

口コミ型情報推薦システムのシミュレーションによる解析 Simulational Analysis of a Word-of-Mouth Recommendation System

保科 一明[†] 武田 利浩[†] 平中 幸雄[†]
Kazuaki Hoshina Toshihiro Taketa Yukio Hiranaka

1. はじめに

インターネットの普及が進み、誰でも様々な情報が得られるようになった。その一方で、膨大な情報の中から目的のものを見つけられない「情報過多」と呼ばれる問題が起きている。その解決のために、検索エンジンが開発され、普及したが、検索キーワードをどこから入手するのか、という問題が残っている。そのため、ユーザの嗜好に合わせた情報を推薦する「情報推薦技術」への期待が高まっている。

著者らはこれまで、各ユーザが興味・関心のある情報や、検索エンジンに利用するキーワードの手軽な入手を目的として、口コミ型情報推薦システムの提案を行った[1]。本稿では、提案システムについて簡単に解説し、シミュレーションによる評価実験と解析の結果を報告する。

2. 口コミ型情報推薦システム

図1に口コミ型情報推薦システムの概要を示す。提案システムでは、各ユーザが自分のプロフィールと知人のプロフィールを持つ。自分のプロフィールは、ユーザの興味・関心のあるタグを登録することで作成する。ここで登録されたタグを「有効タグ」と呼ぶ。知人のプロフィールは情報交換により追加・更新される。また、推薦情報にその内容を表すタグを付加する。推薦の際には、各プロフィールと情報に付いているタグをマッチングすることでフィルタリングを行う。フィルタリングの結果を推薦情報と一緒に送ることで、推薦と同時にプロフィールの交換を行う。ここで送信されたAが持つBのプロフィールに誤りがあった場合、ユーザBは正しい値をユーザAに返信する。

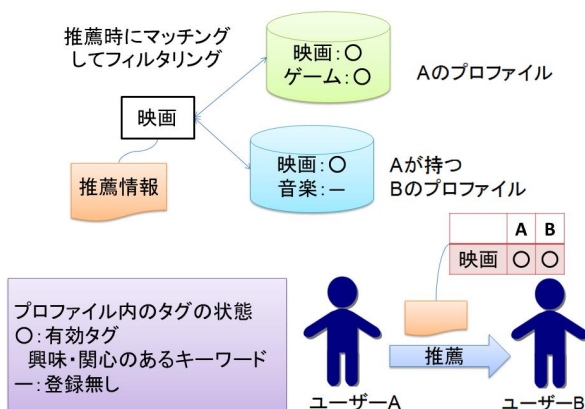


図1 口コミ型情報推薦システム

[†]山形大学 Yamagata University

3. シミュレーションによる評価

3.1 シミュレーションの目的と方法

提案システムの有効性を検証するために、マルチエージェントシミュレーションによる評価を行う。本研究では、Repast Symphony[2]を利用した。

評価には、Novelty な情報の適合率 $P(\text{Novelty})$ と再現率 $R(\text{Novelty})$ や、Serendipity なタグの適合率 $P(\text{Serendipity})$ と再現率 $R(\text{Serendipity})$ を用いる[1]。ここで Novelty とは、ユーザの嗜好に一致し、かつ新規性のある内容のものを指す。また Serendipity とは、ユーザにとって意外性のある内容のことを指し、こういったタグの入手により検索エンジンの利用を補助出来ると考えられる。

シミュレーション上での Novelty な情報とは、ユーザの有効タグが含まれ、新規に推薦された情報とする。Serendipity なタグに関しては、まずユーザの有効タグの一部を、そのユーザが未知なものとして扱う。シミュレーションによりそのタグを推薦出来たとき、Serendipity なタグを推薦したと扱う。

3.2 データセット

シミュレーションに用いるデータセットに必要な項目は、(1)ソーシャルグラフ、(2)各ユーザのプロフィール、(3)推薦される情報である。以下で順に説明する。

- (1) ソーシャルグラフはスケールフリーネットワークモデルで表すことが出来る。本研究では、その生成モデルである BA モデルを利用してソーシャルグラフを作成した。
- (2) プロフィールは、ランダムな英数字文字列を生成し、それをユーザの有効タグとして扱う。ここでソーシャルグラフと各ユーザの嗜好について考えてみると、嗜好の近いユーザ同士が結びついてソーシャルグラフが形成されると考えられる。そこで、あるユーザの有効タグを隣接ユーザにコピーすることで嗜好の共有を再現している。この処理は、ユーザの既知の有効タグと未知の有効タグでそれぞれ別に行う。ここで既知の有効タグの共有率を α 、未知の有効タグの共有率を β とする。
- (3) 推薦される情報は、各ユーザの有効タグをランダムに選択し、組み合わせることで作成する。

4. 実験結果

シミュレーションの条件を表1、結果を表2に示す。ここで表1の n はエージェント数、 t はシミュレーション時間を示す。

結果を見ると、Novelty な情報に関する評価値は共に高くなった。一方 Serendipity なタグに関する評価値は、再現率に比べて適合率が極端に低い値となっている。次章でこの結果の考察を行う。

