Controle Inteligente de Tempo Livre em Tutoria Multissessão: Concepção, Implementação e Avaliação Empírica

Weber Martins[1], **Viviane M. Gomes**[2], **Lauro G. Nalini**[3], **Carolina M. de Carvalho**[3]

[1]Escola de Eng. Elétrica e de Computação – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

[2]Depto de Tecnologia da Informação – Instituto Federal de Goiás (IFGO)
Inhumas – GO – Brasil

[3]Depto de Psicologia – Pontifícia Universidade Católica de Goiás (PUC-GO)
Goiânia – GO – Brasil

weber@eee.ufg.br, viviane@inhumas.ifgo.edu.br, legn@ucg.br

Abstract. *This research proposes the intelligent control of free time (break intervals) in multi-session tutoring. The teaching strategy presents the content in modules with video, exercise, practical suggestion, free time, and revision exercise. Based on the student performance in exercises, the proposed system uses Reinforcement Learning to control pause durations. The experimental group (with intelligent control) has been compared to the control group (where decisions belong to the student). Results have shown significant and equivalent gains in knowledge retention. However, students from experimental group have realized more accurately the role of free time as a component of the teaching strategy.*

Resumo. *Esta pesquisa propõe o controle inteligente de tempo livre (pausas) em tutoria multissessão. A estratégia de ensino apresenta o conteúdo em módulos com vídeo-aula, exercício, sugestão prática, tempo livre e exercício de revisão. Baseado no desempenho do aluno nos exercícios, o sistema proposto utiliza Aprendizagem por Reforço para controlar a duração das pausas. O grupo experimental (com controle inteligente do tempo livre) foi comparado ao grupo controle (onde a decisão pertence ao próprio estudante). Resultados mostram ganhos significativos e equivalentes na retenção de conhecimento. Contudo, alunos do grupo experimental perceberam melhor o tempo livre como componente da estratégia de ensino.*

1. Introdução

Sistemas Tutores Inteligentes (STI) são programas computacionais concebidos para prover instrução personalizada. Tais sistemas agregam os benefícios do ensino individualizado, onde o aluno progride no seu próprio ritmo [Skinner 1972], aos recursos para tratar o aluno de forma única. A partir da interação homem-máquina, métodos de Inteligência Computacional adaptam dinamicamente o conteúdo e/ou o estilo de ensino às características do aprendiz [Murray 1999].

Historicamente, os Sistemas Tutores Inteligentes expõem o material em etapas consecutivas de tutoria, sem pausas [Martins et al. 2004] [Leung and Li 2007]. Conforme ponto de vista comportamental em Psicologia, a pausa influencia o tempo dedicado

ao estudo (sessão de tutoria), sendo a condição de liberdade apreciada pelo estudante [Osborne 1969] [Geiger 1996]. Portanto esta pesquisa introduz e busca controlar inteligentemente o tempo livre entre etapas da tutoria.

O presente artigo apresenta a fundamentação teórica, seguida pela descrição do sistema proposto, o trabalho de experimentação e os resultados obtidos e analisados, encerrando com a seção de conclusão.

2. Fundamentação Teórica

2.1. Aprendizagem por Reforço

Segundo [Kaelbling et al. 1996], Aprendizagem por Reforço é uma técnica adequada ao problema enfrentado por agentes que devem aprender em ambientes dinâmicos através de interações de tentativa e erro. A estrutura (do Inglês, *framework*) da Aprendizagem por Reforço fundamenta-se na relação entre estado-ação-recompensa. Estado, denotado por s, é o contexto de interação entre ambiente e agente a cada momento, onde o agente executa uma ação (a) e, por consequência, recebe uma recompensa (r) do ambiente.

Após explorar as ações, o agente aprende a selecionar boas ações (ou pares estado-ação) para atingir seu objetivo: maximizar o valor total das recompensas recebidas. A cada passo, o agente escolhe alguma ação a partir da política, denotada por π_t, ou seja, pela regra estocástica para seleção de ações como uma função de estados [Sutton and Barto 1998]. Em outras palavras, política é a distribuição de probabilidades das ações num estado. Matematicamente, $\pi_t(s, a)$ é a probabilidade de $a_t = a$ se $s_t = s$.

