

研究指導 中澤 真 准教授

閲覧時間を考慮した高精度なリコメンデーションシステムについて

樋口 達大

1. はじめに

近年、インターネットの普及に伴い、オンラインショッピングの利用が増加している。オンラインショッピングでは品揃えを豊富に扱うことができる一方、顧客が自分にとって最適な商品を効率的に探し出せないという課題も生じている。明確な目的を持って商品を検索する場合はキーワード検索などから適切な商品を探し出すことができるが、顧客が目的の商品を正確に把握していなければ商品の検索は難しい。そのため、「ショッピングサイト選択重視点」の調査では、「欲しい商品の検索が容易」、「サイトが使いやすい」という声が多く上げられている[5]。

そこで、顧客一人ひとりの嗜好に合った商品を自動的に推薦するリコメンデーションシステムが注目されている。現在用いられているリコメンデーションシステムの多くは、顧客から嗜好データを獲得する方法として、主に顧客毎の購入履歴や商品評価、閲覧回数などが用いられてきた[4][8][10]。しかし、購入履歴を利用した場合は十分な情報量が得られない、商品評価の場合は顧客に作業負担がかかる、閲覧回数の場合は興味のない情報も含まれてしまうといった問題が生じる。

よって本研究では、嗜好データの獲得において閲覧回数や購入履歴に加えて閲覧時間を用いることにより、閲覧しただけで興味の度合いを測定し、嗜好データとして獲得することでリコメンデーションシステムの推薦精度を高める方法について提案する。

2. リコメンデーションシステムの概要と問題点

2.1. コンテンツベースフィルタリングと協調フィルタリング

リコメンデーションシステムとは、顧客の嗜好を推定し、その顧客に対して適した商品や情報を推薦するシステムである。リコメンデーションを行う方法として、コンテンツベースフィルタリングと協調フィルタリングの二種類の方法がある[6][7]。

前者は、顧客が閲覧した商品と商品特性が類似する商品を探し出し、その商品を推薦するというものである。コンテンツベースフィルタリングには、システムを使い始めたばかりの顧客へ推薦したり、新しくシステムに登録された商品を推薦対象にしたりする状況において、適切な推薦が可能であるという利点がある。一方、対象顧客のみの嗜好データを利用するため、目新しさや意外性

のある商品を推薦できないという欠点がある。また、システム管理者側で推薦を行うための商品特性情報を登録する作業が必要になるという欠点もある。

もう一つの手法である協調フィルタリングは、顧客が閲覧や購入を行った履歴である行動履歴などを基に個々の顧客の嗜好を推測し、同じような嗜好を持つ他の顧客の行動履歴を参考にして商品を推薦する。協調フィルタリングは、顧客の行動履歴から顧客の嗜好や関心を推察することが可能なため、管理者側で商品特性情報を登録する必要がない。また、コンテンツベースフィルタリングでは顧客が予見できる範囲での予測になりがちであるのに対し、協調フィルタリングでは、他の顧客の行動履歴をうまく利用することができるため、目新しさや意外性のある商品を推薦できる可能性がある。さらに、顧客の行動履歴が増加するほど推薦アイテムの適合性が向上するといった利点もある。そこで本研究では、コンテンツベースフィルタリングと比較して長所の多い協調フィルタリングを用いて、リコメンデーションシステムの推薦精度の向上を図る。

2.2. 協調フィルタリングの手順

協調フィルタリングは以下の3つの手順で実行される。

1. 嗜好データの獲得・・・システムは顧客の行動履歴などをもとに嗜好データを獲得する。
2. 類似度の計算・・・顧客の嗜好データが登録されている顧客DBをもとに、顧客と他の顧客の類似度を求める。
3. 推薦商品の選択・・・類似度と他の顧客の嗜好データをもとに、顧客がまだ閲覧していない商品についての嗜好値を予測し、嗜好値が高いものを顧客に推薦する。

