情報推薦のためのブログの活用法に関する研究

青木　志門†　　白井　靖人‡

静岡大学大学院情報学研究科†　　静岡大学情報学部情報科学科‡

1. はじめに

現在、オンライン通販サイト等で、推薦システムがよく利用されている。推薦システムとは、ユーザにとって有用と思われる対象、情報、または商品などを選び出し、それらをユーザの目的に合わせた形で提示するシステムである[1]。

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(collaborative filtering)と、内容ベースフィルタリング(content-based filtering)が知られている。それぞれの方法には利点と欠点があり、特に協調フィルタリングには、“コールドスタート”という問題があることが指摘されている。

本研究では、このコールドスタート問題を解決するために、ブログを利用した情報推薦の有効性を検証する。

2. 協調フィルタリング

協調フィルタリングとは、推薦システムを利用するユーザに、そのユーザと類似した嗜好パターンを持つ別のユーザが高く評価した対象を推薦するという手法である。

協調フィルタリングを用いて精度の高い推薦を行うためには、多くのユーザの、様々な対象に対する嗜好データを事前に収集しておく必要がある。システムの運用初期段階では、嗜好パターンの情報が不足することから、有効な推薦が期待できないことが指摘されており、これをコールドスタート問題と呼ぶ。

3. 関連研究

情報推薦にブログを利用している例として、小原らの研究が挙げられる[2]。各ブログの作成者であるブロガーを、協調フィルタリングにおける仮想的なユーザとして用いる方式を提案している。推薦対象をウェブ上のニュース記事とし、ブログ内にニュースソースへのリンクがあった場合、ブロガーがそのニュースに興味を持っていると判定している。

Using Blogs as a Basis for Information Recommendation
†Shimon Aoki, Graduate School of Informatics, Shizuoka University
‡Yasuto Shirai, Faculty of Informatics, Shizuoka University

4. 情報推薦へのブログの利用

本研究では、3章で述べた先行研究を踏まえ、ウェブ上のニュース記事以外の推薦対象においても、ブログを用いた情報推薦を行う手法について検証する。

総務省の調査[3]によると、国内のブログ数は毎年増加傾向にある。アクティブなブログ(月一度以上の更新がある)は300万にも上り、ウェブ上でも特に大きな情報源の一つとなっており、利用価値が高い。

本研究では、書評や書籍のレビューを行っているブログ(以下、書評ブログ)に注目し、ブロガーが興味や関心を持った書籍や、著者に関するデータを収集した。こうして収集した情報を基に、協調フィルタリングによって情報推薦を行う。

書評ブログのエントリの多くは図1のような構成を持っている。一つのエントリにつき、一つの書籍が紹介されていることが多い。ブログのエントリを収集、解析することで、そのブロガーがそれまでに読んだ書籍、つまり興味を持った書籍のデータを取得することが可能である。理想的には、ブログ本文中のテキストを分析し、ブロガーの書籍に対する評価の程度を数値化することが望ましいが、本研究においては、ブログで紹介したか否かの2値で評価する。

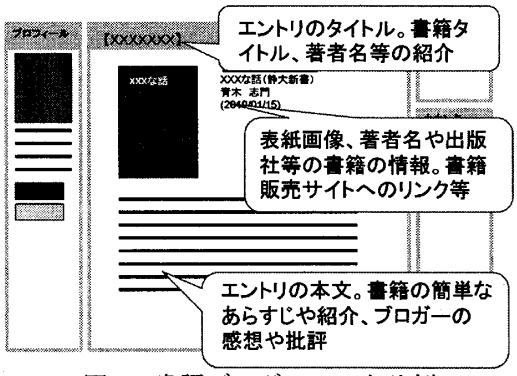

図1　書評ブログのエントリ例

5. ブログの収集と分析

収集の対象となるブログを書評ブログのランキングサイト[4]から選定し、ブログエントリの収集を行った。抽出項目は、エントリ毎のタイトル(書籍タイトル)、著者名、本文、およびURLである。

