた手法では，サイトの運用方針が結果に反映されやすいと言える．しかし一方で，コンテンツに説明文を付加することによって，非常に大きな管理コストが生じる．属性から類似度を計算する場合では，属性の中に大きさ，色といった物理的な属性が含まれることが多く，コンテンツ管理者の特徴は現れにくい．また，管理コストも大きい．

†公立はこだて未来大学 システム情報科学研究科
‡公立はこだて未来大学 システム情報科学部

4. 提案手法

小規模サイトを対象とした場合，コストを抑えつつ特徴を維持することが必要である．そこでコンテンツ管理者がコンテンツに対して抱いた印象を単語で表現し，この主観的な属性のみを利用する．以後この付加情報を主観的概念情報と呼ぶ．この主観的概念情報はサイトの運用方針に依存する．したがって，主観的概念情報を類似コンテンツの判別に利用することで特徴を表現できると考えられる(表 1)．

表 1　類似度計算に用いる付加情報の比較

付加する情報	コスト	特徴
説明文	×	△
付加情報	×	○
主観的概念情報のみ	△	○

提案手法によるふるまい情報統合のフローを説明する．まず，各サイトの持つふるまい情報と主観的概念情報を集約する．次に各サイトのふるまい情報を結合する．しかしこの時点では類似コンテンツが関連づけられていないので，他サイトの情報は利用できない．その後，入力された主観的概念情報の類似度を計算し，類似コンテンツを決定する．最後に，類似コンテンツ間でふるまい情報の統合を行うことで，協調フィルタリングを適用できるデータとなる．

5. 実験

提案手法では類似コンテンツの判別を主観のみで決定するため，精度が向上しないケースが考えられる．そこで提案手法を用いて，データを共有することが安定して推薦精度の向上につながるか検証する必要がある．また，コンテンツ管理者の主観の違い，つまりサイトの運用方針の違いが推薦されるアイテムに影響するか検証する必要がある．

これらの点を評価するために，公開されているデータセットを利用して実験を行った．このデータセットは，ふるまい情報として 5000 人の 100 種類のアイテムに対する評価を記録したものである[4]．

実験では，このデータセットから一部分を取り出した 2 つの小規模サイトを想定した．次に 16 人の被験者に，100 種類の各アイテムに対して 1 つだけ主観的概念情報を付加してもらい，16 個の主観リストを作成した．この主観的概念情報と想定した小規模サイトを組み合わせることで，提案手法で前提とする条件を満たす環境を構築した(図 1)．小規模サイトに組み合わせる主観リストを変えることで，運用方針の異なるサイト間での統合を想定できる．それぞれの状況に対して提案手法と協調フィルタリングを適用し，推薦されるアイテムと推薦の精度を記録した．精度は推薦の正解率を交差確認法を用いて計算した．また，各組み合わせで各利用者に推薦されたアイテムの頻度を測定し，ここから組み合わせごとに多くの利用者に推薦されたアイテムを算出し組み合わせごとにリスト化した．

実行した組み合わせの 9 割程度で推薦の精度が向上した．この結果から，1 つだけの主観的概念情報を用いた場合でも，ある程度安定した推薦を行えると言える．精度の低下した組み合わせでは，被験者が付加した主観的概念情報が同一であるアイテムが多いことわかった．

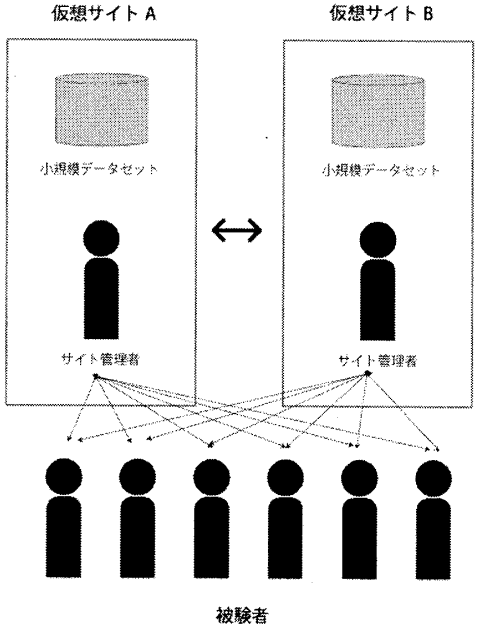

図 1　仮想的な小規模サイト

また，頻度リストの上位アイテムを各組み合わせ間で比較したところ，リストに選択されているアイテムやアイテムの順位などが異なることが確認された．

6. 実験の考察と今後の方針

被験者が付加した主観的概念情報が同一であるアイテムが多いと，それぞれの閾値で類似するとみなされるアイテムがほかの組み合わせよりも多い．これが補完を行う際のノイズとなり精度が低下したと考えられる．これは 1 つのコンテンツに付加する主観的概念情報を複数個に増やすことである程度解決できると考えられる．しかし，付加する主観的概念情報が増えることで，管理コストが増えることや追加で付加した情報がノイズとなって主観が適切に表現されなくなるということが想定される．複数の主観的概念情報を付加したケースで実験を行う必要がある．

頻度リストの比較から，主観リストを作成した被験者の違いによって推薦されるアイテムが変わっていることがわかる．これはサイトの運用方針が推薦結果に影響しており，提案手法を用いた場合でも特徴が反映された推薦となっていると言える．

以上の考察より，提案手法を用いることで，各サイトの特徴を維持して情報推薦を有効に実行するために情報を共有できると考えられる．しかし今回の実験では，提案手法を用いた協調フィルタリングが実際に運用した際にも，利用者に受け入れられるかについての評価と言い難い．今後アンケートなどによる印象評価を行いたい．

参考文献

[1] Amazon.co.jp, http://www.amazon.co.jp/.
[2] アスクル, http://www.askul.co.jp/.
[3] Biyun Hu, Yiming Zhou, "Content Semantic Similarity Boosted Collaborative Filtering," cis, vol. 2, pp.7-11(2008).
[4] T. Kamishima, "Nantonac Collaborative Filtering: Recommendation Based on Order Responses", KDD2003, pp.583-588(2003).