

N-017

教育支援システムにおけるレコメンド機能に関する研究

Study on Recommender System for Education System

瀬川 卓也[†]
Takuya Segawa

和田 雄次[‡]
Yuji Wada

1. はじめに

時と場所を選ばずに学習できる e-Learning は大学や企業研修等で普及し発展してきた。しかし既存のシステムの多くは教材を単純に並べたものや、オンラインテストの実施に留まる等、学習に関して e-Learning ならではの利点が見当たらないシステムが多い。個別復習支援システム AIRS では、学習者毎の学習志向にあったコンテンツの提供を行っている[1]。これはその学習者に適したコンテンツを、その学習者の履歴データや他学習者の履歴データによって学習を効率化させるアルゴリズムを利用している[2]。しかし履歴データが蓄積されていない時点での推薦には対応できないという欠点がある。それを解決する手法としてユーザの背景情報を用いた属性相関手法を提案し、その有用性を評価する。本稿では上記の目的に対して以下の 2 点を行う。

1. ユーザの背景情報を用いた属性相関手法を提案する。
2. 提案手法により被験者実験を実施し、評価する。

2. 関連研究

ユーザの嗜好を推測し、コンテンツを推薦するシステムは現在では研究が進み様々な web サービスで利用されている。例えば EC サービスの Amazon[3]ではユーザの商品ページ閲覧履歴、購買履歴などのデータをもとに、ユーザが好みそうな商品を推薦し提示してくれる。これらの多くは協調フィルタリング (以下、CF) [4]が利用されている。だが教育の現場において、授業など現実ベースの教育に対して CF を利用する研究[5]はあるが、実際に e-Learning システムに実装されている例は少ない。本学では AIRS においてユーザに対して学習コンテンツの推薦を行っている。しかし CF にはユーザがある程度のコンテンツを利用しないと、履歴が得られず精度の高い推薦結果が得られない、という問題がある。これは Cold-Start 問題とされている[6]。講義などで習った内容の不明点の解消や再習得に利用したいユーザにとってこれはマイナスである。本研究ではこの Cold-Start 問題に焦点を当てた属性相関手法を提案する。

3. 協調学習推薦

提案手法の説明に入る前に、簡単に協調学習推薦について触れておきたい。これ以降、ユーザを学習者と、履歴データを学習履歴と呼称する。

[†] 東京電機大学 Tokyo Denki University

[‡] 東京電機大学 Tokyo Denki University

3.1 推薦手法

CF はユーザベース推薦とアイテムベース推薦の二つに大きく分かれる[7]。ユーザベース推薦は推薦対象の学習者と類似した学習者を見つけ出し、類似学習者の学習履歴から有効なコンテンツを抽出して推薦する。これによって単純なコンテンツ毎のアクセス数の分布の違いに左右されずにその学習者に最適なコンテンツを推薦できる可能性が高い。図 1 にユーザベース推薦方式の図を示す。

