Coaching Agent: Making Recommendations for Behavior Change. A Case Study on Improving Eating Habits

Jules Vandeputte, Antoine Cornuéjols, Nicolas Darcel, Fabien Delaere, Christine Martin

Coaching Agent: Making Recommendations for Behavior Change. A Case Study on Improving Eating Habits

Jules Vandeputte
UMR MIA-Paris, AgroParisTech, INRAe, Université Paris-Saclay
Paris, France
jules.vandeputte@agroparistech.fr

Antoine Cornuéjols
UMR MIA-Paris, AgroParisTech, INRAe, Université Paris-Saclay
Paris, France
antoine.cornuejols@agroparistech.fr

Nicolas Darcel
UMR PNCA, AgroParisTech, INRAe, Université Paris-Saclay
Paris, France
nicolas.darcel@agroparistech.fr

Fabien Delaere
Danone Nutricia Research
Palaiseau, France
Fabien.DELAERE@danone.com

Christine Martin
UMR MIA-Paris, AgroParisTech, INRAe, Université Paris-Saclay
Paris, France
christine.martin@agroparistech.fr

ABSTRACT

In many applications, the desire is to change the behavior of users over a period of repeated episodes of similar decision making. For example, in a nutritionist scenario, one may try to encourage the user to adopt better food consumption habits. In this paper, we propose to model this recommendation scenario with a long-term goal, that we call coaching, as an iterative two-player game. At each decision time, the user makes a proposal (e.g., a meal) based on her/his preferences, the coach can then suggest a change in this proposal that the user may or may not accept. After each episode, the user updates her/his preferences and the coach adapts its model of the user.

We propose a formalization of the coaching problem and discuss several possible criteria to measure the performance of a coach. Different coaching strategies are described in the paper. They are then tested and compared using a real-world dataset in the field of nutrition, where a user is simulated using a simple, but general, model of decision making and adaptation. Results show that it pays to adapt to user characteristics and use non-myopic strategies, which aim for long-term gains, when the number of interactions becomes large.

Although illustrated on choice sequences for food consumption, the scope of the proposed method goes far beyond this use case, as in sports, health, entertainment or tourist activity choices, etc.

KEYWORDS

Automated assistant; Teacher-student scenario; Recommendation; Sequential Recommendation; Personalized Recommendation.

ACM Reference Format:
Jules Vandeputte, Antoine Cornuéjols, Nicolas Darcel, Fabien Delaere, and Christine Martin. 2022. Coaching Agent: Making Recommendations for Behavior Change. A Case Study on Improving Eating Habits. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 9 pages.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

1 INTRODUCTION

Recommender systems have become crucial in helping users to identify relevant items in the enormous choice sets they face, specially on online platforms. The classical scenario is characterized by the facts that the items chosen are generally consumed only once and that the recommendations are usually one-off recommendations that only take into account the set of past choices and not their history.

This scenario where the user is seen only as a potential consumer whose preferences need to be identified in order to maximize his/her consumption activity contrasts with another perspective in which the user is considered as *a person who aims at changing his/her own preferences* using a recommender system that acts like a coach.

In this context, the user is repeatedly confronted with a set of choices and wishes to improve his/her behavior over time. This is the case for food choices that must be made several times a day and that have a large impact on health, as well as for sports activities or cultural choices, for example. Here, the performance metrics is no longer the number of purchased items, but is related to observed changes, if possible lasting ones, toward "better" behavior. We use the term "coaching" to refer to such situations because the recommender system acts as a personal assistant to which the user turns to in order to improve his/her behavior over time through repeated interactions.

The coach can influence the user's repeated choices through two different channels: either by altering the set of possible choices or the context of the choices, or by modifying the user's preferences. Only the latter may allow the user to ultimately get rid of the coach.

In this paper, we propose to model this new recommendation setup, called *coaching*, with an iterated two-player game. At each decision time, the user proposes a solution (e.g. a meal) according to his/her current preferences. The coach may then suggest a change in the solution (e.g. a change of dish) that the user, in turn, may or may not accept. After each episode, both players update their policy. The user may change his/her preferences, and the coach may alter his/her recommendation's strategy.

This scenario where the recommendation is based upon the user's expressed preferences is well suited to coaching because (*i*) the user makes choices that most of the time depend upon the

circumstances (e.g. what is available in the refrigerator) or upon his/her own characteristics (e.g. being vegetarian or diabetic) and therefore the coach must address the possible choices within the given context, or risk to appear completely arbitrary and irrelevant, and (*ii*) because it is known that subjects tend to change their behavior more when they are engaged by their own choices and not passively following external recommendations ([8, 11]).

In this work, our contribution is fourfold.

(1) We define a *novel recommendation scenario* where the user faces repeated decision making episodes and seeks the help of a "coach" in order to improve his/her behavior (Section 2).
(2) We formulate this scenario as an iterated two-player game, where each player adapts its own strategy after each interaction step, and we *propose performance criteria* that enable to evaluate the coaching strategies (Section 2).
(3) We introduce the idea of guiding the choices of the coach using estimates of the *acceptability of substitutions* from one item i to another one j, based on past observed behavior. This yields better informed strategies (Section 4).
(4) We *propose and compare several possible strategies* for the coach (Section 3) and compare them empirically (Section 4).

In the rest of the paper, we first state the problem more formally in Section 2. In Section 3, we examine possible strategies for the coach to optimize the change in the user's habits. The section 4 reports experiments to test these strategies and the results obtained. Section 5 positions the coaching scenario with respect to existing works. Section 6 concludes.