表1 シミュレーションの条件

		n	t
40%	20%	300	120

表2 シミュレーション結果

P(Novelty)	R(Novelty)	P(Serendipity)	R(Serendipity)
92.45%	98.92%	0.90%	96.25%

表3 各エージェントが受けた推薦数の平均

総受信数	総受信タグ数
411.38	864.18

5. 解析結果・考察

5.1 データセットの解析

実験に用いたデータセットの解析を行った。まず、今回生成されたタグは全部で4134種類あり、その内2141種類が既知のタグとして登録され、1921種類が未知のタグとして登録されていた。これらのタグが各エージェントにきちんと共有されているか確認するために、各タグを何人のエージェントが持っているのか調べた。結果を図2に示す。

図2より、一人しか持っていないタグが2710種類あることが分かる。つまり、全てのタグの6割以上がエージェント間で共有されていないことになる。現実のデータでもこのようなタグが存在することが考えられるが、どの程度発生するのが妥当なのか不明なので、アンケート等の方法で傾向を調べる必要がある。

また、各タグが既知と未知で上手く分かれて分布しているか調べるために、各タグを持つ合計人数に対する既知の所有人数を調べた。これを図3に示す。

図3より、10%未満と90%以上が極端に高くなっていることが分かる。タグの分布と同様に、これもどのような分布になるのが妥当か不明ではあるが、10%以上90%未満のタグがもう少し増えるのが妥当だと考えられる。

5.2 シミュレーション結果の解析

シミュレーションの結果、各エージェントが何件の推薦を受けたのか調査した。結果を表3に示す。表3より、5日間で各エージェントは平均441件の情報と、平均864件のタグを推薦されたことが分かる。5日間という期間で考えると、推薦数が多すぎると言える。P(Serendipity)が低くなったのも、そもそも10個程度しか用意していないSerendipityなタグに対して、多すぎるタグを推薦してしまったためだと考えられる。フィルタリングを再検討するか、ユーザに見せる情報を絞るといった対応が必要である。また、Noveltyな情報の基準が、「有効タグが一件あれば良い」という緩いものなので、推薦された情報のほとんどが嗜好に一致するという結果になっている。しかし、実際は好きなタグが付いている情報でも、内容が嗜好に一致しない場合があると考えられる。そこで、推薦情報の内容を追加し、判定基準を厳しくする等の改良が必要である。

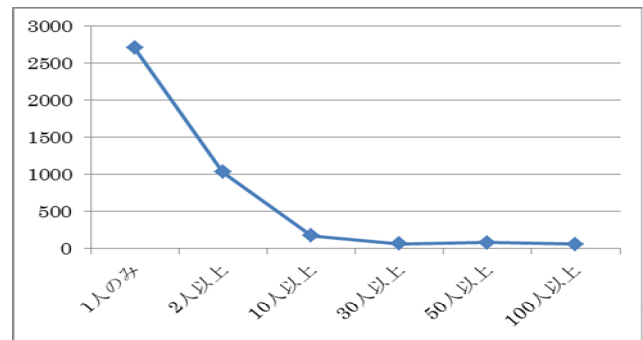


図2 タグの分布

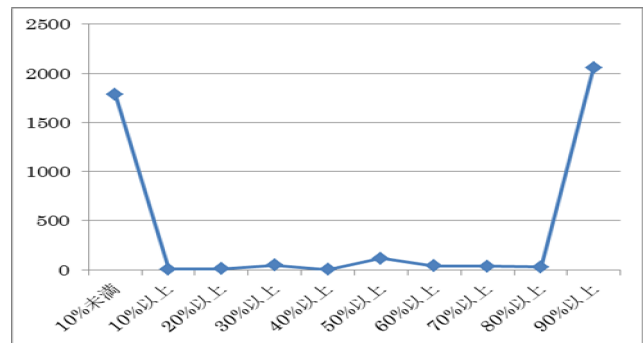


図3 既知タグの割合

6. まとめ

本稿では、提案システムのシミュレーションによる評価実験を行った。また、シミュレーション結果やデータセットを解析した。

提案システムのフィルタリング方法では情報が流れすぎる傾向があり、タグの扱いなどを再検討する必要がある。また、今回利用したデータセットはタグの分布が極端なものになっているが、現実にはどのようなものか調べる必要がある。

今後は、同じデータセットを用いて提案システムと協調フィルタリング[5]等の既存の手法の比較実験を行う予定である。

参考文献

- [1]保科 一明, 吉津 翔太, 武田 利浩, 平中 幸雄, 「ロコミ型情報推薦システムの提案」, FIT2009(第8回情報科学技術フォーラム), 第2分冊 405-406, F007 (2009).
- [2]Repast Agent Simulation Toolkit <http://repast.sourceforge.net/>
- [3]神鷹敏弘, 「協調フィルタリングの課題: プライバシー, サクラ攻撃, 評価値のゆらぎ」, 情報処理, Vol.48, No.9(2007).
- [4]土方嘉徳, 「嗜好抽出と情報推薦技術」, 情報処理, Vol.48, No.9 (2007).
- [5]Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl, 「Item-Based Collaborative Filtering Recommendation Algorithms」, WWW10 (2001).
- [6]吉田匡史, 伊藤雄介, 沼尾正行, 「ロコミによる分散型情報推薦システム WAVE を起こそう -Word-of-mouth-Assisting Virtual Environment」, MACC2001 論文集(2001).