

協調フィルタリングに基づく情報推薦システムにおける擬似投票方式

宇田 隆幸^{†,††} 藤井 敦^{††} 石川 徹也^{††}

[†] 図書館情報大学 情報メディア研究科

^{††} 筑波大学 図書館情報学系

^{†††} 株式会社ネオジェイエスケー

{udat, fujii, ishikawa}@slis.tsukuba.ac.jp

1. はじめに

ユーザの嗜好に関する情報を用いて新規情報に対する嗜好を予測し、嗜好に合うアイテムを提示するシステムとして、情報推薦システムが研究されてきた[6, 8]. しかし、研究システムや Amazon.com[3] などに代表される電子商取引サイトが中心であり、現実社会におけるユーザの行動範囲に的を絞った地域情報を対象としたシステムは、著者らが知る限りない.

情報推薦システムのアルゴリズムとして協調フィルタリング (以下 CF) が中心に扱われてきた[8]. CF の問題点として、アイテムおよびユーザ数が大規模になるほど、評価値疎ら問題 (sparsity problem) や再生起問題 (recurring startup problem) 等が指摘されている. この 2 点の問題を解消する方式として、筆者らは擬似投票方式を提案した[9]. 擬似投票方式は「あるアイテムを嗜好するユーザは、類似する別のアイテムも同じように嗜好する」という考えに基づいている.

擬似投票方式の最適化を行う目的で、本研究では地域情報アイテムを対象とする. ユーザによって評価されたアイテムへの投票値とシステムによる擬似投票値間の比較を行い、最適な類似度計算方式および類似度判定基準のための閾値について検討する.

実験は、アイテム間の類似度を計算する選択肢を評価対象にする. 特に、擬似投票値の正確性と計算量について比較する. また、類似度判定基準となる閾値については、網羅性向上と正確性の低下がトレードオフの関係にあるので、これらの交差点 (閾値) を探す.

以上の比較実験に基づいて、擬似投票方式における最適なアイテム間の類似度計算方式と、類似度判定基準となる閾値を提案する.

以下 2 章で先行研究を検討し、3 章で擬似投票方式について概説する. 4 章で本研究システムについて説明し、5 章で評価実験結果を示す.

2. 先行研究

既存の情報推薦システムは、推薦アイテムを提示するとき使用する属性によって表 1 のように分類できる. 表 1 における「狭義の情報推薦方式」は、ユーザの行動履歴を使用して推薦情報を生成するので、シス

テム稼働のために、

- ・手動で行う準備作業が不要である点
 - ・ユーザ数やアイテム数の増加に追従できる点
 - ・意外性のあるアイテムを推薦できる点
 - ・推薦アイテムの予測計算を自動で行える点
 - ・ユーザはクリックを除くキー入力が必要でない点
- において他の方式よりも優位性が高く、多くの先行研究が存在する[1,2,5,6,7,8].

狭義の情報推薦方式は、表 2 のように分類できる. 内容ベース方式は、ユーザの閲覧したアイテムと類似するアイテムを提示する. CF 方式は、ユーザの行動履歴に基づいてユーザ間の相関性を判定し、高い相関がある他のユーザの行動履歴を使用して、新規アイテムに対する嗜好を予測する. 相関係数法は、「ユーザーアイテム」マトリクスにユーザの評価投票値を記録する. これに対し、クラスタモデルはユーザおよびアイテムを分割することで、マトリクスの次元を減少させている.

相関係数法に基づく CF 方式は、表 2 における他の方式と比べ、

- ・予測精度を低下させない点
 - ・意外性のあるアイテムを推薦できる点
 - ・ユーザの嗜好特性や情報特性に影響されない点
- において、他の方式より優位性が高い. そのため、本研究では相関係数法を用いた CF 方式に基づいている.