このような手順で行われるシステムの多くでは、表1のようなアイテム - 顧客評価マトリクスにより嗜好データを表記する。ここで、顧客の集合を $C = \{c_1, \dots, c_i, \dots, c_{|C|}\}$ 、商品の集合を $G = \{g_1, \dots, g_j, \dots, g_{|G|}\}$ で表し、顧客 c_i の商品 g_j に対する嗜好値は w_{ij} で表す。顧客 c_i の嗜好データは嗜好ベクトル c_i で表し、次のように定義する。

$$c_i = (w_{i1}, w_{i2}, \dots, w_{ij}, \dots, w_{i|G|})$$

表1: アイテム - 顧客評価マトリクス

	g_1	...	g_j	...	$g_{ G }$
c_1	w_{11}		w_{1j}		$w_{1 G }$
$\vdots$					
c_i	w_{i1}		w_{ij}		$w_{i G }$
$\vdots$					
$c_{ C }$	$w_{ C 1}$		$w_{ C j}$		$w_{ C G }$

協調フィルタリングの手順において、 w で表される嗜好データの獲得が正確に行われなければ、その後の類似度の算出、推薦商品の選択も正確に行われない。そのため、推薦の精度を高めるためには、嗜好データの獲得においてより正確に顧客の嗜好を反映させることが最も重要だと思われる。そこで、次に現在用いられている嗜好データの獲得方法について検討する。

2.3. 嗜好データの獲得

顧客の嗜好予測を正確にするには、より多くの正確なデータを収集できた方が有利となる。現在協調フィルタリングで用いられている嗜好データを獲得する方法には、顧客の行動から顧客の嗜好や関心を推察することで嗜好データを得る方法がある。具体的には、顧客の商品購入履歴や商品閲覧回数などを基に顧客の嗜好や関心を推察する方法などが挙げられる。商品購入履歴を用いる場合は、購入の有無により w を 1 か 0 で設定している[4]。購入行動というのは顧客の行動の中でも明確な意思表示であるので、正確に嗜好データを獲得できる。しかし、顧客が実際に商品を購入するのはごく一部であるので、購入履歴のみでは十分なサンプル数を得られないという問題がある。もう一方の商品閲覧回数を用いる場合は、閲覧回数そのものを w の値として設定している[10]。この方法では、閲覧時間にかかわらず1回の閲覧とみなされるため、誤って興味のないページを閲覧した場合や、閲覧したが興味を持たなかった場合でもその商品に興味を持ったとみなされてしまう。

よって、閲覧回数や購入履歴に閲覧時間も加味して嗜好データを獲得する必要があると考えられる。そこで、次節ではこの考えに基づき、リコメンデーションの精度を上げ、それぞれの顧客に適した商品を推薦するための方法を示す。

3. 高精度な新しいリコメンデーションシステムの提案

3.1. 顧客の嗜好の概要と収集データ

本研究では顧客の嗜好を獲得する方法として、商品閲覧回数や購入履歴に加えて閲覧時間も用いる。ここでの閲覧時間とは、顧客がwebサイト上の商品ページを表示していた時間を意味する。この閲覧時間を用いることにより、次のような利点が得られる。

まず、顧客への負担が軽減できるという点である。

アンケートやフォーム入力などにより嗜好データを獲得する場合は顧客への作業負担が生じるが、閲覧時間を用いた場合にはそのような作業を必要としないため、顧客への負担なしで嗜好データを獲得できる。また、一度の閲覧だけではその商品に興味を持ったか正確に判断したり、興味の程度を判断したりすることは不可能だが、閲覧時間も考慮すればそれらを判断することもできる。一般的に、顧客が有用でないページを閲覧した場合は短時間の閲覧となる。このような状況において、閲覧回数を利用した方法では一度閲覧したということに興味をもったと判定されてしまうのに対し、閲覧時間を用いればそれを顧客の興味として反映させないことが可能となる。さらに、閲覧回数では何度か閲覧しなければ興味の度合いを測ることは不可能だが、閲覧時間を用いれば一度の閲覧で顧客の興味の度合いを推測することも可能である。そのほかにも、閲覧行動で嗜好データの獲得を行うため、購入履歴のみを用いるよりも多くのデータを獲得することが可能であるという長所もある。以上のような理由から、閲覧時間を用いればより正確に顧客の嗜好を反映させることができると考えられる。