A probabilidade $\pi_t(s, a)$ resulta de uma função-valor combinada ao critério de seleção específico da política π. Como características essenciais da Aprendizagem por Reforço, funções-valor fornecem o retorno esperado de cada estado (R_t) - a função de recompensas futuras que o agente busca maximizar, conforme mostra a Equação 1, onde T é o passo final no tempo.

Especificações da interface dos passos definem tarefas, episódicas ou contínuas. Em tarefas contínuas, ao contrário das episódicas, agente e ambiente interagem indefinidamente ($T \rightarrow \infty$). Nesse caso, o retorno apresenta recompensas descontadas, valores presentes de recompensas futuras. A taxa de desconto γ ($0 \leq \gamma \leq 1$) minimiza a influência de recompensas incertas, distantes do momento atual. Para tarefas episódicas, $\gamma = 1$ e T é o número de passos.

$$R_t = \sum_{k=0}^{T} \gamma^k r_{t+k+1} \tag{1}$$

Se o sinal de estado do ambiente sintetiza de forma compacta o passado sem degradar a habilidade de predizer o futuro, diz-se que o ambiente satisfaz a Propriedade de Markov e, portanto, define um Processo de Decisão de Markov (do Inglês, Markov Decision Process, MDP). Tal processo constitui prerrequisito essencial para o agente aprender a política ótima, π^*, sendo a predição do estado futuro do agente baseada somente no estado atual e na ação escolhida [Saeb et al. 2009]. A política ótima π^* é resultado de funções-valor ótimas (de estado V^* e de ação Q^*), ou seja, com maior retorno esperado.

Métodos de seleção de ação definem $\pi_t(s, a)$ baseados no valor da ação $Q_t(s, a)$, tais como *Greedy*, ϵ-*Greedy* e *Softmax*. O método *Softmax* determina a política como uma distribuição ordenada de probabilidades a partir das estimativas de ação-valor. Comumente utilizada no *Softmax*, a Distribuição de Gibbs (ver Equação 2) possui o parâmetro τ, temperatura, responsável pela diferenciação das ações. Temperaturas altas tornam as ações equiprováveis, incentivando a exploração extensiva das ações (*exploration*). À medida que a temperatura diminui, as probabilidades das ações diferenciam-se até convergir (exploração intensiva ou *exploitation*), evidência da aprendizagem computacional.

$$\pi_t(a) = \frac{e^{\frac{Q_t(a)}{\tau}}}{\sum_{b=1}^{n} e^{\frac{Q_t(b)}{\tau}}} \tag{2}$$

No paradigma da Aprendizagem por Reforço, a caracterização do problema contribui diretamente na escolha do método de solução: Programação Dinâmica, Monte Carlo e Aprendizagem por Diferença Temporal. A modelagem revela o nível de conhecimento, completo ou incompleto, do ambiente dinâmico. Métodos de Programação Dinâmica exigem completo e acurado modelo do ambiente. Ao contrário, métodos de Monte Carlo e de Aprendizagem por Diferença Temporal aprendem da experiência, mesmo em ambiente totalmente desconhecido [Sutton and Barto 1998].

2.2. Análise Experimental do Comportamento

A Análise Experimental do Comportamento (AEC) é uma disciplina científica surgida em 1938, no contexto da Psicologia como ciência natural, com o objetivo geral de descrever e explicar as interações entre o organismo/indivíduo e o ambiente [Catania 1999].

Nos casos de comportamentos emitidos espontaneamente, as respostas agem ou "operam" sobre o ambiente. Em AEC, recebem o nome de comportamento operante. Ao operar o ambiente, o comportamento sofre as consequências da ação: se a resposta for imediatamente seguida por evento reforçador, será emitida mais frequentemente em circuntâncias similares. Situações de reforçamento aumentam a frequência do comportamento operante [Holland e Skinner 1975] [Skinner 2003].