表 1: 既情報推薦システムの種類

	方式	使用する属性
(1)	チェックボックス方式	ユーザの登録した嗜好情報
(2)	ルールベース方式	サイト運営者の登録した推薦ルール
(3)	狭義の情報推薦方式 [1, 2, 5, 6, 7, 8]	ユーザの行動履歴

表 2: 狭義の情報推薦方式の種類

	方式
(1)	内容ベースフィルタリング方式[7]
(2)	協調フィルタリング方式
	相関係数法[1, 7, 8]
	クラスタモデル[2, 5]
(3)	ハイブリッドフィルタ方式[6, 7]

3. 擬似投票方式

筆者らは CF における「評価値疎ら問題」と「再生

起問題」の解消方式として擬似投票方式を提案し、有効性を示した[9]。擬似投票方式は、ユーザはアイテムの内容に基づいて評価投票を実施していること、内容的に類似するアイテムへは同じ評価投票値が与えられることを前提としている。すなわち、擬似投票は「あるアイテムを嗜好するユーザは、内容的に類似する他のアイテムも同じように嗜好する」という考えに基づいている。

擬似投票は、投票の有無を判定するために、既登録アイテムペア全ての類似度を計算する。類似度計算は、索引語の出現頻度に基づく。ユーザがアイテムに対して評価投票を行ったとき、類似するアイテムに対して、ユーザと同じ評価値にて擬似投票する。擬似投票対象かどうかの判定は、類似性判定基準の閾値に従う。

類似性判定基準の閾値を低く設定すると、類似性有り判定されるアイテムペア数が増加するので、「アイテム-ユーザ」評価マトリクスのスパース性は解消される反面、擬似投票によるノイズが増加し、正確な推薦アイテムを提示できない。また、閾値を低くすると、類似するアイテムが増加するので、推薦アイテムの計算処理量が増加する。

計算処理量が増加する問題点は、システム実装において致命的になる。CFにおける推薦アイテムの計算処理は、推薦アイテムの予測計算時に「アイテム-ユーザ」評価マトリクスを使用する。そのため、 U を異なりユーザ数、 I をアイテム数とした場合、その計算量は $O(U \times I)$ の空間サイズに比例して増大する。しかし、擬似投票を行わない場合、評価マトリクスはスパースであるから、空間サイズは $O(U + I)$ に近似できるので計算量は問題とならない。しかし、擬似投票を実施した場合、評価マトリクスのスパース性はある程度解消し、時間経過と共にその空間サイズは、 $O(U \times I)$ に近づく。このため、ユーザ u_i に対するアイテム k への嗜好度の予測値 $P_{i,k}$ の計算と、上位 N アイテムの選択処理を分離している。具体的には前者をオフラインで実行し、後者をオンラインで実行する。特に計算処理量の多い前者をオンライン計算から切り離すことで、応答速度の高速化と、ユーザ数およびアイテム数の増加に対する大規模化を可能にした。

4. 研究システム

4.1. システムの概要

図1は、実験を行う目的で構築したシステムの全体図である。擬似投票処理は、ユーザ評価履歴を入力として、アイテムペアの類似度テーブルを参照し、類似度の高いアイテムに対して擬似投票を行い、結果をアイテム-ユーザ評価マトリクスに記録する。

