3G8-2

共有コンテンツのアクセス履歴分析に基づく情報推薦エンジン

Content Portal Recommendation Engine Based on Access Log Analysis

石川 雅之[*1] Masayuki Ishikawa
森田武史[*1] Takeshi Morita
和泉 憲明[*2] Noriaki Izumi
山口 高平[*1] Takahira Yamaguchi

[*1] 慶應義塾大学 Keio University
[*2] 独立行政法人 産業技術総合研究所 National Institute of Advanced Industrial Science and Technology

In this paper, we propose collaborative filtering algorithm which materialize recommendation with novelty to it. Our collaborative filtering method has been evaluated by the simulation with large scale of Intranet log data. This case study shows us that our method tries to solve the Cold Start Problem with conventional collaborative filtering methods.

1. はじめに

情報爆発により、インターネット上に文章、映像、音楽、ソフトウェア等、様々なコンテンツが日々増加している。現在は、自分の求めている情報が他の情報に埋もれてしまい、自分の嗜好にあった情報を探し出すことが困難になる、情報過多の状況にある。その為、膨大な量の情報の中からユーザにとって有用な情報の推薦を行う、推薦システムが注目されている。

推薦システムの実現の為には、ユーザの嗜好を考慮した情報を大量の情報の中から取り出す、情報フィルタリングが必要となる。情報フィルタリングには主に、「内容に基づくフィルタリング」と「協調フィルタリング」の2種類の手法がある。この内、協調フィルタリングは Amazon.com、Tivo 等、多くの商用システムに採用され、有効であるとされている。

協調フィルタリングはユーザAに対する推薦を行う場合、ユーザAと嗜好の類似しているユーザBを探し、ユーザBが高評価したコンテンツの内、ユーザAが体験していないコンテンツを推薦する。この手法は他人の嗜好を利用している為、推薦を行うユーザにとって意外性のあるコンテンツが推薦される可能性がある[神嶌 2006]。しかし、多くの推薦システムでは全てのユーザの嗜好データを使用し、ある程度推薦の精度が高くなってから推薦を行っている為、推薦の精度は上がるが、目新しさがなくなっている。

本稿では目新しさのある推薦を実現する為、新規コンテンツに早くアクセスを行うユーザのアクセス履歴を利用した協調フィルタリング方式を提案する。また、提案する推薦システムの有用性をイントラネット内のアクセスログを使用した実験を行い、評価した。

2. 特定ユーザのアクセスに基づく推薦システム

提案システムでは、主にマーケティングで用いられるイノベーター理論から、ある新規コンテンツに早くアクセスするユーザは、他の新規コンテンツに対しても早くアクセスを行う特徴があり、他のユーザに対してのオピニオン・リーダーとしての機能があると仮定した。そして、これらのユーザのアクセス履歴を利用した協調フィルタリング方式により目新しさのある推薦が可能であると考えた。

2.1 イノベーター理論

イノベーター理論とは1962年にE. M.ロジャースが提唱したイノベーションの普及に関する理論である。ここで、イノベーションとはカラーテレビ、電子レンジ、コンピュータ、インターネットといった技術革新のことであり、イノベーションの採用の頻度を時間の経過に従ってプロットしていくとベルカーブになっている。ロジャースはベルカーブをイノベーションの採用の早い順から、イノベーター(2．5%)、アーリー・アダプター(13．5%)、アーリー・マジョリティ(34%)、レイト・マジョリティ(34%)、ラガード(16%)という5つのタイプに分類した。そして、採用者数を累積すると曲線はS字カーブになることを示し、S字カーブとベルカーブを比較すると、イノベーターとアーリー・アダプターの割合を足した16%のラインとS字カーブが急激に上昇するラインとほぼ一致することから、アーリー・アダプターへの普及がイノベーション普及のポイントであることを見出した。ロジャースはこれを「普及率16%の論理」として提唱している。