2 PROBLEM STATEMENT AND FORMULATION

2.1 The Coaching Scenario

One typical example of the coaching scenario is the choice of meals that each of us has to make day after day. In order to improve our consumption behavior we may ask the help of a personal coach. Here, for simplicity of exposure, we will suppose that at each time step t, the user U chooses only one item in a set of items $\mathcal{I}$ (e.g. one has to choose a dessert). We will assume that the preferences of the user at t can be expressed as a probability distribution Π_t defined over the set of items. We have $\Pi_t = (\pi_t(i))_{i\in\mathcal{I}}$ where $\pi_t(i)$ represents the probability that the user selects item i at time t.

The goal of coaching is to modify the behavior of the user, hence Π_t, so as to improve it. In this scenario, we assume that each item $i \in \mathcal{I}$ is associated with a score $s(i) \in \mathbb{R}$. For instance, nutritional rating systems such as the Nutri-Score in Europe evaluate each type of food with respect to their content in energy, fibers, saturated fatty acids and sodium, among other things.

Given that a user is characterized by a preference distribution over $\mathcal{I}$, the associated mean score value at t can be expressed as:

$$\mathcal{V}(\Pi_t) = \sum_{i\in\mathcal{I}} \pi_t(i)\cdot s(i) \tag{1}$$

The goal of the coach, and of the user, is to improve as much as possible this instantaneous mean score in the shortest possible time. But only the coach knows the score function $s(\cdot)$ and, conversely, only the user "knows" Π_t. The coach may only estimate it from observations of the user's sequence of choices.

2.2 An Iterated Two-player Game

In a coaching situation, as for instance, when one is learning to play tennis, the player plays a shot first, and only then can the coach make a comment like "no, no, no, you should have played that shot instead". Thus, the coach does not show what should have been played in the first place. The player is therefore not in a situation to imitate a better player as in [17, 18]. But the coach uses the student's choice to suggest a better move or a better solution. It is thus hoped that the student will gradually change his/her habits by adopting the suggested moves in place of the initially preferred choices. Accordingly, we formulate the coaching scenario as an *iterated two-player game* between the user U and the coach C.

(1) U makes an item proposal, for example i, using his/her vector of preferences Π_t.
(2) C analyzes U's proposal, and suggests, if judged useful, a modified proposal j, using his/her knowledge of the value of the items through the score function $s(\cdot)$, and his/her estimation of U's ability to accept the proposal.
(3) U accepts or rejects the substitution proposal provided by C.
   - If U accepts C's proposal (replacing item i by item j), he/she modifies the preference vector Π_t according to his/her learning capacity, so as to propose the recommended item more frequently in the future.
   - Otherwise, U does not modify the preference vector.

This is how we account for the fact that U can learn as he/she interacts with C, the idea being that U, if he/she accepts the modification $i \to j$ proposed by C, is more ready to choose j in the same context in the future, instead of i.

Formally, the two players are modeled as follows:

Model of the user U. A user is characterized by three components.

(1) A probability distribution over the set of items that may change over time: Π_t. This expresses the *preferences* of the user, and dictates the behavior of U.
(2) A matrix $\mathbf{M}_t\colon \mathcal{I}\times\mathcal{I} \to [0,1]$ of which each element $m^t_{i,j}$ expresses the probability that U *accepts a suggestion* of change from an item $i \in \mathcal{I}$ to an item $j \in \mathcal{I}$ at time t. If the suggestion is not accepted, U stays with his/her choice.
(3) A *propensity to modify* Π_t when a suggestion of change $(i \to j)$ by the coach C has been accepted by U. In our model, we consider changes of preferences of the form:

$$\begin{cases} \pi_{t+1}(i) &= (1-\lambda)\,\pi_t(i) \\ \pi_{t+1}(j) &= \pi_t(j) + \lambda\,\pi_t(i) \end{cases} \tag{2}$$

where $\lambda \in [0,1]$. This formula guarantees that, if Π_t is a probability distribution, then so is Π_{t+1}. One can consider λ as a parameter that controls or characterizes the learning rate of the user. A value $\lambda = 0$ means that no learning takes place, while the closer to 1 the value of λ, the larger the effect of accepting a recommendation $i \to j$ of C. When $\lambda = 1$, there is a complete transfer of the probability to chose i to the probability of choosing j. When $j = i$, then there is no change in the probability: $\pi_{t+1}(i) = \pi_t(i)$. In the following, we note $f_{i\to j}(\Pi)$, the preference vector after the user, starting from Π, has accepted the substitution $i \to j$.

Model of the coach C. The coach C examines the choice i made by U at time t and computes the expected gain $G(i,j)$ of each possible substitution $i \to j$ (including "no substitution": $i \to i$). C choses the recommendation $i \to c_t(i)$ according to:

$$c_t(i) \;=\; \underset{j \in \mathcal{I}}{\text{ArgMax}}\, G(i,j)$$

The expected gain $G(i,j)$ depends on the respective quality $s(i)$ and $s(j)$ of i and j (e.g., their nutritional quality) and on other parameters depending on the recommending strategy used by C as described in Section 3.

The coach C may maintain an estimate $\widehat{\mathbf{M}}_t$ of the acceptability matrix $\mathbf{M}_t$ of the user, an estimate $\widehat{\Pi}_t$ of Π_t and an estimate of λ, the learning rate that characterizes U. Given these characteristics, the iterated two-player game is described in Algorithm 1.

begin
  $t = 0$
  while *coaching in play* **do**
    $t \leftarrow t+1$
    Decision making phase
      U chooses item i according to policy Π_t.
      C suggests substitution $i \to c_t(i)$ according to the strategy c_t.
      U accepts the substitution $i \to c_t(i)$ with probability $m_{i,c_t(i)}$.
    Learning phase for the user
      U changes the preference vector:
      $\Pi_{t+1} \leftarrow f\big(\Pi_t, i, c_t(i)\big)$
      (f defined according to Eq. 2)
    Learning phase for the coach
      C changes its strategy: $c_{t+1} \leftarrow g\big(c_t, i, c_t(i)\big)$
      (see Section 3)
  end
end

Algorithm 1: The two-player game between U and C

The updating function g used by the coach to modify his/her strategy is the subject of Section 3.

