2D3-3

ソーシャルブックマークにおけるイノベータに注目した情報推薦手法の提案
Information Recommendation based on Innovators' Activity for Social Bookmarking Service

大力　慶祐[*1]
Keisuke Dairiki

大向　一輝[*2*3]
Ikki Ohmukai

武田英明[*2*4]
Hideaki Takeda

[*1] 東京大学大学院経済学研究科
Graduate School of Economics, The University of Tokyo

[*2] 国立情報学研究所
National Institute of Informatics

[*3] 総合研究大学院大学
The Graduate University for Advanced Studies

[*4] 東京大学人工物工学センター
Research into Artifacts, Center for Engineering, The University of Tokyo

This paper proposes an effective way to recommend "New URL" for Social Bookmark Services (SBM). In SBM environment, most of collaborative recommendation methods do not work well because the users of SBM are mainly interested in newest articles. Our method focuses behavior of active users in SBM and recommends their latest bookmark to the rest. This method does not require huge computational power to keep updating data set for recommendation and provides an efficient result.

1. 研究背景

近年、情報技術の発達によりインターネット上には多種多様な情報があふれている。しかし、人々は自分の興味のあるコンテンツを探すことが困難になる情報過多の問題にも直面している。

そのため、情報検索の手段として Google や、Yahoo!に代表される「検索システム」の技術が発達し、大規模な計算機資源を用いてインターネット上の情報を集め、情報の信頼性と人気の高いものから順番に検索を行えるようになった。

だが一方で、検索システムが出力する結果は一般的なものが多く、自分の嗜好に沿ったものを探すとなると、融通が利かないものとなる。

そのため検索システムと同様に「推薦システム」の研究が盛んになっている。

本研究はその中でソーシャルブックマーク（以下 SBM）内のデータを用いた情報推薦の手法について調査した。

2. 目的

本研究の目的は、効率よく、逐次変動するユーザの嗜好に即時対応できる推薦である。

ユーザが新しい URL をブックマークすることにより、ユーザの嗜好が変化する。また、古いブックマークは新しいブックマークに比べて、ユーザの興味を引きにくい。そのため、ユーザのニーズを満たすためには、なるべく新しく登録された URL を含めた推薦をすべきである。

しかし、SBM 内に入ってきたばかりの新しい URL を推薦する場合、推薦システムはそのつど最新のデータに更新して再計算をする必要が出る。そこで、更新コストをなるべく少なく「新しい URL」の推薦を望むユーザのニーズを満たすことのできる推薦手法を構築することが、本研究の目的である。

Table1 SBM の登録件数など諸情報

ユーザ数	7,579（人）
URL 数	81,123（件）
SBM の総登録件数	139,602（件）
データ期間	2005/10/25～2006/12/5（407 日）

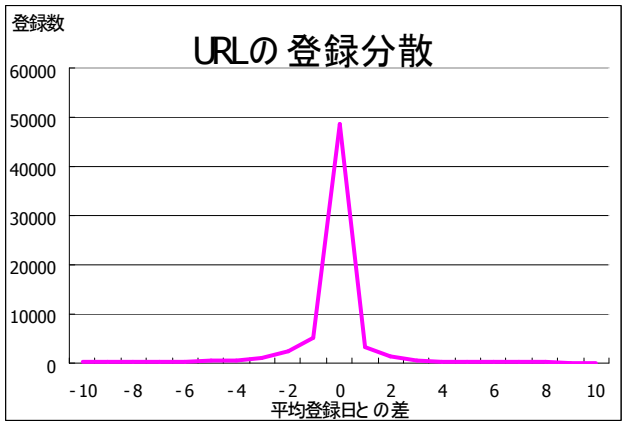

Fig.1 URL の登録分散

3. 研究対象

使用した SBM は Buzzurl（http://buzzurl.jp）の 2005 年 10 月から 2006 年 12 月までのデータである。Buzzurl は、元 EC ナビ人気ニュースというニュース系の SBM から始まったため、ブックマークされているデータはニュース形が大半を占めている。