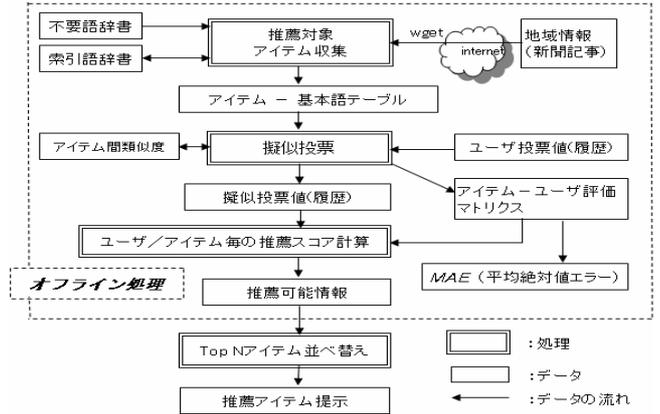


図1：システム全体図

4.2. アイテム間の類似度計算

本研究では、擬似投票の正確性を向上させる目的で下記の点について評価実験を行った。

研究において対象とする4種類のアイテム間類似度計算式[4]を表3に示す。表3に表れる $w_{s,t}$ と $w_{u,t}$ は、アイテム内における索引語の重みを意味する(表4)。表4では、TFのみで類似度計算する場合(表4-A)、TF-IDFで類似度計算する場合(表4-B)を示している。表5と表6は、表3と表4に表れた変数を説明している。本研究におけるアイテム間類似度計算の組み合わせは、「表3x-表4y」とする。ただし、xは表3の式A~Dを指し、yは表4のAまたはBを指す。従って、本研究では8種類の方式を評価する。

表3：類似度計算式の選択肢

	コサイン関数方式(Cosine measure)
A	$S_{u,s} = \frac{\sum_{t \in T_{u,s}} (w_{u,t} \cdot w_{s,t})}{W_u \cdot W_s}$
B	Dice係数方式(Dice formulation) $S_{u,s} = \frac{2 \sum_{t \in T_{u,s}} (w_{u,t} \cdot w_{s,t})}{W_u^2 + W_s^2}$
C	Jaccard計数方式(Jaccard formulation) $S_{u,s} = \frac{\sum_{t \in T_{u,s}} (w_{u,t} \cdot w_{s,t})}{W_u^2 + W_s^2 - \sum_{t \in T_{u,s}} (w_{u,t} \cdot w_{s,t})}$
D	部分一致公式(Overlap formulation) $S_{u,s} = \frac{\sum_{t \in T_{u,s}} (w_{u,t} \cdot w_{s,t})}{\min(W_u^2, W_s^2)}$

表4：索引語の重み付けに関する選択肢

A	TFのみで計算(TF-only formulation)
	$w_{u,t} = r_{u,t} \quad w_{s,t} = r_{s,t}$
B	TF-IDFで計算(Standard formulation, TF-IDF)
	$w_{u,t} = r_{u,t} \cdot w_t \quad w_{s,t} = r_{s,t} \cdot w_t$

表5：変数説明(その1)

A	索引語の重み(Term weights)
	$w_t = \ln(1 + \frac{f^m}{f_t})$
B	関連語の出現頻度(Relative term frequencies)
	$r_{u,t} = 1 + \ln(f_{u,t}) \quad r_{s,t} = 1 + \ln(f_{s,t})$
C	アイテムの長さ(Item length)
	$W_u = \sqrt{\sum_{t \in T_u} w_{u,t}^2} \quad W_s = \sqrt{\sum_{t \in T_s} w_{s,t}^2}$

表 6: 変数の説明 (その 2)

$S_{u,s}$	アイテム u とアイテム s の類似度
T	異なり語の集合
$t \in T_{u,t}$	t はアイテム u とアイテム s の両方に含まれる語
$w_{u,t}, w_{s,t}$	アイテム u またはアイテム s に含まれる語 t の重み
$f_{u,t}, f_{s,t}$	アイテム u またはアイテム s における語 t の出現頻度
f_t	語 t を含むアイテムの数
f^m	アイテム集合における最大 f_t

4.3. アイテム間類似度計算方式の評価方法

アイテム間の類似度計算式の優劣は、計算処理量と、評価値の絶対誤差とを用いて評価する。

式 1 は、MAE の計算式を示している。MAE (mean absolute error) とは、ユーザによる評価投票値と予測値間の平均絶対値エラーである。先行研究において、システム評価方式として提案されている。

$$MAE = \sum_{j \in U} \frac{\sum_{k \in T_{u_j, s_j}} \frac{|r_{j,k} - sv_{j,k}|}{N_k}}{N_j} \quad (1)$$