In the rest of the paper, we assume that the matrix $\mathbf{M}_t$ that controls the acceptability of suggestions $i \to j$ by U is constant over time, hence the notation $\mathbf{M}$.

2.3 Evaluating the Coach

The goal of coaching is to make the user U follow a trajectory in the space of preferences, i.e. in the space of probability vectors Π_t. Given a space of preference vectors $\mathcal{P}$, we call $\Pi^\star$ the set of probability distributions in this space that are associated with the maximal value of $\mathcal{V}$: $\Pi^\star = \text{ArgMax}_{\Pi \in \mathcal{P}}\, \mathcal{V}(\Pi)$.

Let us note Π_0 the starting preference vector of the user U. Given the matrix of acceptability of suggestions $\mathbf{M}$ and λ the propensity to learn that characterize U, the space of preference vectors that can be attained by U may vary. For instance, U may not be ready to change his/her choice of *French fries* under any circumstances and, given the model of the user described in Section 2.2, this implies that $\forall t \geq 0, \pi_t(\textit{French fries}) \geq \pi_0(\textit{French fries})$.

Accordingly, the set of optimal preference vector(s) reachable from Π_0 is dependent upon Π_0, and we will denote it by $\{\Pi_0^\star\}$ and by $\mathcal{V}_0^\star$ the value of any one of the optimal preference vector: $\mathcal{V}_0^\star = \mathcal{V}(\Pi_0^\star)$.

There are several ways to assess the merit of a coaching strategy.

(1) The first method stresses the *level of performance that one wants to obtain* $\eta\mathcal{V}_0^\star$ with $\eta \in (0,1)$ and measures the mean number of interactions that the coach needs to guide the user towards this performance level: $\overline{T}_\eta$ starting from Π_0. The problem is that it is difficult to evaluate $\mathcal{V}_0^\star$, the best performance that U may attain from Π_0.
(2) A dual method consists in *giving a "budget" to the coach* in terms of a number T of interactions, and measures the mean gain in performance $\overline{\mathcal{V}}_T = \text{mean}\big(\mathcal{V}(\Pi_T) - \mathcal{V}(\Pi_0)\big)$ after T interactions. The mean here is taken from repeated episodes of T interactions starting from Π_0 since an episode stems from stochastic choices from the user.
(3) It is also possible to consider a *criterion based on the whole trajectory* in the preference vector space, for instance a cumulated gain: $G(T) = \sum_{t=1}^{T}\big(\mathcal{V}(\Pi_t) - \mathcal{V}(\Pi_0)\big)$.

In the rest of this article, we will focus on the second criterion $\overline{\mathcal{V}}_T$. It allows easy comparisons, especially in the case of a real user having a limited number of interactions with the system.

3 THE SPACE OF COACHING STRATEGIES

At each time step, the coach must choose a suggestion to the user that takes the form of a substitution $i \to j$ with $(i,j) \in \mathcal{I}^2$ and possibly $j = i$ when the coach is satisfied with the choice of the user. The goal of the coach is to "push" the user towards better preferences, as measured by Equation 1.

It is important to note that the coach does not have a priori knowledge of an optimal preference vector $\Pi^\star$ since this depends on the characteristics of each user, specially $\mathbf{M}$ and λ and of his/her starting preferences.

Suppose that $V(\Pi)$ represents the desirability that the user is in state Π from the perspective of the coach who looks at long term expected benefits if the coach follows the optimal policy defined below by Equation 3. Then, for each possible choice of item $i \in \mathcal{I}$ by U, the coach should choose the substitution $i \to j^\star$ such that:

$$j^\star \;=\; \underset{j \in \mathcal{I}}{\text{ArgMax}} \Big\{ m_{i,j}\,\big[\big(s(j)-s(i)\big) + V\big(f_{i\to j}(\Pi)\big)\big] \;+\; (1-m_{i,j})\,V(\Pi) \Big\} \tag{3}$$

Now, the expected value $V(\Pi)$ of all preference vectors Π are fixed point of the Bellman equation that relates the updated evaluation $V_{t+1}(\Pi)$ with the current evaluations of the preference vectors $V_t(\Pi')$ that may ensue a recommendation by the coach (see Figure 1).

$$V_{t+1}(\Pi) \;=\; \sum_{i \in \mathcal{I}} \pi(i)\; \underset{j \in \mathcal{I}}{\text{Max}} \Big\{ m_{i,j}\,\big[\big(s(j)-s(i)\big) + V_t\big(f_{i\to j}(\Pi)\big)\big] \;+\; (1-m_{i,j})\,V_t(\Pi) \Big\} \tag{4}$$

where $f_{i\to j}(\Pi)$, the preference vector resulting from the acceptation of the suggestion $i \to j$, is defined by Equation 2.

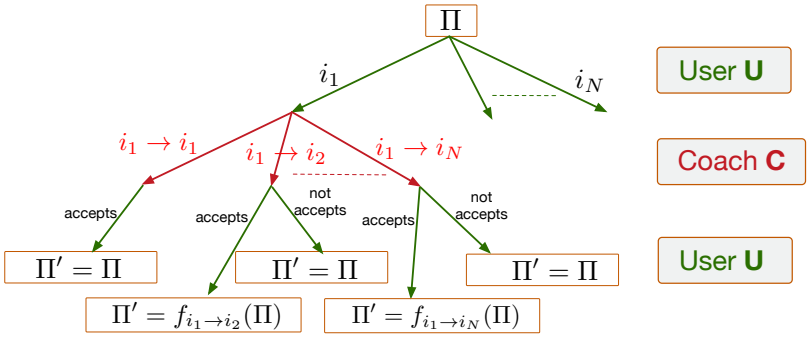

Figure 1: A decision step faced by the coach. Following his/her preference Π_t, the user chooses one item, and then the coach must select a substitution, which, in turn, can be accepted or rejected by the user. After this turn, the preference vector is updated.