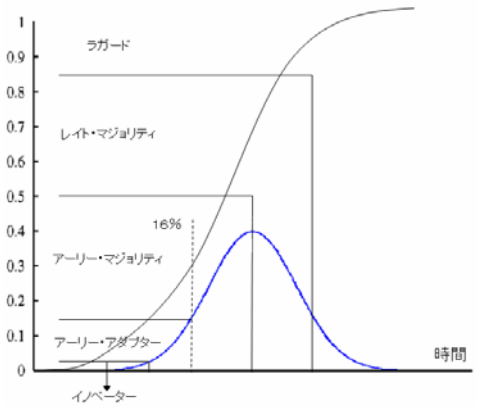

図1 ベルカーブとS字カーブ

2.2 サイバースペース上のイノベーター

ロジャースの提唱した「普及率16%の論理」では、イノベーションの普及が最も急激に進むポイントがイノベーターとアーリー・アダプターの割合を足した16%のラインであることから、アーリー・アダプターが他のユーザにとってのオピニオン・リーダーとしての機能があるとした。ロジャースは、イノベーターの特徴の一つとして、イノベーションを早くに採用しすぎる為信頼度が低いことを挙げている。しかし、イノベーションのサイクルが早いサイバースペース上のコンテンツに関してはむしろ新規コンテンツに早くアクセスを行うユーザであるイノベーターの方がオピニオン・リーダーとしての機能があると考えられる。そこで提案システ

連絡先:石川雅之<sekisen@ae.keio.ac.jp>
〒223－8522神奈川県横浜市港北区日吉3－4－1慶応義塾大学理工学部管理工学科矢上キャンパス24号棟6階619A号室 TEL:045－566－1614

ムではイノベーターのアクセス履歴に基づいて、イノベーター以外のユーザに推薦を行うことで鮮度が高く、且つ予測精度の高い推薦が行えると考えた。

2.3 システム処理手順

システムの処理手順は以下の通りである(図2)。ユーザ1がページAを閲覧したとする。システムはまずページAのプロファイルからページAを過去に訪れたユーザのうちイノベーターであるとされたユーザのIPアドレスを取得する。次に、全ユーザの嗜好DBから推薦候補を列挙する。また、ユーザ1の過去のアクセス履歴を全ユーザの嗜好DBから取得し、ユーザ1のプロファイルを作成する。最後に推薦候補とユーザ1のプロファイルから、ユーザ1がまだアクセスをしたことのないページで、推薦候補に上がっているページをマッチメーカーにより選定し、アクセス数順位が上位に位置しているページが推薦される。

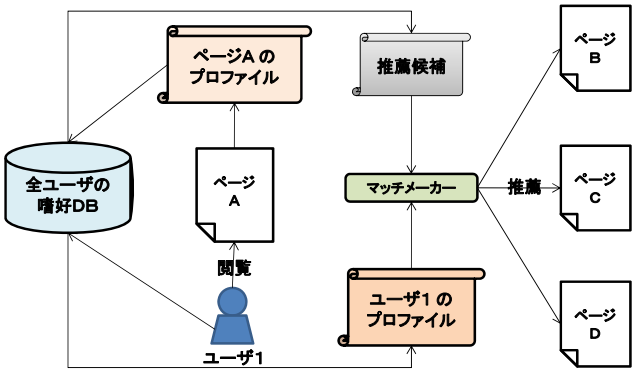

図2 推薦システム概要

(1) 全ユーザの嗜好DB

全ユーザのアクセス時間、URLの情報が蓄積されている。推薦候補、ユーザ1のプロファイルを作成するのに参照する。

(2) ページAのプロファイル

ページAにアクセスしたユーザの内、X番目(イノベーターの数)のユーザまでがイノベーターと定義する。ページAのプロファイルには、イノベーターと定義されたユーザのIPアドレスが記述されている。

(3) ユーザ1のプロファイル

ユーザ1のプロファイルには全ユーザの嗜好DBを参照し、作成されるユーザ1が6ヶ月前までアクセスしたページIDが保持されている。

(4) 推薦候補

ページAにアクセスしたユーザ1に対する推薦候補はページAのプロファイルに記述されているイノベーターがアクセスした時間の1ヶ月前からユーザ1がアクセスした時間までにアクセスした全ページである。

(5) マッチメーカー

推薦候補の内、ユーザがアクセス済みのページをユーザ1のプロファイルを参照し削除し、アクセス数の多い順からY番目(推薦を行う数)までのページを推薦する。アクセス回数が同じだった場合、最終アクセス時間が最新のページを優先する。