収集したデータの概要を表1に示す。

表1　収集データ

項目		件数
ブログ数(仮想ユーザ数)	(a)	80
エントリ数	(b)	10310
平均エントリ数	(b/a)	128.9
紹介書籍数	(c)	4994
平均紹介書籍数	(c/a)	62.4
紹介著者数	(d)	3402
平均紹介著者数	(d/a)	42.5

協調フィルタリングにおいては、ユーザ間の類似度を算出するので、ユーザ間において紹介している要素にどれだけの重なりがあるかが重要である。書籍タイトルと著者の二つの要素について、同一要素の出現するブログ数の分布を調査した。

どちらの要素も一つのブログにだけ出現するものが圧倒的に多く、出現するブログ数が増えるにつれて該当する要素数は急激に減少する。特に、複数のブログに登場する書籍タイトルが少なく、書籍タイトルだけに着目するとユーザ間の類似度が見出しにくい。著者名に関しても同様の傾向が見られるが、書籍タイトルよりはユーザ間の重なりが大きいといえる。

そこで、本研究では、書籍タイトルだけでなく、それよりも推薦対象に関する抽象度が高く、ユーザ間の重なりが発生しやすい著者名にも着目して、ユーザ間の類似度を評価することにした。二つの観点を用いることによって、ユーザ間の類似性を見出しやすくなると考えられる。

6. 推薦提示

収集したデータを基に、ブログ間の類似度の算出を行い、協調フィルタリングによって推薦対象を決定する。類似度の算出にはコサインベースの類似度計算を行う。この手法では、各ユーザの要素への評価をベクトル化し、ベクトルの内積をとることで類似度を得る。

今回収集した80のブログ(b001～b080とする)のうち、ブログ b001(エントリ数 168、紹介書籍数 153、紹介著者数 94)との類似度が高いブログ 4 件を求めた結果を表2に示す。表2の左半分は書籍タイトルに着目した類似度、右半分は書社名に注目した類似度に基づく結果である。

表2　ユーザ間の類似度の一例

書籍タイトル		著者名	
ブログ	類似度	ブログ	類似度
b009	0.110	b009	0.158
b006	0.095	b006	0.124
b032	0.053	b050	0.094
b050	0.041	b017	0.071

こうして得られた結果を基に、推薦対象を求める。

書籍タイトルを用いた推薦プロセスでは、ブログ b001 と類似度の高いブログで紹介されている書籍タイトルのうち、ブログ b001 で紹介されていない書籍タイトルを抽出し、その中での紹介回数の多い順にリストアップし、提示する。

著者名を用いた推薦プロセスでは、ブログ b001 と類似度の高いブログで紹介されている著者名のうち、ブログ b001 で紹介されていない著者名を抽出し、その中で紹介回数の多い著者名をリストアップし、その著作を提示する。

7. 考察

本研究では、書籍という推薦対象を扱うにあたり、書籍タイトルと著者名という二つの属性に着目した2種類の類似度を算出し、協調フィルタリングを行った。ユーザ間の重なりが希薄な状況でも、より抽象度の高い要素に着目することでユーザ間の類似度を見出すことが可能になった。

また、本研究では情報のリソースとして、ブログを利用しているが、コメント数やトラックバック数といったブログ特有の機能を考慮することでより質の高い推薦ができるのではないかと考えられ、今後検討が必要である。

8. まとめ

本研究では、ブログの作成者を仮想ユーザとみなし、書籍推薦を行う手法について提案した。情報源としてブログを利用することで、システムを利用したユーザのデータがない状態でも推薦を行うことが可能であるという結果が得られた。また、書籍タイトルと著者名との二つの要素を用いて、ユーザ間の類似度の算出を行った。

9. 参考文献

[1] 神嶌敏弘 "推薦システムのアルゴリズム(1)～(3)" 人工知能学会誌, vol.22,No.6～vol.23,No.2

[2] 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" 2005 年度人工知能学会全国大会、2C2-02

[3] 総務省：ブログの実態に関する調査研究の結果 ,http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2008/2008-1-02-2.pdf

[4] にほんブログ村：書評・レビュー人気ランキング、http://book.blogmura.com/bookreview/