These equations require that the coach knows the matrix $\mathbf{M} = [m_{i,j}]$ $(1 \leq i, j \leq |\mathcal{I}|)$ and the current preference vector Π of the user.

From this ideal optimization criterion, we can derive heuristic ones, and the ensuing strategies, where simplifications are made or factors are ignored.

Strategy Greedy Score (GS)

The simplest strategy for the coach is to ignore the characteristics of the user altogether and, at each interaction, to suggest the substitution $i \longrightarrow j$ associated with the highest score gain: $s(j) - s(i)$.

$$j^\star \;=\; \underset{j \in \mathcal{I}}{\text{ArgMax}} \big[s(j) - s(i) \big] \tag{5}$$

Strategy Greedy Expected Score (GES)

A second strategy considers the $m_{i,j}$ values, but ignores the possible change in the preference vector of the user (i.e. $f_{i \to j}(\Pi_t) = \Pi_t$), which gives:

$$j^\star \;=\; \underset{j \in \mathcal{I}}{\text{ArgMax}} \Big\{ m_{i,j} \, \big[s(j) - s(i) \big] \Big\} \tag{6}$$

We call the corresponding strategy *greedy-expected-score* (GES) because it does not consider rewards beyond the immediate one.

Strategy Greedy Acceptation (GA)

This strategy maximizes the probability of the user accepting $i \longrightarrow j$ as long as the corresponding change in score is positive or null: $s(j) - s(i) \geq 0$. In this way, it is hoped that the user changes his/her behavior more easily and that, in the longer term, this will overcome a lack of high gain in the short term.

$$j^\star \;=\; \underset{j \in \mathcal{I}}{\text{ArgMax}} \Big\{ m_{i,j} \;\big|\; \big(s(j) - s(i) \big) \;\geq\; 0 \Big\} \tag{7}$$

Both the GA and the GES strategies maintain an estimation $\widehat{\mathbf{M}}$ of the matrix $\mathbf{M}$ based on the interactions with the user. More specifically, each element $m_{i,j}$ of the matrix is evaluated using the following equation:

$$\widehat{m}_{i,j}^{t+1} = \begin{cases} \dfrac{\widehat{m}_{i,j}^{t} + 1}{n_{i,j}} & \text{if the substitution } i \longrightarrow j \text{ is accepted} \\ \dfrac{\widehat{m}_{i,j}^{t}}{n_{i,j}} & \text{otherwise} \end{cases}$$

with $\widehat{m}_{i,j}^{t}$ the current estimate of $m_{i,j}$ and $n_{i,j}$ the number of times the recommendation $i \longrightarrow j$ has been proposed to the user.

Informed strategies for GA (iGA) and for GES (iGES)

In order to test the effect of a more or less precise knowledge of the matrix $\mathbf{M}$, we considered fully informed versions of the GA and GES strategies: **iGA** and **iGES** who have a perfect knowledge of $\mathbf{M}$ and do not have to learn it.

All of *the above strategies are myopic*, in that they do not explicitly take into account the gains that a substitution $i \longrightarrow j$ can bring in the long term. They do not try to estimate the values $V_t(\Pi)$, a feat that indeed requires the exploration of the possible consequences of the choice j to learn $V_t(\Pi)$ $(\forall t)$.

We thus introduce a *reinforcement learning* (as defined in [12]) algorithm in order to assess the merit of estimating longer term gains when choosing a suggestion of substitution. This type of approaches have been popularised in recommender systems to tackle the sequential nature of recommendation [1, 3]

Q-learning

The equation that evaluates the merit of suggesting j when the user has chosen i and accepts the proposed substitution is:

$$Q(i,j) \;\longleftarrow\; (1-\alpha)\, Q(i,j) \;+\; \alpha \Big\{ \big(s(j) - s(i) \big) + \gamma \underset{k \in \mathcal{I}}{\text{Max}}\; Q(j,k) \Big\} \tag{8}$$

where α controls the learning rate, and γ is a discount factor used to value short-term gains more than longer-term ones.

The Q values gradually reflect the long term potential of the choices of substitutions. When the user selects the item i, the coach suggests the item $j^\star$ according to:

$$j^\star \;=\; \underset{j \in \mathcal{I}}{\text{ArgMax}} \big\{ Q(i,j) \big\} \tag{9}$$

It is important to note that this strategy does not directly use knowledge of the acceptability matrix $\mathbf{M}$. On the one hand, this avoids the necessity to estimate it, and is therefore a more general approach to the coaching problem. On the other hand, this usually has to be paid by a longer learning phase.

tQL-5000 and tQL-10000 We considered two additional strategies: tQL-5000 and tQL-10000, which have been pre-trained for $N' = 5{,}000$ and $N' = 10{,}000$ iterations respectively with a prototypical

user as if they had benefited from past interactions with many more users, which would likely be the case for a realistic coach. They are then used as coaching strategies in our experiments (see Section 4).

Item-Based Collaborative Filtering strategy (IBCF)

A baseline strategy is the one used in a standard recommendation scenario: the item-based collaborative filtering strategy [10]. In this approach, a similarity $sim(\cdot,\cdot)$ between the items is precomputed using the expressed choices of the users (e.g. food item consumption). Then, when the user selects an item i, the recommending system suggests the item $j^\star$ according to equation:

$$j^\star = \underset{j \in \mathcal{I}}{\text{ArgMax}} \left\{ \frac{\sum_{n \in \mathcal{I}} \left(sim(j,n) * R_{u,n} \right)}{\sum_{n \in \mathcal{I}} sim(j,n)} \right\} \quad (10)$$

where $R_{u,n}$ is the rating of item n by user U (here, this rating is estimated by the consumption frequency of n by U) and the similarity is computed as usual in recommender systems ([10]).