3.2. 閲覧時間を用いた顧客の嗜好抽出方法

閲覧時間の測定方法は、あるページ P_1 をリクエストした時間と、同じサイト内にある異なるページ P_2 をリクエストしたリクエスト時間の差により、 P_1 の閲覧時間を求めることができる[9]。この方法では、ブラウザのキャッシュが利用された場合など正しく閲覧時間が計測できない場合があるが、本研究では閲覧時間が正しく計測されなかった場合のために、類似度に影響を与えない嗜好値を設定することで対応する。

まず、閲覧時間と顧客の嗜好との関係を明らかにするために、「wwwにおける顧客の興味対象と閲覧時間の関係の調査」の実験結果[1]を用いる。この調査では、まず顧客に興味のあるテーマに関して www で調べてもらい、次に閲覧したページを a. 大変興味があったページ、b. 少し興味があったページ、c. 興味がなかったページに分類し、最後にページ数と興味度合い別の閲覧時間を度数分布で表している。

この実験結果では、顧客の興味の度合いが高くなるに従い閲覧時間も長くなるという相関関係が示されている。また、それぞれの興味の度合い別の平均閲覧時間は約 40 秒、約 15 秒、約 5 秒となっている。

本研究ではこの調査結果を利用し、顧客の興味の度合いを閲覧時間に応じて次のような5段階に区分することとした。

5 秒未満の場合	必要としない
5 秒以上 15 秒未満の場合	どちらでもない
15 秒以上 40 秒未満の場合	少し興味がある
40 秒以上 110 秒未満の場合	興味がある
110 秒以上の場合	測定不能

以下で、このような 5 段階に区分した理由について説明する。まず、閲覧時間が 5 秒以内である場合、調査結果から興味がなかったページの平均値以下であるから、その時点では顧客にとってその商品が必要としない商品であるとした。閲覧時間が 5 秒以上である場合、調査結果から興味がなかったページの閲覧時間の平均値と興味があったページの閲覧時間の平均値の間であるため、顧客はその商品に対して興味があるわけではないが不要でもない、つまり、どちらでもない状態であるとした。閲覧時間が 15 秒以上、40 秒以上の場合、調査結果から少し興味があったページ、大変興味があったページの閲覧時間の平均値以上であるから、それぞれ多少興味がある、興味があるとした。最後に、調査結果では興味があったページの閲覧時間は 110 秒までとなっていることから、この時間を超えた場合は、顧客が商品ページを表示したまま実際には閲覧をしていない状態か、そのショッピングサイト外のページに移動したと考え、正確に閲覧時間が測定できなかった状態とした。この区分を設けることにより、閲覧時間の異常計測時において、閲覧時間が長時間である場合に興味があると反映されるのを防止することができる。

次に、システムは商品のページ閲覧時間が求められた後、次のような嗜好値を設定する。ただし、 w の初期値は *null* として、未閲覧の商品と閲覧済みの商品を区別できるようにする。これにより、商品推薦の際に未閲覧の商品のみを効率的に抽出することが可能となる。

必要としない場合	$w = -1$
どちらでもない場合	$w = 0$
少し興味がある場合	$w = 1$
興味がある場合	$w = 2$
測定不能の場合	$w = 0$

このような値を設定した理由について区分別に説明する。まず、どちらでもない状態であると考えられる場合と測定不能の場合には、顧客の興味があるかないか判断できない状態であるため、類似度の計算に影響を与えないように 0 を設定する。その理由として、このモデルでは顧客間の類似度を測定する際にコサイン類似度を用いるが、コサイン類似度で内積を計算すると、0 の成分は相手の値にかかわらず 0 となるからである。つまり、コサイン類似度では 0 が割り当てられている商品の嗜好値は類似度に影響を与えないことになる。次に、少し興味がある、興味がある場合、顧客にとってその商品が有用であると考えられるため、その度合いに応じて正の値