Premack, em experimentos com macacos *Cebus* e crianças, verificou que atividades com elevada frequência podem funcionar como reforçadores para atividades de baixa frequência. Tal constatação foi formulada como o princípio de Premack [Premack 1959]. Baseado neste princípio, Osborne (1969) contingenciou tempo livre das atividades escolares como consequência para a permanência nos assentos na sala de aula (resposta operante), numa turma de seis alunos surdos. Como resultado, a frequência do comportamento dos alunos de ficarem sentados aumentou, mesmo após a suspensão da contingência (comportamento observado por seis semanas) [Osborne 1969]. Geiger (1996) aplicou o princípio de Premack em turmas de sétima e oitava séries. Caso o grupo/turma terminasse as atividades de 5 a 10 minutos antes do término da aula, era liberado para brincar no *playground*. Evidenciou-se maior disciplina e maior aproveitamento escolar para o grupo experimental, comparado ao grupo controle (aula tradicional) [Geiger 1996].

Frequentemente, crianças e adolescentes utilizam o computador como plataforma de entretenimento [Subrahmanyam et al. 2001]. Baseado nesse cenário, aplicou-se o princípio de Premack como fundamento para o sistema proposto desta pesquisa. No

presente estudo, tempo livre, quando o aprendiz podia ouvir música ou jogar (atividades comumente mais frequentes), foi contingenciado como consequência às atividades de estudo no tutor (atividades comumente menos frequentes).

3. Sistema Proposto

A tutoria constitui a interação entre aluno e sistema tutor inteligente. Neste trabalho, a estratégia de ensino apresenta o conteúdo em módulos, formados por vídeo-aula, exercício, sugestão prática, tempo livre e exercício de revisão. Como fase especial do ensino, o tempo livre possibilita aprendizagem pela experimentação, afinal o aluno pode explorar novas habilidades durante a pausa. A vídeo-aula, de pequena duração, expõe o assunto do módulo através de esquemas, analogias e linguagem apropriada ao aprendiz. Nos exercícios, o aluno prossegue somente após acertar. E, imediatamente antes do tempo livre, o STI propõe uma sugestão prática (procedimento).

Baseado no desempenho do aprendiz, o sistema proposto controla a duração das pausas conforme decisão do agente inteligente. No contexto da tutoria, os estados referem-se aos módulos do curso, as ações são as escolhas do controle (conjunto de ações A, $\{diminuir, manter, aumentar\}$ a duração da pausa) e as recompensas resultam das notas ($N1$ e $N2$) nos exercícios.

Em cada estado s, o agente 1) seleciona uma ação a a partir da política $\pi_t(s, a)$, 2) determina r pela função-recompensa e 3) atualiza o valor da ação $Q(s, a)$, de forma incremental, conforme mostra o algoritmo da Figura 1. A política π fornece, pelo método *Softmax*, as probabilidades de escolha de cada ação, como resultado da Distribuição de Gibbs (ver a Equação 2).

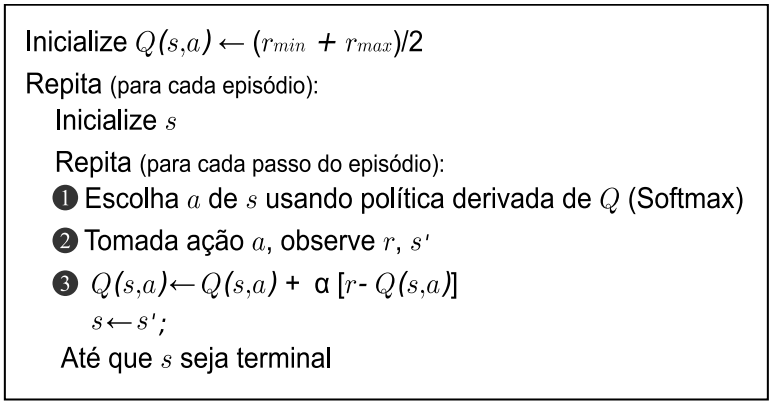

Figura 1. Algoritmo do sistema proposto.

Como o ajuste da temperatura é um ponto crucial para o processo de convergência, este trabalho adota caimento linear da temperatura com razões distintas, q_1 e q_2, para dois momentos da aprendizagem computacional: 1) exploração extensiva e 2) exploração intensiva. Simulações definem as temperaturas inicial, intermediária e final, correspondentes aos estados inicial (referente ao primeiro módulo), intermediário e final (referente ao último módulo), respectivamente. Como base dessas simulações, a modelagem do problema da Aprendizagem por Reforço pressupõe perfis distintos de aluno: teórico, equilibrado e pragmático. Acredita-se que quando o tempo livre diminui, aprendizes teóricos desempenham melhor. Outros aprendizes, equilibrados, obtém mais benefícios quando

o tamanho da pausa é mantido. Por fim, supõe-se que os pragmáticos preferem pausas maiores.