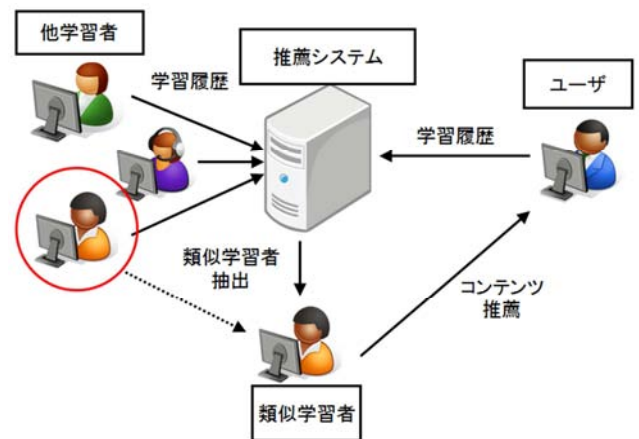


図 1 ユーザベース推薦

3.2 類似学習者抽出

全学習者 $\{U|u_0, u_1, u_2, \dots, u_n\}$ の学習履歴から推薦対象の学習者 (u) と学習傾向が類似している類似学習者を導き出す。学習者 u と学習者 u_x それぞれの学習履歴を比較し、相関係数などで評価する。これを学習者 u 以外の全ての学習者に対して行い、学習者 u に対して類似度の高い順番で順序付けし、特に高い学習者を類似学習者とする。表 1 の場合、基準となる学習者 N と学習者 X, Y を比較している。学習者 X は学習履歴全一致であるのに対し学習者 Y は一件一致のみであるため、学習者 X は学習者 Y と比較して相関係数が高くなり、類似学習者として抽出される。これを全学習者 U に対して実施し、それぞれの学習者の類似学習者を抽出する。

表 1 学習履歴

	Content 1	Content 2	Content 3
User N	A	B	C
User X	A	B	C
User Y	C	D	E

3.3 推薦コンテンツの抽出

類似学習者の学習履歴において多く閲覧されたコンテンツは類似学習者にとって苦手であったと仮定して、当該コンテンツを苦手克服コンテンツとして推薦する。類似学習者がコンテンツ A を 9 回、コンテンツ B を 3 回、コンテンツ C を 5 回閲覧していた場合、最も閲覧回数が多いコンテンツ B を当該学習者に推薦する。複数の類似学習者を抽出した場合にはそれぞれの結果の平均を取るという手法が使われている[1]。

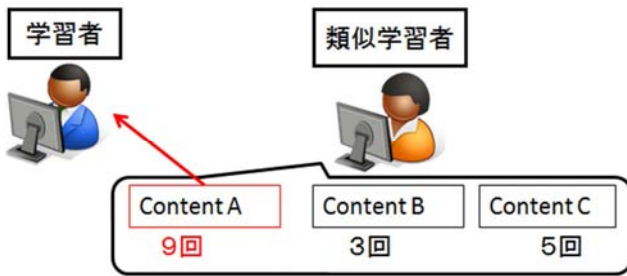


図2 推薦コンテンツの抽出

4. 属性相関手法

以上の通り、協調学習推薦は学習者の履歴をベースに推薦コンテンツを選定するものであるため、学習履歴が存在しない学習者や学習履歴が一定以上蓄積されていない学習者（以下、新規学習者）に対してうまく類似学習者を抽出できず、結果として精度の高い推薦結果を提示することができない。

そこで新規学習者の背景情報を収集し属性データとして扱い、推薦を行う学習者と属性データが似通っている属性類似者の抽出を行う手法を提案する。

4.1 概要

協調学習推薦手法において Cold-Start 問題が発生してしまう主な原因は、3.2 節に述べた類似学習者抽出において新規学習者の履歴が少ないことにある。つまり、新規学習者と既存の学習者の間の相関を別の方法で評価できれば、この問題を解決できると考える。図 3 は提案手法を取り入れた全体のフローチャートである。

4.2 属性データ

属性データはメタデータ、例えば年齢、性別、趣味嗜好、得意科目、苦手科目などの個人情報から取得する。これらは学習履歴が無くとも新規学習者が必ず保持している情報になる。

4.3 属性データの体系化

属性データは各データ間で関連性がある。例えば、高校教育においてライティングと物理に関連性は見出せないが数学 IA と物理には理系科目という点において一定の関連

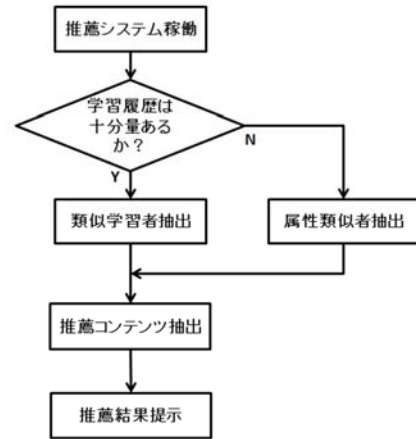


図3 属性相関手法フローチャート

性が見出せる。このように属性データの内部を体系化して表現する必要があると考える。これは学習履歴と比較して利用できる属性データの量が少ないことが予想されるからである。