Item-Based Collaborative Filtering strategy with score (IBCFs)

A natural question is whether a classical recommendation strategy, such as IBCF, could be tweaked in order to make recommendations aimed at changing the behavior of the users. One simple way to do so is to modify Equation 10 to include the score gain associated with a recommendation:

$$j^\star = \underset{j \in \mathcal{I}}{\text{ArgMax}} \left\{ \frac{\sum_{n \in \mathcal{I}} \left(sim(j,n) * R_{u,n} \right)}{\sum_{n \in \mathcal{I}} sim(j,n)} * \left(s(j) - s(i) \right) \right\} \quad (11)$$

In this way, IBCFs will tend to recommend to user U substitutions j in $i \longrightarrow j$ that are closed to the ones already consumed by U and that bring as much gain as possible.

4 EXPERIMENTAL EVALUATION

Given the coaching scenario several questions arise.

(1) *What can achieve a coach which does not take into account the characteristics of the user?* Do these types of strategies fare significantly worse than strategies that adapt to the users? One can distinguish here strategies like IBCF and IBCFs that do take into account the past consumption of the user but nothing else about U, and strategies like GA and GES and their variants that maintain an estimate of the acceptability matrix **M** of U.

(2) Considering *strategies that explicitly take into account (an estimate of) the matrix* **M** and possibly the learning rate λ of the user, what are the best ones? And what is the sensitivity of the performance attained in function of the quality of the initial estimation of these characteristics? Here, we carried out experiments with simulated users with various profiles and with different initial estimates of **M**.

(3) Finally, *how myopic strategies*, like GA and GES, that try to maximize only the immediate gain, *fare against non myopic strategies* like Q-learning, but which do not explicitly maintain an estimate of the characteristics of the users? We may expect that the second will prevail, but at the price of lots of training. Do the experiments confirm this?

To answer these questions, we specially look at (*i*) the *mean gain in performance* with respect to the number of interactions, (*ii*) the *recommendation rate* of the coach: it should be decreasing after a while when the user does not have anything more to learn from the coach or he/she does not accept his/her suggestions, and (*iii*) the *acceptance rate* of the suggestions by the user: it should increase as the coach learns the user's characteristics.

4.1 The Experimental Protocol

The experiments simulate interactions between the coach, using a given strategy, and users characterized by their matrix **M**, a propensity to learn λ and a starting preference vector Π_0 over the available items. In order to have simulated users with realistic characteristics, we derived the latter from real data in the field of nutrition as is explained below.

We considered different user profiles, different strategies for the coach, and different initialization settings. In our experiments, the number of interactions was set to $N = 2000$ in order to measure the long term trends of each coaching strategy. The results show that most effects are already obtained after 500 interactions or less, which is realistic for a coaching scenario. All results are obtained from 200 simulations for each situation.

4.2 The Simulated Users

The Individual and National Food Consumption Survey (INCA2) database provides a snapshot of the food consumption habits of the population of metropolitan France gathered between 2006 and 2007[1]. From this population, we retained only the adults since children are not the main target for food coaching. This resulted in a database containing the consumption of 2552 users and 365,621 registered meals. We further focused on the choice of one dessert by consumers o, among 267 possibilities, so as to satisfy the assumption that only one item is chosen at each time step instead of several as would be the case with complete meals.

In order to get a rather homogeneous set of users, we selected women (who represent more than 80% of the respondents in the survey) over 20 years of age, yielding 1497 users. This group was split into two sub-groups: women with "bad" nutritional habits (i.e. with an average score in the lower third among women), and one with "good" habits (the top third). For each of these sub-groups, a matrix **M** was estimated (see below), representing to which extent the corresponding users were ready to accept to substitute one item by another one. We also estimated their preference vector Π_0.

We set $\lambda = 0.2$. We found this value to be reasonably representative of the change of habits under the suggestions of a coach, but above all, our experiments show that λ mainly controls the speed but not so much the overall behavior of the evolution of the users. Results for $\lambda = 0.5$ and $\lambda = 0.9$ are available in the supplementary material [15].

Computing the matrix M of substitutability acceptance rates

We estimate **M** directly from the database of food consumptions by a set of users following the proposition of [2]. Their hypothesis is that two items are highly substitutable if they are consumed in similar contexts, but not together (e.g., butter can be substituted to margarine since they are consumed in similar contexts, but, usually, not consumed together).

[1] See https://www.anses.fr/en/content/anses-food-consumption-data-made-available-open-data

Let us thus denote, for an item i, the context set C_i as the set of contexts in which i is a substitutable item. If $|C_i|$ is high, then i is substitutable in many contexts.

For two items i and j, the intersection of C_i and C_j: $|C_i \cap C_j|$ provides an estimate of the number of contexts in which either i or j can be found. If $|C_i \cap C_j|$ is high, then i and j are consumed in similar contexts. Denoting by $A_{i:j}$ the set of contexts of i where j appears:

$$A_{i:j} = \{c \in C_i | j \in c\} \tag{12}$$

The cardinality of $A_{i:j}$ denotes how j is associated to i.

Taking into account these considerations, the authors of [2] propose the following score inspired from the Jaccard index :

$$m_{i,j} = \frac{|C_i \cap C_j|}{|C_i \cup C_j| + |A_{i:j}| + |A_{j:i}|} \tag{13}$$

The score equals 1 when i and j appear in exactly the same contexts and de facto $A_{i:j} = A_{j:i} = \emptyset$. If i and j are never consumed in the same context, the score equals 0. The higher $|A_{i:j}| + |A_{j:i}|$ is, the higher the association of i and j and the lesser the score $m_{i,j}$.