Para encontrar a melhor ação, o agente modifica a duração da pausa a partir do tempo livre inicial tl_0. As ações executadas recebem recompensas do ambiente, definidas pela função-recompensa $r = f(N1, N2)$. Tanto $N1$ quanto $N2$ são valores numéricos para o desempenho do aluno, entre 0 e 1 (de nenhum a total conhecimento da resposta).

Devido à obrigatoriedade do acerto, $N1$ e $N2$ dependem da quantidade de tentativas combinadas ao tipo de resposta (Correta, Semelhante à Correta, Incorreta). Os erros consistem em perdas no valor máximo da nota, como mostra a Equação 3, onde p_i é a perda por tentativa, p_R, a perda de referência (valor da perda total no pior caso) e ta é a tentativa do acerto.

$$Nota = 1 - \frac{\sum_{i=1}^{ta} p_i}{p_R} \tag{3}$$

Para cada tipo de resposta, existe um peso específico no cálculo da perda, definido pela função $H(x)$. Por exemplo, se x for igual à resposta correta, então $H(x) = 0$, pois não há perda neste caso. Calcula-se a perda pela equação $p_i = H(x).\frac{1}{T-i+1}$, sendo *T* o total de alternativas (tentativas possíveis). À medida que o estudante erra, as alternativas escolhidas desaparecem da tela e a perda por tentativa aumenta.

O sistema proposto estima o valor das ações como a média das recompensas observadas. A função-valor é atualizada de forma incremental, pela equação $Q(s, a) \leftarrow Q(s, a) + \alpha[r - Q(s, a)]$, onde o parâmetro α, conhecido como o tamanho do passo (do Inglês, *step-size*), é igual a $\frac{1}{k}$, para k-ésima recompensa recebida para ação a.

4. Experimento

Para testar o sistema proposto, desenvolveu-se o tema "Introdução à Microinformática no Linux". As etapas da tutoria compreendem o conteúdo acrescido de tempo livre. Basicamente, a sessão de ensino contém um questionário ou um módulo do curso, como mostra a Figura 2. Quanto ao tempo livre, sua duração é predefinida para as etapas gerais da tutoria e controlada durante o curso. Baseado em estimativa da duração total da atividade e na pesquisa-piloto, a pausa dura quatro minutos ($tl_0 = 4min$) após as Etapas 1 e 2 e um minuto ($tl = 1min$) após as Etapas 4 e 5.

No experimento, os estados são caracterizados pelos módulos do curso. A sequência de estados $s_1, s_2, s_3, ..., s_{15}$, referente aos módulos, é fixa para o ambiente dinâmico. Quanto às ações, o agente pode decidir por diminuir, manter ou aumentar a duração do tempo livre. No casos de diminuição e aumento, a pausa varia em 25% (a menos ou a mais) do valor anterior. Portanto, numericamente, o conjunto de ações é: $A = \{0,75; 1; 1,25\}$. Assim o controle inteligente escolhe a ação a_t para cálculo do tempo livre, pela equação $tl_t = a_t.tl_{t-1}$, com $a_t \epsilon A$, onde tl_t é o tempo livre atual[1], tl_{t-1} é o tempo livre anterior.

[1]O tempo livre atual possui limites mínimo (1 min) e máximo, estimativa do tempo livre médio disponível por módulo, tendo em vista o tempo livre total estimado (100 min) e o tempo gasto até o momento.