学習属性の場合、例えば高校における教育を例にとると、授業科目をルートノードとして、子ノードを理系と文系という系統、さらにその子ノードに大まかな授業区分、さらにその下にはより詳細な授業の名称と定義付けられ、階層化される。下の層に移るにつれデータの具体性が上がるので、情報としての重みが増していく。この重みを、ルートノード直下の第一層を 0.5 点、以下層を下る毎に 1 点、2 点、4 点と 2 倍ずつ増える点数で表現する。これは深層部の評価値を高めるためである。図 4 は学習に関する属性データを体系化した例である。これを「学習」や「職種」などと、評価に必要な属性の種類だけ作成する。

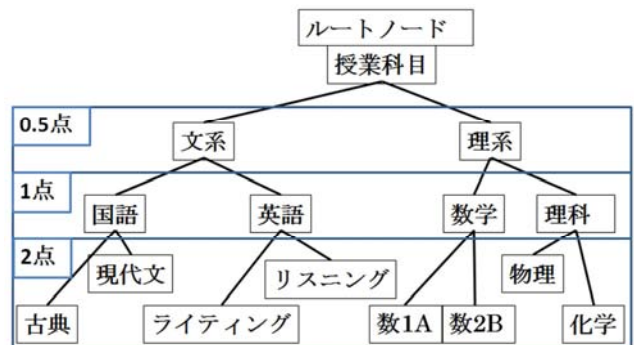


図4 階層的な属性データ体系

4.4 属性データ類似度と属性類似者の抽出

新規学習者と他の全学習者の間で属性の適合を比較し全ての属性における点数を合算し、点数が高かった順番で順序付けし特に高い学習者を属性類似者とする。表 2 の場合、学習者 N は物理を得意科目としているので上位ノードである理科、理系も情報として持っている。学習者 X は全一致であるので 2 点。学習者 Y は理系科目のみの一致であるので 0.5 点の点数が与えられる。従ってこの段階では X が属性類似者となる。これを用意した属性だけ繰り返して合

算し、最終的に点数が高かった学習者を属性類似者として抽出することになる

表 2 属性類似度テーブル

	理系	理科	物理	化学	数学
UserN	0.5	1	2	0	0
UserX	0.5	1	2	0	0
UserY	0.5	0	0	0	1

4.5 類似学習者と属性類似者の関係

図 3 のフローチャートでは学習履歴が十分量ある場合は属性相関手法をスルーして通常の協調学習推薦のアルゴリズムで動くことが示されている。これは属性相関手法が学習履歴による類似学習者の抽出よりも優れた結果を導くことはないという推測からである。なぜなら、類似学習者の抽出には推薦するコンテンツそのものを履歴として使用しているのに対し、属性相関はそれとは直接関連のない、学習者の背景情報を使用している。この二つを比較したとき、学習履歴の方がシステムに対して純粋な情報であることは間違いない。例えば、今まで本を読んだことの無い人に本を推薦するために、別の趣味や特技などの要素からその人が好みそうな本を推薦するという考え方を取り入れているからだ。この手法は Cold-Start 問題を解決することを第一義としている。

5. 実験

提案手法の効果を明らかにするために被験者実験を行った。評価したのは以下の 2 点。

- 提案手法に有効性はあるか
- 属性データ選定に関しての仮説の立証ができたか

5.1 属性データ選定に関する仮説

属性データは 3.1 節で述べた通り学習者のメタデータになるのだが、学習者の個人情報全てが必要とは限らない。例えば健康を保つための運動を推薦するなら身長や体重などの身体的な情報は重要になるが、小説を推薦するのにこれらの情報は必要ないはずである。つまり、推薦するコンテンツに対して関連がある属性が高い相関を示すのではないだろうかという仮説を立てた。ここでは教育支援システムにおける推薦システムとなるため、学習に関する情報は特に学習と関係のない属性と比較して高い相関を示す、ということになる。