1, 2 をそれぞれ設定する。なぜなら、少し興味があった平均値が約 15 秒であり、それが 2 回行われれば約 30 秒以上となり、大変興味があった平均値の約 40 秒程度と同等になるため、顧客がある商品について少し興味を持って 2 回程度閲覧すれば顧客はその商品に対して興味を持っていると考えられるからである。最後に、顧客にとってその商品が不要であると考えられる場合、-1 を設定する。なぜなら、嗜好値を正の値のみとすると、コサイン類似度では正の類似度、つまりどれだけ嗜好の傾向が類似しているかということしか測定できないが、負の値も設定すれば、負の類似度、つまりどれだけ嗜好の傾向が反対であるかということも測定することが可能となる。このことにより、推薦商品の選択において、嗜好が反対である顧客が興味を持たなかったものについては、推薦対象の顧客が興味を持つと考えることができる。ただし、誤ってクリックしてしまったなどの操作ミスも考えられるので、負の値は -1 程度とする。

3.3. 嗜好値の更新

ここでは、同じ商品を複数回閲覧した場合の嗜好値の更新について説明する。嗜好値 w は $w' = w + \alpha$ の式によって更新する。ここで α は以下のように設定されるパラメータであり、次のように設定する。

15 秒未満の場合	$\alpha = 0$
15 秒以上 40 秒未満の場合	$\alpha = 1$
40 秒以上 110 秒未満の場合	$\alpha = 2$
110 秒以上の場合	$\alpha = 0$

このように嗜好値の更新について設定したのは、顧客が商品に興味を持った場合、複数回閲覧することが考えられるため、一定時間以上閲覧した際にその嗜好値を加算すれば、閲覧回数を重ねることによる嗜好値の増加を表現することが可能となるためである。ただし、このモデルは購入履歴も嗜好データの対象としているので、閲覧した商品を購入した際には、購入済みと未購入の商品を区別するために嗜好値の更新に上限値を設け、ここでは 6 とした。今回は 1 ページあたりの情報量が多く、十分に理解するのにはある程度時間のかかるような商品を想定しており、そのような商品に対しては 2, 3 回程度閲覧すると考えると、嗜好値が高々 6 であるため、この値を上限値とした。また、閲覧した商品を購入した際には上限値の 6 に 1 を追加して嗜好値を 7 と設定するものとする。これは、閲覧していた商品を購入すれば、何度も時間をかけて閲覧した場合よりもさらに興味があると考えられるので、このような設定とした。一方、2 度目以降の 5 秒以内の閲覧では、商品を比較している場合や、既に閲覧済みで十分理解しているため閲覧時間が短時間になる場合などが大半であると考えられるため、減算は行わないこととした。最後に、5 秒以上 15 秒未満や 110 秒以上の場合であるが、初期値の設定と

同様に類似度の計算に影響を与えないために 0 とする。

3.4. 閲覧時間の補正

顧客の閲覧行動には個人差があり、短時間で必要な情報が判断できる顧客や、時間をかけて判断するような顧客などがいると考えられる。そこで、顧客ごとに異なる基準を設定するために、顧客全体の閲覧時間の平均値と、その顧客の閲覧時間の平均値を元に係数を求め、嗜好の判定基準を補正する。顧客全体の閲覧時間の平均値を u 、顧客 c_i の閲覧時間の平均値を u_i とすると、係数 $= u_i / u$ と定義し、閲覧時間の区切り方は以下のよう

$5 \times (u_i / u)$ 秒未満の場合	必要としない
$5 \times (u_i / u)$ 秒以上の場合	どちらでもない
$15 \times (u_i / u)$ 秒以上の場合	少し興味がある
$40 \times (u_i / u)$ 秒以上の場合	興味がある
$110 \times (u_i / u)$ 秒以上の場合	測定不能

これにより、全体の閲覧時間の平均値よりも閲覧時間が長い顧客は閲覧時間の区切り方における基準値を長くすることができる。逆に、平均値よりも短い顧客は基準値を短くすることができる。例えば、全体の閲覧時間の平均値が 20 秒、顧客 c_i の閲覧時間の平均値が 10 秒であった場合、係数は 0.5 となり、必要としないものであると判断する基準は 2.5 秒未満というように補正される。