式 1 において、 j はアイテムへの評価投票を行ったユーザの集合で、 N_j は集合 j の要素数 (異なりユーザ数)、 k はユーザ j が評価したアイテムとシステムがユーザ j のために擬似投票したアイテムの積集合、 $r_{j,k}$ はユーザ j がアイテム k に与えた評価値で、 $sv_{j,k}$ はシステムがユーザ j のアイテム k に対して行った擬似投票値の平均値、 N_k は積集 k の要素数 (アイテム数) である。なお、計算に先立ち、 $r_{j,k}$ と $sv_{j,k}$ は、 $-1.0 \sim +1.0$ の値域に基準化した。

MAE がゼロに近づくほど、ユーザによるアイテムへの評価値とシステムによる擬似投票値の嗜好パターンが類似していることを意味する。この結果を用いて、最適なアイテム間の類似度計算方式を特定する。

5. 評価実験

5.1. 研究データ

研究用データは、「えるこみ」¹ に蓄積されているアイテムのうち、2002 年 10 月 27 日～2003 年 2 月 15 日 (12 週間) の間 1 回以上表示されたアイテムを対象として、その内 6% を標本化した。アイテムの標本数は、2,424 アイテムである。

図 2 にアイテムページの例を示す。ユーザの評価値は、とても (3 点)・まあまあ (2 点)・いまいち (1 点) とした。研究用ユーザ評価値投票履歴データは、上記 12 週間においてユーザが直接評価投票した履歴を標本化した。標本数は 1,275 投票であり、異なりユーザ



本文

投票ボタン

図 2: アイテムページの例

数は 756 人である。なお、「同一ユーザによる同一アイテムへの評価投票」は、第一印象を重視して、初回の評価投票値のみを利用した。

評価実験では、上記のアイテムデータとユーザ評価値投票履歴データを用いて擬似投票を行い、その結果とユーザによる投票値とを比較した。

5.2. 考察

擬似投票の最適な方式は、「アイテム-ユーザ評価マトリクス」の未評価セル (欠損値) を補完しても擬似投票の正確性が低下しない方式である。

1) 類似度計算方式の評価

図 3 は、同じ擬似投票件数に対して、各類似度計算方式における MAE を比較した。最も好成績な方式は Dice 係数方式であった。Dice 係数方式と Jaccard の式は、両方式におけるアイテムペア (比較ペア) 集合の捉え方が違うのみであり、MAE については僅差であった。

部分一致公式の MAE が低い原因は、同方式の分母に起因する。すなわち、アイテムの長さが短い方に照準を合わせて類似度の計算を行っているため、部分的に一致する文書に高いスコアが与えられる。擬似投票方式はこのスコアに応じて擬似投票の有無を判断する。そのため、文書的に長いアイテムの一部を含む他のアイテムに擬似投票が行われる。しかし、部分的に一致する文書のみがユーザの嗜好を反映しているため、ノイズ的擬似投票によって MAE が低下した。

¹ サンケイリビング新聞社が提供している地域情報サイト「えるこみ」 <http://www.lcomi.ne.jp/>

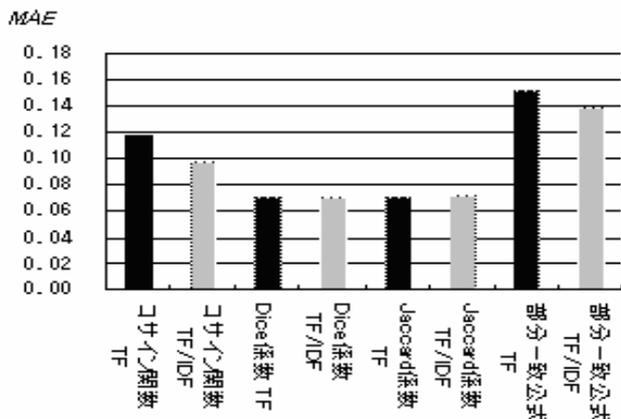


図3: 類似度計算方式の平均絶対値誤差 (MAE)

TF と TF-IDF を比較した場合、各方式は僅差である。その理由は、全アイテム数が 2,424 であり、アイテムの標本サイズが小さいことにより IDF の効果が現れないためである。