5.2 実験内容

5.2.1 事前準備

予め学習に関する属性として高等学校の授業科目を、それ以外の属性として趣味をそれぞれ体系化して用意する。趣味を選んだ根拠は、身長などの生来的な情報よりも学習者自身が後天的かつ能動的に決定しているため、学習者の

嗜好がうまく反映されているのではという推測に基づいている。学習属性に高等学校の授業科目を選択したのは、実験参加者である大学生の学年による差異を出さないためである。

趣味に関しては体系化されたモデルが存在していなかったため憶測で体系化した。高等学校の授業科目に関しては文部科学省が定めている高等学校学習指導要領を参考にしている[8]。属性の階層はそれぞれ 3 層に統一する。これは、階層の深さによって得点の重みが変わることによって発生する得点差のバイアスを抑えるためである。よって得られる属性データは 2 属性 3 階層のものになる。

5.2.2 被験者について

2 グループ合計 18 人の学生の被験学習者集団に対して、実験直前にアンケートを行った。質問は以下の 2 点。

- 高校生時に得意だった教科(学習属性)
- 現在の趣味(趣味属性)

このアンケートによって被験学習者の属性を収集した。回答形式は、選択する属性の量が多かったため自由記述にしたが、正規の回答ではなかった場合こちらから再度質問をすることで属性情報の正規化を図った。

その後被験学習者に実験の内容について以下の 3 点を説明し、実施した。

- 学習時間は 15 分
- コンテンツは「データベース」
- 学習時間終了後に実力テストあり

また、コンテンツ「データベース」は 15 分間で学習しきれものではないので、各自わからない部分を選択し重点的に学習してもらうようアナウンスを行った。これはコンテンツの 1 番最初から順番に学習するという事態を避けるためである。

また、この実力テストも被験学習者にモチベーションを高めて短時間で勉強を効率化してもらうためのものであり、実験の評価自体には影響しない。

5.2.3 分析方法

図 5 に実験の概略図を示す。15 分間の学習で得た学習履歴による類似度(類似学習者)と、収集した属性データによる類似度(属性類似者)をそれぞれ算出、比較し評価する。具体的には、上記 2 つの類似尺度からそれぞれ得られた類似学習者と属性類似者の順位に対する順位相関を被験学習者の数だけ取得する。学習履歴による類似度の算出には jaccard 係数を用い[7]、順位相関の算出にはケンドールの順位相関係数を用いる[9]。属性相関手法はあくまで新規学習者に向けられたものである。学習履歴による類似度は信頼性が高いためこちらを基準として扱う。つまり学習履歴による類似度(類似学習者)における順位と、提案手法によって算出された類似度(属性類似者)における順位の間で順位相関を取ることで、これの被験学習者全員分の平均が高い場合は、学習者の抽出に関して類似学習者の抽出に迫る信頼性が得られたとみなせるため、有効性が認められる事になる。