Even though the INCA2 database represents a large survey, still uncommon in food consumption studies, it is nonetheless limited in scope. As a result, the matrix $\mathbf{M}$ computed from it using Equation 13 is sparse and does not fully represent the true propensity of users to accept suggestions of substitutions. In order to remedy this, we took into account not only the score between items, but also the score computed from higher level categorization of food items in INCA2 (e.g., *chocolate brownie* belongs to the *cakes* category and the *pastries and cakes* super-category). We added the score computed for the items to the score computed for their category and their super-category to obtain the substitutability from an item to another. For both the GA and GES strategies, we also set a threshold τ such that if $m_{i,j} < \tau$, the substitution $i \longrightarrow j$ is not proposed. On the following experiments, we set $\tau = 0.05$.

4.3 The Nutritional Score

In this paper, we assumed that a score could be assigned to each food. For this, we used the nutritional score designed by Rayner and colleagues [9]. In our case, we used a mapping from each food item present in the INCA2 database to the nutrients registered in the Ciqual food composition database[2], to compute a score for each food item.

4.4 The Results and their Analysis

The main characteristics of the coach's strategies are summed up in Table 1. Below, we assess the influence of these characteristics on the observed results.[3]

Overall results (see Table 2 and Figure 2)

Table 2 provides a comparison of the benefits for the user of using the various coaching strategies when interacting 2000 times. The mean value of $\mathcal{V}(\Pi_t)$ for $0 \leq t \leq T$ and the standard deviation for each situation are reported for 200 simulations. Several conclusions appear. *First*, the potential gains for the users are a function of the quality of the starting habit. As can be expected, the higher the initial quality (e.g. *Good tier* of the consumers), the lower the potential gain (see also Figure 2). *Second*, non guided strategies, like IBCF which does not take into account the nutritional score, cannot guide the user towards better habits. Even IBCFs, which does take into account the nutritional score, is inefficient because it does not consider the acceptability of substitutions by the user. GS which only looks at the potential score's gain is also very inefficient. Conversely, GA, which only looks at acceptability and suggests only positive substitutions, but does not consider the value of these substitutions, is surprisingly good, and even better than GES on the *Bad tier* consumers. One reason may be that it tends to favor any positive move of the user, and this may accelerate changes of behavior in the right direction as compared with GES which tends to select the best suggestions, perhaps at the cost of their acceptability. On *Good tier* consumers, the starting preference vector of the users is better and GES overcomes GA. *Finally*, Q-learning is good if it has benefited from previous training (see tQL-10000), and poor otherwise, which is not surprising given that Q-Learning starts with no explicit knowledge about the user. Most remarkably, tQL-10000 outperforms even iGES which starts with a perfect knowledge of the matrix $\mathbf{M}$ of the user. This is due to the non-myopic character of Q-Learning.

	Knowledge of nutritional score	Takes ∪ into account	Explicit estimation of $\mathbf{M}$	Informed	Myopic
IBCF	-	√	-	-	√
IBCFs	√	√	-	-	√
GS	√	-	-	-	√
GA	inc.	√	√	-	√
iGA	-	√	√	√	√
GES	√	√	√	-	√
iGES	√	√	√	√	√
Q-Learning	√	indirectly	-	-	-
tQ-Learning	√	indirectly	-	pre-trained	-

Table 1: ***Main characteristics of the coach's strategies.***

The behavior of the strategies (see Figures 2 and 3)

(1) Regarding the *recommendation rate* (see Figure 3), one can note that the worst strategies: QL and GES, which both have to explore possible recommendations in order to learn from the user, keep a high recommendation rate, whereas the better strategies iGES, tQL-5000 and tQL-10000 tend to quickly not to have to make recommendations since the user is rapidly improving his/her behavior.

(2) Regarding the *acceptance rate* by the users (see Figure 3), it is interesting to see that iGES, which is fully informed about the matrix $\mathbf{M}$, and tQL-5000 and tQL-10000, which have been trained, have the highest acceptance rate by far. And they tend to keep it that way during the 2000 iterations, while the poorly informed strategies GES and QL make recommendations that are rarely followed by the user. It may appear that iGES, with its highest acceptance rate

[2] see https://ciqual.anses.fr

[3] The reported results have been obtained on simulated users as explained. Work is under way to replicate the experiments with real users, specifically in university restaurants.

Consumer prototype		GES	GA	IBCFs	IBCF	GS	QL	iGES	tQL-10000
Good tier	μ	2.91	2.57	1.56	1.67	0.56	-0.27	4.38	**4.49**
	σ	1.57	1.42	1.63	1.64	1.50	1.40	1.14	1.04
Bad tier	μ	3.24	3.32	1.50	1.53	0.36	0.64	7.55	**8.25**
	σ	3.11	2.48	2.80	2.86	2.84	2.41	2.29	2.46

Table 2: Table of the mean μ and standard deviation σ of the expected score (Eq. 1): $\mathcal{V}(\Pi_{T=2000}) - \mathcal{V}(\Pi_0)$, for *Good tier* and *Bad tier* consumers depending on the coaching strategy.

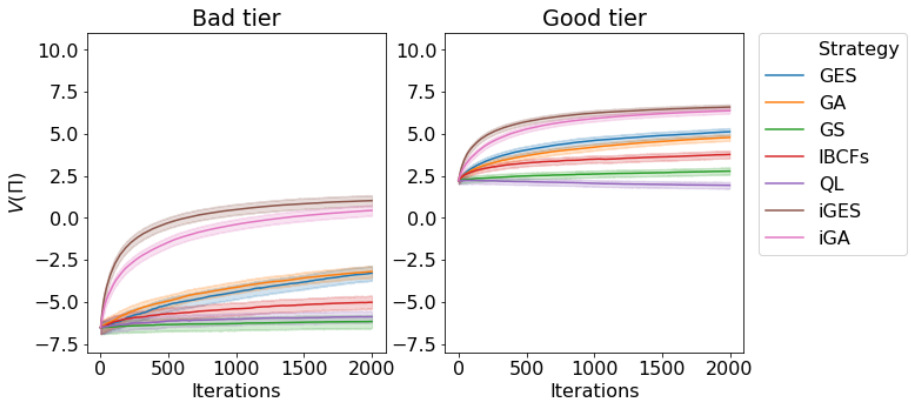

Figure 2: Comparison of $\mathcal{V}(\Pi_t)$, $(0 \leq t \leq T = 2000)$ for two informed strategies (iGA and iGES) and five uninformed strategies (GA, GES, IBCFs, GS and QL) for both *Bad tier* (left) and *Good tier* (right) prototype users. The colored area around the curves represent the 95% confidence interval.