従って、本実験データについては、TF に基づく Dice 係数方式で計算すべきと考える。

2) 類似度判定基準閾値の評価

類似度判定基準閾値については、類似度計算方式で好成績をあげた Dice 係数方式について実験した。MAE が同じになるという条件下で、(1)それぞれのアイテムペア集合の類似度上位 $n\%$ へ擬似投票を行ったときの擬似投票件数と (n は 0.1~50 の範囲の複数值を観測)、(2) Dice 係数の計算結果に対する類似性判定基準に基づいた擬似投票件数とを比較した。この結果、(2)の方が投票件数が多く優位である。すなわち、擬似投票件数を増加させても MAE の低下速度が遅いという点で、(2)は(1)より優位であった。

次に好成績を示した(2)について、Dice 係数の計算結果に対する類似性判定基準を比較した。類似性判定基準の閾値緩和 (擬似投票件数増加) と正確性向上は、トレードオフの関係にある。そこで、(A)類似性判定基準 (閾値) を 0 (アイテムは類似しない) から 1.0 に向けて変化させ、擬似投票件数を観測した。また、(B) 正確性を 2 (誤差率 100%) からゼロ (誤差率 0%) に向けて変化させ、擬似投票件数を観測した。観測結果から、(A)(B)いずれにも変異点が存在することがわかった。変異点は、擬似投票件数が急増する反面、正確性が急低下するポイントである。この変位点が、本研究データにおける類似度判定基準閾値であると判断した。

3) 計算処理量の評価

類似度計算にかかる処理時間を計測した。Dice 係数方式を 100 とした処理時間の割合は、Dice 係数、Jaccard 計数、コサイン関数は、ほぼ同じで計算量であるのに対し、部分一致公式は 1.2 倍の処理量であった。

6. おわりに

本研究では、擬似投票における類似度計算の最適な方式が TF に基づく Dice 係数方式であることを示した。そして、同方式における類似度判定基準閾値の求め方について示した。今後は、本研究の結果に基づいて、先に構築したシステムの擬似投票アルゴリズムに本研究で得た最適方式を基に大規模化し、再度評価を行う。

謝辞

評価実験用データを貸与していただきました サンケイリビング新聞社の皆様に感謝します。

文 献

- [1] B. Sarwar, G. Karypis, J. Konstan and J. Riedl. Item-Based Collaborative Filtering Recommendation Algorithms. *Proc. the 10th International World Wide Web Conference (WWW10)*, pp.285-295, 2001.
- [2] C. Basu, H. Hirsh and W. Cohen. Recommendation as Classification: Using Social and Content-based Information in Recommendation. *Proc. Recommender System Workshop AAAI 1998*, pp.11-15, 1998.
- [3] G. Linden, B. Smith and J. York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *Proc. IEEE Internet Computing*, 2003. <http://dsonline.computer.org/0301/d/w1lind.htm>
- [4] J. Zobel and A. Moffat. Exploring the Similarity Space. *Proc. the ACM SIGIR FORUM*, 32(1), pp.18-34, Spring, 1998.
- [5] L. H. Ungar and D. P. Foster. Clustering Methods for Collaborative Filtering. *Proc. Workshop on Recommender Systems at the 15th National Conference on Artificial Intelligence*, pp.114-129, 1998.
- [6] M. Balabanovic and Y. Shoham. Fab: Content-Based Collaborative Recommendation. *Proc. CACM*, vol.40, no.3, pp.66-72, 1997.
- [7] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes and M. Sartin. Combining Content-Based and Collaborative Filters in an Online Newspaper. *Proc. Recommender Systems Workshop at ACM SIGIR 1999*. <http://www.csee.umbc.edu/~ian/sigir99-rec/>
- [8] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Proc. ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, NC*, pp.175-186, 1994.
- [9] 宇田隆幸, 藤井敦, 石川徹也. ユーザ投票と情報アイテム間類似度を併用した情報推薦システム: 擬似投票方式の提案と評価. 情報処理学会研究報告, 2004-DBS-132, pp.105-112, 2004.