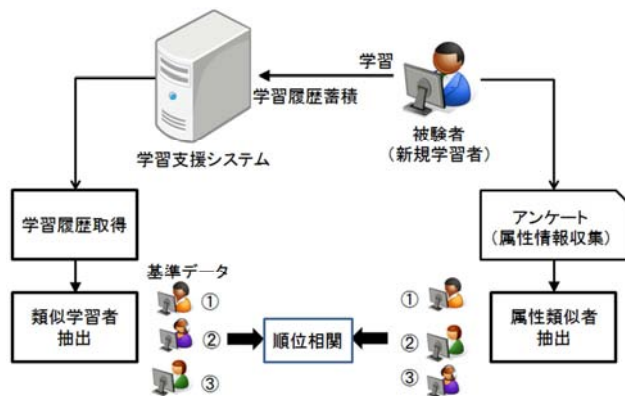


図5 分析方法の概要

5.3 実験結果

表3, 4は、それぞれのグループにおける実験結果であり、数値はそれぞれの属性類似者と類似学習者の順位との間で順位相関を取ったものをグループ全体の平均と標準偏差を取ったものである。順位相関は相関係数であるため-1から1の間を取る。1に近ければ相関が強く、-1に近ければ逆相関、0に近ければ相関が弱いになる。それぞれの表を見れば分かる通り、どちらも平均は|0.1|以下であったため相関なしということになり有意な結果は得られなかった。また、学習属性と趣味属性では誤差ほどの差異しか無かったため、仮説は否定された。標準偏差も一つを除いて0.2以下にまとまっているため、この結論は妥当であると考えられる。

表3 グループ1の実験結果

	学習属性	趣味属性	総合属性
平均	-0.077	-0.044	-0.095
標準偏差	0.1805	0.2204	0.1798

表4 グループ2の実験結果

	学習属性	趣味属性	総合属性
平均	0.1032	0.0238	-0.0397
標準偏差	0.4074	0.2901	0.3748

5.4 考察

このような結果になった原因を考えると、属性データの量が不足していたという可能性が浮かび上がる。そこで被験者のデータを個別に見ると、属性相関による順位において、同率順位の発生が被験者の多くに対して確認された。中には趣味属性が他被験学習者全員に対して一律の数値が出てしまい順位相関が求められないケースも数例あった。

だがサンプルデータが少なかつたとはいえ相関係数の平均の値が0に近かった結果は無視できない。点数計算の方法を変更したりなどの総合的なアルゴリズムの変更も視野に入れる必要である可能性もある。

6. おわりに

本稿では、属性相関手法についての提案を行い、その有用性を評価するために実験を行い、そこで背景情報と推薦対象のコンテンツの関係性についての仮説の検証も行った。その結果、提案手法ではCold-Start問題の解決に対して有効性を示せず、仮説も否定された。

これはこの手法によって推薦を行うことが難しいことを示しているのと同時に、学習の推薦に関して過去に学習した事柄の情報と、それとは関連の無い情報の間に、現在学習している事柄との直接的な関係は弱いことを表していると考えられる。今後は属性データを増やした状態での再実験を行うとともに、新しいCold-Start問題を解決する枠組みの開発に取り掛かりたい。

謝辞

本研究は、科学研究費補助金(基盤研究(C)課題番号21500908「学習教材向け二方向遷移推薦方式における適応型推薦技術に関する研究」)の支援による。

また、実験に際してAIRSの利用とデータ提供をしてくださった東京電機大学情報環境学部の土肥紳一准教授、及びオーサリング研究室の皆様には感謝いたします。

参考文献

- [1] 松澤俊典, 山口未来, 和田雄次, 土肥紳一:教材コンテンツ双方向推薦システムの実装, 情報処理学会研究報告, 情報処理学会研究報, コンピュータと教育研究会報告 2008(13), pp.127-132 (2008)
- [2] 神島敏弘:推薦システムのアルゴリズム, 人工知能学会誌, vol22, No6, pp.826-837(2008)
- [3] amazon: <http://www.amazon.com>
- [4] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P. and Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *proc. ACM Conf. on computer Suspend Coparative Work*, pp.175-186(1994).
- [5] 北村智:協調学習研究における理論的関心と分析方法の整合性: 階層的データを扱う統計的分析手法の整理, 日本教育工学会論文誌, vol33, No3, pp.342-352(2010)
- [6] Schein, A., Popescul, A., Unger, L. and Pennock, D.: Methods and Metrics for Cold-Start Recommendations, *25th Annual ACM SIGIR Conf.*, pp253-260(2002)
- [7] Toby Segaran (著), 富山仁健, 鴨澤眞夫(翻訳): 集合知プログラミング
- [8] 文部科学省ホームページ: http://www.mext.go.jp/b_menu/shuppan/sonota/990301d.htm
- [9] 授業評価・市場調査のための「アンケート」調査・分析ができる本: 第9章