3. システム評価方法

提案システムを評価する為、産業技術総合研究所の1年間(2005/4/1～2006/3/31)のイントラネットのログを使用した実験を行った。

3.1 イントラネットのログデータ

イントラネットは経営、運営、会計、研究資料のデータベース、出勤確認、掲示板、ドキュメントのダウンロードに主に使用され、大半のユーザはナレッジワーカーである。全体のログデータは315,005,952レコードあるが、これらは監視ソフトウェアによるログが含まれていたので、それらを取り除いた。フィルタ後、残ったログは126,483,295レコードであった。これらのログデータに含まれる情報はIPアドレス、URL、アクセス時間である。表1にログデータの統計をまとめた[Geczy 06]。

表1 ログデータの統計

Log Records		315,005,952
Clean Log Records		126,483,295
Unique IP Addresses		22,077
Unique URLs		3,015,848
	Scripts	2,855,549
	HTML Documents	35,532
	PDF Documents	33,305
	DOC Documents	4,385
	Others	87,077

3.2 実験方法

1. 2005/5/1～2005/9/30 の期間に30以上のユニークユーザがアクセスしたページを2000ページ、ランダムに抽出した。

2. 抽出したページにアクセスをしたユーザの内、最も早くアクセスしたX人のユーザをイノベーターとした。なお実験ではイノベーターの数Xを1～50人とし、イノベーターの数と推薦の正確性を検討した。

3. イノベーターが抽出したページにアクセスした時間よりも1ヶ月前の期間からイノベーター以外のユーザがアクセスした時間までに、イノベーターがアクセスしたページの統計を取り、推薦候補とした。

4. 推薦候補の内、イノベーター以外のユーザがすでにアクセスしているページを推薦候補から除外し、最もアクセスの多かったページ上位10個を推薦したと仮定した。

5. 最後に、推薦したページにイノベーター以外のユーザが推薦した時間から3ヶ月以内にアクセスしているかどうかを調べた。

4. 実験結果

実験ではユーザへの推薦数を10個と固定し、推薦を行うページに対するイノベーター数を1人から50人まで変化させ、適合率の推移を調べた。適合率とはユーザに推薦を行った10ページへ、ユーザが3ヶ月の期間の間にアクセスを行った割合である。

4.1 イノベーターの数と適合率

図3はイノベーター数と推薦を行った全てのユーザの適合率の平均値の関係をグラフ化したものである。図3より、イノベーターが1人の時、適合率は12.4%であり、イノベーターの人数が増加すると適合率が上昇する傾向にある。しかし、イノベーター

数を21人とした時の適合率21．19%以降はイノベーター数を増加させる毎に適合率は低下し、イノベーター数50人の時の適合率は20．66%であった。これは、イノベーター数を10人とした時の適合率、20．80%よりも低い。このことから、提案システムで得られる適合率は約20%であると言える。しかし、イノベーター数が6人以上であれば、適合率は20%を越えることから、提案システムでは6人程度の少数の人数であっても、推薦を行うページにアクセスすれば、20%程度の適合率は得られることが分かる。このことは、協調フィルタリングの課題である、コールド・スタート問題に対して有効である可能性が高いことを示している。コールド・スタート問題とは、新規ユーザや新規コンテンツのように、ユーザの嗜好データが少ない、アイテムを評価しているユーザが少ない場合、適切に推薦が行われない状態である[Schein 2002]。この内、新規コンテンツの推薦に関してはイノベーターのアクセス数順に推薦を行っている為、今回提案した推薦システムには向かない。しかし、新規ユーザに関しては、そのユーザの嗜好情報が全く無くても、アクセスしたページに10人以上のユニークユーザがアクセスしていれば、適合率20%程度の推薦が可能となり、1つの解決策となりえる。