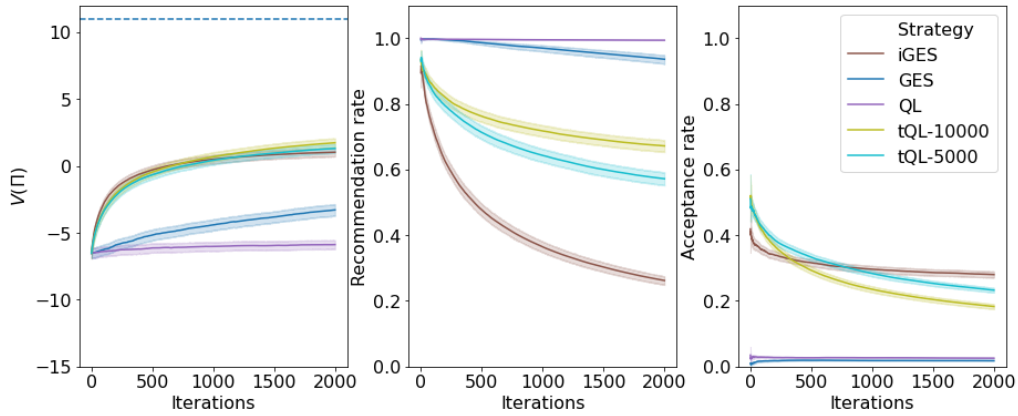

Figure 3: Comparison over 2000 interactions of $\mathcal{V}(\Pi_t)$, $(0 \leq t \leq T = 2000)$ (left), the recommendation rate (center) and the acceptance rate (right) for GES, Q-Learning, iGES and trained Q-Learning on 5,000 and 10,000 steps, for *Bad tier* users. The colored area around the curves represent the 95% confidence interval.

than tQL-2000 and tQL-10000, is better. But this is an illusion. Indeed iGES makes less recommendations, and the fewer remaining recommendations are well accepted by the users since iGES knows **M** perfectly. Conversely, the strategies tQL-5000 and tQL-10000 evolve over time and they explore a larger space of choices by the users. Hence the recommendation rate stays high, but because these strategies do not have a perfect knowledge of **M**, the acceptance rate of the more adventurous recommendations falls down more rapidly than iGES.

Influence of having prior knowledge of the user (see Table 2 and Figures 2 and 3)

One important question is whether prior knowledge by the coach about the user brings a significant gain in the user's performance. Experimental results show that *it is very beneficial to have a good*

prior knowledge of the user's characteristics. While it can be expected that the adaptive strategies GA and GES tend to the performances of iGA and iGES for a large number of interactions, for less than 2000 interactions, the difference in performance is striking. The same effect can be seen for Q-learning algorithms. The pre-trained tQ-Learning algorithms show increasing levels of performance when the number of interactions in pre-training goes from 5,000 steps to 10,000 steps.

It must be noticed that in a realistic setting, a digital coach will benefit from interactions with thousands of users simultaneously, which will provide knowledge about prototypical users, and will thus result in high quality prior knowledge. It can thus be expected that the performances obtained will tend to the higher end of the spectrum of possible performances.

Myopic vs. non myopic strategies (see Figure 3)

It is expected that non-myopic strategies, like Q-Learning, outperform myopic strategies. The question is by how much. Our results confirm the advantage of these strategies. In the case of Q-learning, pre-training permits to significantly overcome the performance obtained with iGES which has perfect knowledge about the matrix $\mathbf{M}$ of the user. This is a remarkable feat given the high level of performance exhibited by iGES. We also noticed that the advantage of non-myopic strategies is even larger with a higher user learning rate λ (see Supplementary material [15]). In the context of a digital coach, these results advocate the use of reinforcement learning algorithms.

5 RELATED WORKS

Even though the literature on Recommender Systems has become quite large, there are few studies on recommendations aimed at changing behavior in a lasting way. Among the latter, one can distinguish between those that aim to broaden the user experience, as in the case of music recommendation [7], and those whose goal, like ours, is to change an individual's behavior, such as security [5], physical activity or healthy food consumption.

Because one common underlying assumption is that changing habits is difficult, if not painful, the main concern of the existing various studies is how to suggest recommendations that will be followed by the user. Hence, [16] makes a distinction between the "responders" and the "non responders" and proposes to make recommendations corresponding to small, incremental and achievable goals accordingly. Farrel et al. [4] explicitly seek lifestyle and behavior changes. They propose to look for stable patterns of behavior in the past history of the user and to look for associations of these patterns with profitable and unprofitable behaviors, so that recommendations can be made when a user exhibits a plateau or a drop in the performance criterion. In [14] or [13], it is mentioned that recommendations can take the form of substitutions, but in both works these take place inside recipes when an ingredient (e.g. butter) is replaced by another one (e.g. margarine). Other works rely on the Rasch scale to balance the engagement and the motivation of the user [8, 11]. Following this model, engagement is maximized by proposing very feasible behaviors while motivation is maximized by matching the difficulty of the advice with the ability of the user.