実際に登録されたデータをドメインごとに登録回数をチェックしたところ、上位 20 件中 16 件のドメインがニュース系のサーバであった。

また、登録データの時間による移り変わりを見たところ、URL の登録は平均登録日のほぼ前後 1 日の間に集中していることが分かった。つまり、Buzzurl に登録される URL には「鮮度」と呼ぶべき基準があり、登録して 2,3 日でほぼ鮮度がなくなりその URL は登録されなくなる、と言える。

連絡先：大力慶祐，東京大学，hinathi@yahoo.co.jp

4. SBM における協調フィルタリングの有効性

この章では、新手法と精度を比較するため、一般の協調フィルタリングの推薦精度を求めた。

4.1 比較実験

新しい推薦手法と比較をするために、2 つの協調フィルタリングによる推薦の精度を Buzzurl のデータを利用して求めた。

1 つ目に、一般的な協調フィルタリングによる推薦の精度を求めた。推薦対象のデータをランダムで半分にわけ、半分を検証データとした。残りのデータを利用し協調フィルタリングにかけ推薦対象に URL を推薦する。その推薦結果が実際に登録されるかどうか調べるために、先ほど分けた検証データとマッチングを行い、精度を出した。

2 つ目は、計算量を減らすため推薦結果を更新せずに推薦を続けた場合の精度を求めた。2006 年 8 月 1 日の時点で推薦の計算を行い、それ以降同じ結果を推薦し続けるとする。計算方法は、Buzzurl のデータを、2006 年 8 月 1 日を基準に前後にわけ、前半データを協調フィルタリングにかけ、各ユーザに URL の推薦を行った。その推薦結果が実際に登録されるかどうか、後半データと推薦データのマッチングを行い、その精度を出した。

4.2 結果

2 種類の推薦結果を Table2,3 に示す。

Table2 一般的な協調フィルタリング推薦の精度

ユーザ数	597
適合数	1842
推薦数	59600
検証 URL 数	27870
精度	3.09%
再現率	6.61%
F1 値	4.21%

今回考案する新しい手法は、計算量をなるべく低く落としたまま、F1 値を 4%程度まで上ることを目標とした。

Table3 前半データだけの協調フィルタリング推薦の精度

ユーザ数	597
適合数	21
推薦数	59,700
検証 URL 数	33,876
精度	0.04%
再現率	0.06%
F1 値	0.04%

協調フィルタリングだと推薦結果の更新を行なわない場合、推薦の精度・再現率は著しく悪い値を取っている。これは推薦した URL の鮮度が時間とともに落ち、ユーザが推薦された URL を登録しなくなったことが原因だと考えられる。つまり、協調フィルタリングは逐次変動するユーザの嗜好に即時対応するのは難しいということがわかった。

5. イノベータによるブックマーク登録

本研究では、全ユーザのデータを計算する代わりに、一部のユーザのみに注目して計算量を減らすことを考えた。そこで、注目したユーザがイノベータである。

イノベータとは、「URL の大部分を、他のユーザより先に登録するユーザ群」と、今回定義した。このようなユーザ群が Buzzurl に存在する場合、そのユーザのみに注目することに以下の 2 つ利点がある。

1. 大部分の URL をイノベータ群が登録するため、推薦漏れがほとんどない。
2. 各 URL は最初に登録されてから 2,3 日で誰にも登録されなくなるので、なるべく早く推薦しないと精度は上がらない。イノベータ群は大部分の URL を最初に登録するので、鮮度の高い URL を推薦することができる。

そこで、Buzzurl にイノベータ群が存在するかどうか求めた。

5.1 実験手法

Buzzurl にこのイノベータ群がいるかどうかを、以下の手順で求めた。

1. SBM の各 URL で、それぞれ最初に登録をしたユーザを求める。

（以降、最初に URL を登録することを、「1 ゲットする」と呼ぶ）

2. 1 で求めたユーザのうち、1 ゲットした URL が多いユーザを求め、順番に並べる。
3. 2 で求めた順位から上位 n 人が SBM の URL 全体の何%を 1 ゲットしたか求める。