3.5. コサイン類似度を用いた類似度の計算方法

顧客の嗜好値 w が定まると、システムはコサイン類似度[2]を用いて、推薦対象となる顧客 c_a と DB 中に登録されている顧客である顧客 c_i の嗜好ベクトルから類似度を求める。顧客 c_a と顧客 c_i の類似度 r_{ai} は以下の式で求められる。

$$r_{ai} = \text{Cos}(\mathbf{c}_a, \mathbf{c}_i) = \frac{\mathbf{c}_a \cdot \mathbf{c}_i}{\|\mathbf{c}_a\| \|\mathbf{c}_i\|}$$

これにより、顧客間の嗜好の傾向がどれだけ類似しているかということや、嗜好の傾向がどれだけ反対であるかということを計算することができる。

3.6. 推薦スコアを用いた推薦方法

最後に、推薦対象の顧客はまだ閲覧していない商品についての推薦スコアを他の顧客の嗜好値と類似度をもとに算出し、このスコアに基づき推薦商品を決定する。顧客 c_a の嗜好値の平均を w_a とすると、商品 g_j に対する顧客 c_a の推薦スコア s_{aj} は

$$s_{aj} = \frac{\sum_{i \in I_j} r_{ai} (w_{aj} - w_a)}{\sum_{i \in I_j} |r_{ai}|}$$

で求められる。ただし、 $I_j = \{i | w_{ij} \neq \text{null}\}$ とする。

この計算により、顧客 c_a のまだ閲覧していない商品に

対する嗜好値が他の顧客の嗜好値と類似度の加重平均によって推定され、システムは推薦スコアの高い商品から順に顧客へ提示する。

4. まとめ

本研究では精度を高め、より顧客の嗜好にあった商品を推薦するリコメンデーションシステムを提案した。このシステムでは、顧客の嗜好データ獲得方法として閲覧時間を利用し、既存のシステムにおける顧客への負担軽減を図るとともに、購入履歴でのデータ量が少ないという問題、閲覧回数での興味が低い情報も含まれてしまうといった問題などを改善することができた。現在、嗜好データの獲得方法として、閲覧中におけるマウス操作、閲覧中の視線を用いるなどの手法が提案されているが[2]、顧客側で特別なプログラムや装置を導入する必要がある。しかし、本研究は特別なプログラム等を必要としないため、実現可能性が高いという利点もある。今後そのような手法が実現できれば、閲覧時間に加えてマウス操作などによる嗜好データ獲得方法を組み合わせることにより、より正確に顧客の嗜好を獲得できるシステムの開発が期待される。

参考文献

- [1]清水勇真, "www における顧客の興味対象と閲覧時間の関係の調査", 情報処理学会第 57 回全国大会, pp.3-191-3-192, 1998
- [2]土方嘉徳, "情報推薦・情報フィルタリングのためのユーザプロファイリング技術", 人工知能学会誌 Vol.19 No.3, pp.365-372, 2004
- [3]神島敏広, "何となく協調フィルタリング - 応答順序に基づく推薦", 人工知能学会研究会資料, pp.1-8, 2004
- [4]高山毅・小熊浩史・池田哲夫・三浦良介, "商品購入の前後関係を考慮したレコメンド方式と評価", 電子情報通信学会, pp.25-30, 2004
- [5]株式会社日本能率協会総合研究所, "IT・ネット社会データ総覧 2005", 2005
- [6]神島敏弘, "推薦システム - 情報過多時代をのりきる", 情報の科学と技術, 56 巻 10 号, pp.452-457, 2006
- [7]石川徹也・宇田隆幸, "情報フィルタリングの利用システム: 情報推薦システム", 情報の科学と技術, 56 巻 10 号, pp.458-463, 2006
- [8]石野雅彦・山木直一・市川照久, "顧客の嗜好に合わせた商品のリコメンデーションシステムの提案", 情報処理学会第 67 回全国大会, pp.4-619-4-622, 2007
- [9]三原宏一郎・寺邊正大・橋本和夫, "ページ閲覧時間を考慮した web アクセスパターンの抽出とその分析", 情報処理学会第 69 回全国大会, pp.1-519-1-530, 2007
- [10]加藤美希, "オンラインモールにおける顧客の負担を軽減したリコメンデーションシステム", 会津大学短期大学部卒業研究論文要旨集, 2007