In all these studies, the evaluation of the effectiveness of the recommendations is done either through classical measurements like acceptance rate, bookmarks or comments (e.g. for recipes), ratings, purchases and queries or through so called user-centric evaluation metrics such as the pleasure or easiness felt by the user while using the recommending system [6]. But all these measurements are immediate indicators that do not provide information about the long term effects of the recommendations, and therefore do not capture the actual changes of habits that they may induce. By contrast, the criteria we define in Section 2.3 offer ways to measure these.

Another line of research related to our work deals with *teacher-student models*. The general idea is that both the student and the teacher are reinforcement learners. The teacher is charged with the task of teaching another agent, the student, to perform a particular task (see for instance [17, 18]). There are common features with the coaching setting: in both, the student announces its intended action and the teacher "corrects" it if the action seems non optimal, furthermore, the teacher tries to learn a teaching policy while interacting with the student. Also, the works cited put forward the notion of a budget that can be spent by the teacher in the form of a limited number of interactions with the student. However, there are differences too: in coaching, we suppose that the user does not change his/her preferences independently of the interactions with the teacher while he/she can learn by himself in the teacher-student framework. In addition, in coaching, the coach does not know the optimal strategy to solve the task (e.g., reach the healthiest diet), but only knows a score function over the items, and can only makes suggestions of changes.

6 CONCLUSION

This paper has presented a new recommendation scenario in which the objective is to *sustainably modify the user's preferences* instead of eliciting instantaneous and independent actions, such as buying items on a website. We have proposed a formalization of this scenario, that we call *coaching*, as an iterated two-player game where the user expresses a choice and the coach may suggests a modification. From its mathematical analysis, we have derived several coaching strategies and we have compared them using experiments in the field of dietary choices. We have looked at the gain in performance for the user as well as the evolution of the recommendation and acceptance rates. The main lesson is that non myopic strategies are more effective, provided that they may benefit from some prior exposition to the characteristics of the user, or users with the same profile.

We strongly believe that *coaching* corresponds to a large spectrum of recommendation problems for which this work represents a first step.

In the future, we plan to study contexts where the user has to choose a combination of items, for example the dishes constituting a meal. Indeed, in this context, the score could be non-additive, for instance by taking into account the interactions between the chosen items, or even the history of the choices such as what was consumed over a week.

ACKNOWLEDGEMENTS

This research has been funded by the French National Agency for Research (ANR), grant number ANR-18-CE21-0008.

REFERENCES

[1] Afsar, M.M., Crump, T., Far, B.: Reinforcement learning based recommender systems: A survey. arXiv preprint arXiv:2101.06286 (2021)

[2] Akkoyunlu, S., Manfredotti, C., Cornuéjols, A., Darcel, N., Delaere, F.: Investigating substitutability of food items in consumption data. In: Second International Workshop on Health Recommender Systems co-located with ACM RecSys. vol. 5 (2017)

[3] Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., Chi, E.H.: Top-k off-policy correction for a reinforce recommender system. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. pp. 456–464 (2019)

[4] Farrell, R.G., Danis, C.M., Ramakrishnan, S., Kellogg, W.A.: Intrapersonal retrospective recommendation: lifestyle change recommendations using stable patterns of personal behavior. In: Proceedings of the First International Workshop on Recommendation Technologies for Lifestyle Change (LIFESTYLE 2012), Dublin, Ireland. p. 24. Citeseer (2012)

[5] Guo, S., Ding, L., Zhang, Y., Skibniewski, M.J., Liang, K.: Hybrid recommendation approach for behavior modification in the chinese construction industry. Journal of construction engineering and management **145**(6), 04019035 (2019)

[6] Knijnenburg, B.P., Willemsen, M.C.: Evaluating recommender systems with user experiments. In: Recommender Systems Handbook, pp. 309–352. Springer (2015)

[7] Liang, Y.: Recommender system for developing new preferences and goals. In: Proceedings of the 13th ACM Conference on Recommender Systems. p. 611–615. RecSys '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3298689.3347054, https://doi.org/10.1145/3298689.3347054

[8] Radha, M., Willemsen, M.C., Boerhof, M., IJsselsteijn, W.A.: Lifestyle recommendations for hypertension through rasch-based feasibility modeling. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization. pp. 239–247 (2016)

[9] Rayner, M.: Nutrient profiling for regulatory purposes. Proceedings of the Nutrition Society **76**(3), 230–236 (2017)

[10] Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web. pp. 285–295 (2001)

[11] Schäfer, H., Willemsen, M.C.: Rasch-based tailored goals for nutrition assistance systems. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. pp. 18–29 (2019)

[12] Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)

[13] Teng, C.Y., Lin, Y.R., Adamic, L.A.: Recipe recommendation using ingredient networks. In: Proceedings of the 4th Annual ACM Web Science Conference. pp. 298–307 (2012)

[14] Trattner, C., Elsweiler, D.: Food recommender systems: important contributions, challenges and future research directions. arXiv preprint arXiv:1711.02760 (2017)

[15] Vandeputte, J., Cornuéjols, A., Darcel, N., Delaere, F., Martin, C.: Coaching Agent: Making Recommendations for Behavior Change. A Case Study on Improving Eating Habits (Supplementary material) (Jan 2022). https://doi.org/10.5281/zenodo.5907599, https://doi.org/10.5281/zenodo.5907599

[16] Yürüten, O.: Recommender systems for healthy behavior change. Tech. rep., EPFL (2017)

[17] Zhan, Y., Fachantidis, A., Vlahavas, I., Taylor, M.E.: Agents teaching humans in reinforcement learning tasks. In: Proceedings of the Adaptive and Learning Agents Workshop (AAMAS) (2014)

[18] Zimmer, M., Viappiani, P., Weng, P.: Teacher-student framework: a reinforcement learning approach. In: AAMAS Workshop Autonomous Robots and Multirobot Systems (2014)