以上の手順にて、上位 n 人で SBM の何%の URL を 1 ゲットしているか、その占有率のグラフを Fig2 に示す。

Fig2 より、ほとんど少数のユーザでほとんどの URL を 1 ゲットしていることがわかる。1 ゲットしている回数が 100 回以上のユーザは 83 人で、全ユーザの 2.37%。そのユーザ群が 1 ゲットしている URL は、全 URL の 91.91%となっている。これは十分イノベータの条件を満たしている。

実験 2 の結果を Fig.3 に示す。Fig.3 は 1 ゲット回数の多いユーザから順番に n 人までのユーザで、全 URL の何%を 1 ゲットしているかを表したグラフである。Fig.3 より、少数のユーザによって全 URL の 9 割近くが 1 ゲットされていることがわかる。よって、移り変わりの速い SBM にはイノベータが存在していることがわかった。

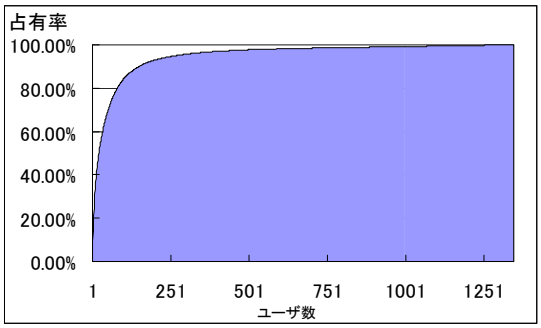

Fig.2 上位 n 人の SBM 占有率

6. イノベータに注目した情報推薦

5.により求めたイノベータ(上位 83 人)に注目することで推薦を行った。

6.1 実験手法

2006 年 8 月 1 日の時点で推薦の計算を行い、それ以降はイノベータの情報だけで再計算を行う。2006 年 7 月 31 日までのデータを利用して、各ユーザと 83 人のイノベータとのコサイン距離を求め、その距離と 2006 年 8 月 1 日以降のイノベータの情報から推薦を行う。

3 種類のイノベータ群を利用した推薦を行い、その精度と再現率を調べた。

手法 1 では、推薦対象ユーザに近い n 人のイノベータの登録した URL の積集合を推薦した。

2006 年 8 月 1 日の時点で計算した、各ユーザと 83 人のイノベータとのコサイン距離から、推薦対象ユーザに最も近い n 人のイノベータを選択し、そのイノベータ達が 2006 年 8 月 1 日以降に登録した URL の積集合を推薦 URL とした。

手法 2 ではイノベータとの類似度が規定値 distance より近いイノベータの登録した URL の積集合を推薦した。

手法 1 と同様に計算した 83 人のイノベータとの距離が規定値より近いイノベータを選択し、そのイノベータ達が 2006 年 8 月 1 日以降に登録した URL の積酒豪を推薦 URL とした。

手法 3 は、イノベータとの類似度を登録 URL に付加し、URL の重さが規定値 weight を超えたものだけを推薦した。

手法 1 と同様に求めたイノベータとの距離を、各イノベータが 2006 年 8 月 1 日以降に登録した URL に重さとして付加させる。二人以上のイノベータが登録した URL は、付加された距離の合計がその URL の重さとなる。付加された重さが規定値を越えたものを推薦 URL とした。

6.2 結果

手法 1 の結果を Table.4 に示す。利用するイノベータの人数 n が増えるに連れて、精度は上がり、再現率は落ちていく。これは、推薦 URL がイノベータの登録した URL の積集合なので、利用する人数が増えるほど推薦 URL が減っていくためである。この手法の F1 値を、最も高い値をとる n=2 のときの 1.00%とする。