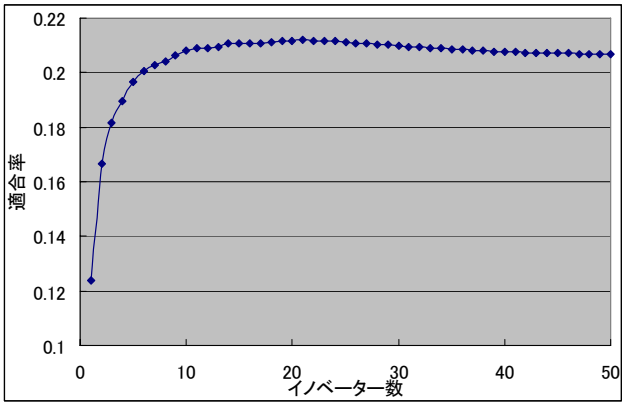

図3 イノベーター数と平均適合率

4.2 イノベーターへの追従度

提案システムでは、オピニオン・リーダーとなるイノベーターにその他のユーザが追従すると仮定している。しかし、イノベーターが10人の時の平均適合率20．80%からはイノベーターのアクセスにその他のユーザが追従しているとは言い難い。そこで、イノベーター数10人の時の全ユーザの適合率の分布と累積分布関数をグラフ化した(図4)。この結果、適合率が0%のユーザが全体の30．44%、適合率10%以下のユーザは52．87%もいることが分かった。これらのユーザがイノベーターのアクセスに追従していないことにより、全体の適合率が低下していると考えられる。これは一般的な協調フィルタリングの手法では、ユーザ同士の嗜好の類似度を複数のコンテンツの評価により求めているのに対し、提案システムでは推薦を行うページにアクセスをした全てのユーザの嗜好は類似していると仮定しているからである。しかし、逆に約47%以上のユーザに対しては適合率20%以上の推薦が可能であり、この場合のユーザの平均適合率は38．15%となる。また、適合率50%以上のユーザも約10%いることからユーザによってはイノベーターに対して高い追従率があり、推薦を行うページにアクセスをした全てのユーザの嗜好は類似しているという仮定は、一部のユーザに取っては有効であることが分かる。

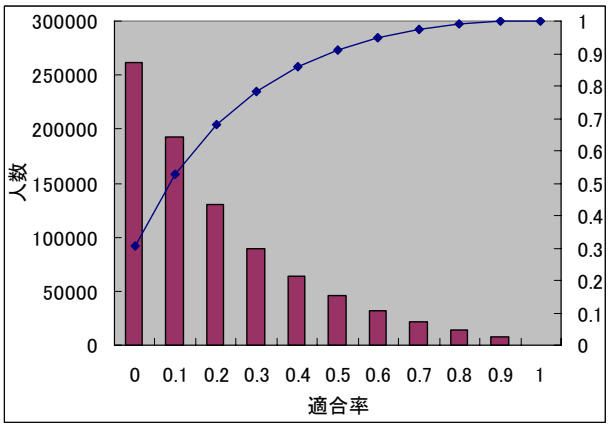

図4 全ユーザの適合率の分布と累積度数分布

5. おわりに

本稿では、目新しさのある推薦を実現する為、イノベーターという新規コンテンツに早くアクセスを行うユーザのアクセス履歴を利用した推薦システムを提案した。また、提案した推薦システムの有用性をイントラネットのアクセスログを使用し、実験を行った。その結果、提案システムで得られる適合率は推薦数10個の場合、約20%であった。しかし、ユニークユーザが6人以上であれば適合率は20%を超えることから、6人程度の少数の人数であっても、推薦を行うページにアクセスすれば、20%程度の適合率は得られ、提案システムがコールド・スタート問題の新規ユーザに対する推薦に有用であることが分かった。次に、全ユーザの適合率の分布から、適合率10%以下のユーザが52．87%存在することが判明した。しかし、逆に約47%以上のユーザに対しては適合率20%以上の推薦が可能であり、適合率50%以上のユーザも約10%いることから、推薦を行うページにアクセスをした全てのユーザの嗜好は類似しているという仮定は、一部のユーザに取っては有効であることが分かった。

今後の課題は、イノベーターに対する追従率の低いユーザに対する推薦の洗練を行うことである。また、本稿ではイノベーターの推薦したコンテンツは他のユーザにとって目新しいということが前提となっているが、この前提が正しいのかをコンテンツの内容から検討する必要がある。

参考文献

[神嶌 2006] 神嶌敏弘:推薦システム―情報過多時代を乗り切る,情報の科学と技術,vol.56,no.10,pp.452-457, 2006

[Sarwar 2001] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J: Item-Based Collaborative Filtering Recommendation Algorithms, In Proc. of the 10th International World Wide Web Conference, Hong Kong, 2001

[Schein 2002] A. Schein, A. Popescul, L. Ungar, D. Pennock: Methods and metrics for cold-start recommendations, 25th Annual ACM SIGIR Conference, pp. 253-260, 2002

[Geczy 06] Extraction and Analysis of Knowledge Worker Activities on Intranet, P. Geczy and S. Akaho and N. Izumi and K. Hasida , Practical Aspects of Knowledge Management (LNAI) pp73--85, Springer-Verlag, Heidelberg, 2006