Table.4 手法 1 の精度と再現率

n	1	2	3	4	5
ユーザ数	597	597	597	597	597
有効ユーザ数	483	483	483	483	483
適合数	169	25	7	2	0
推薦数	55499	2270	324	57	16
検証 URL 数	2591	2591	2591	2591	2591
精度	0.30%	1.10%	2.16%	3.51%	0.00%
再現率	6.52%	0.96%	0.27%	0.08%	0.00%
F1 値	0.58%	1.00%	0.48%	0.15%	0.00%

手法 2 の結果を Table.5 に示す。手法 1 と違い、distance を増加していくと、精度と再現率も増加していくのがわかる。しかし、有効ユーザ数に注目してみると、有効ユーザ数が減少していくことがわかる。この手法の最高の F1 値は 21.99%と非常に高いが、このときの有効ユーザ数はたったの 9 人である。推薦システムに利用する場合、過半数のユーザに推薦を行いたい。そこで、この推薦手法の F1 値は distance=0.05 のときの 0.65%とする。

Table.5 手法 2 の精度と再現率

distance	0	0.05	0.1	0.2	0.3
ユーザ数	597	597	597	597	597
有効ユーザ数	483	365	92	9	1
適合数	3	69	128	62	0
推薦数	24125	17488	6021	282	0
検証 URL 数	4267	3783	2555	536	0
精度	0.01%	0.39%	2.13%	21.99%	0.00%
再現率	0.07%	1.82%	5.01%	11.57%	0.00%
F1 値	0.02%	0.65%	2.99%	15.16%	0.00%

手法 3 の結果を Table.6 に示す。この手法の精度と再現率の動きは、手法 1 と同様に Weight を増加させると精度は増加し、再現率は減少する。また、有効ユーザ数は手法 1 と同様に一定の値を保っている。そこで、この手法の F1 値を weight=0.5 のときの 8.17%とする。

Table.6 手法 3 の精度と再現率

weight	0.2	0.3	0.4	0.5	0.6
ユーザ数	597	597	597	597	597
有効ユーザ数	483	483	483	483	483
適合数	838	561	396	297	231
推薦数	30932	11765	5547	3005	1771
検証 URL 数	4267	4267	4267	4267	4267
精度	2.71%	4.77%	7.14%	9.88%	13.04%
再現率	19.64%	13.15%	9.28%	6.96%	5.41%
F1 値	4.76%	7.00%	8.07%	8.17%	7.65%

3 つの手法の中で最も F1 値が高いのは手法 3 となった。

7. 結論

イノベータを利用した手法は協調フィルタリングを利用した推薦に比べ 3 つの利点がある。

① 利用するデータが少なくてすむこと。
② 計算量が少なくて済むこと。
③ ユーザ・URL の増加に対する耐久性の高さ。

このように、この手法は推薦の質をなるべく落とさずに、管理側には負荷の少ない手法となっていることがわかった。

今後の課題として、最新のデータに更新して協調フィルタリングの推薦を行った場合にかかる時間を出し、手法 3 と比較を行いたい。また、手法 3 の弱点であった既に登録している URL を推薦してしまうという問題を、毎回推薦する URL が登録されているかどうかをチェックして推薦するようなシステム変更することによって解決し、そのときの精度の変化を比較したい。

参考文献

[神嶌 2007] 神嶌敏弘: 推薦システムのアルゴリズム, 人工知能学会誌 Vol.22.No.6, 人工知能学会 ,2007.

[白土 2002] 白土 慧, 吉井 伸一郎, 古川 正志: ソーシャルブックマークサービスを利用した情報レコメンデーション, 情報処理学会論文誌 Vol.2006. No.84, 情報処理学会 , 2002.

[丹羽 2006] 丹羽智史, 土肥拓生, 本位田真一: Folksonomy マイニングに基づく Web ページ推薦システム, 報処理学会論文誌 Vol.47 No.5, 報処理学会 , 2006.