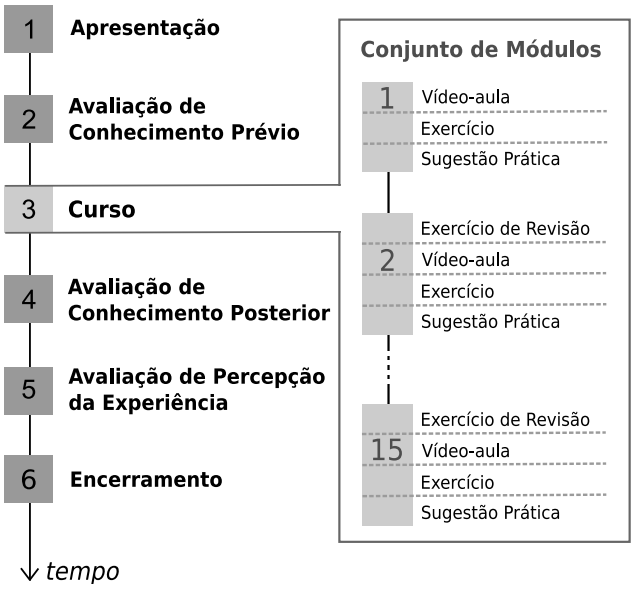

Figura 2. Etapas da tutoria.

Para avaliar as ações (escolhas), o agente utiliza uma função-recompensa, a média ponderada das notas $N1$ e $N2$ nos exercícios, após vídeo-aula e após tempo livre (exercício de revisão). Os pesos de $N1$ e $N2$ são constantes, com ênfase para o momento posterior à ação a_t executada, conforme mostra a equação $r_t = \frac{N1_t + 2N2_t}{3}$.

As notas $N1$ e $N2$ são calculadas a partir da mesma fórmula, apresentada na Equação 3. Obtém-se o valor da perda por tentativa (p_i) pela Equação 4, onde $H(x)$ é a função que caracteriza o tipo de resposta, Correta (C), Semelhante à Correta (SC) e Incorreta (I). Quando a perda total é zero, a nota no exercício é igual a 1 (nota máxima), e vice-versa. Vale ressaltar que os valores de $H(x)$ e p_R foram encontrados empiricamente. Para quatro alternativas e três tipos de respostas, existem nove cenários distintos de tentativas até o acerto. O pior caso definiu o valor da perda de referência em $1,792$.

$$p_i = H(x).\frac{1}{T-i+1}, \quad sendo \quad que \quad H(x) = \begin{cases} 0 & \text{se} \quad x = C \\ 0,5 & \text{se} \quad x = SC \\ 2 & \text{se} \quad x = I \end{cases} \tag{4}$$

A modelagem do problema da Aprendizagem por Reforço, conforme mostra a Tabela 1, possibilitou o ajuste da temperatura. A tabela apresenta as situações de acerto para cada tipo de aluno a partir da ação realizada, ou seja, a sequência de respostas dadas nos exercícios como consequência da ação. As simulações (com 10.000 iterações para cada caso) apontaram para o comportamento desejado da aprendizagem computacional. Para temperatura inicial igual a 2, as ações são praticamente equiprováveis. Houve diferenciação satisfatória da política π para os valores de τ menores que $0,2$. Portanto a temperatura varia de 2 até $0,2$ do Estado 1 ao 8 (s_1-s_8) e de $0,2$ (temperatura intermediária) a $0,1$ (temperatura final) para os estados restantes, s_8-s_{15}, com as razões de caimento $q_1 = -0,257$ e $q_2 = -0,014$.

Tabela 1. Modelagem do problema do Aprendizagem por Reforço.

Ação \ Perfil do Aluno	Teórico	Equilibrado	Pragmático
Diminuir Tempo Livre	C	SC-I-I-C	SC-I-I-C
Manter Tempo Livre	SC-C ou I-C	C	SC-C ou I-C
Aumentar Tempo Livre	SC-I-I-C	SC-I-I-C	C

Tipos de resposta: C-Correta, SC-Semelhante à Correta e I-Incorreta

O STI deste estudo foi implementado em Java e executado durante a coleta de dados em ambiente Linux, distribuição Ubuntu (ver Figura 3). Os participantes (amostra) são adolescentes entre 14 e 17 anos com pouco ou nenhum conhecimento sobre Microinformática. No experimento, a amostra foi dividida em dois grupos equivalentes de 32 estudantes, diferenciados apenas pelo controle aplicado ao tempo livre: inteligente (decisão computacional), grupo experimental, e livre (decisão humana), grupo controle.

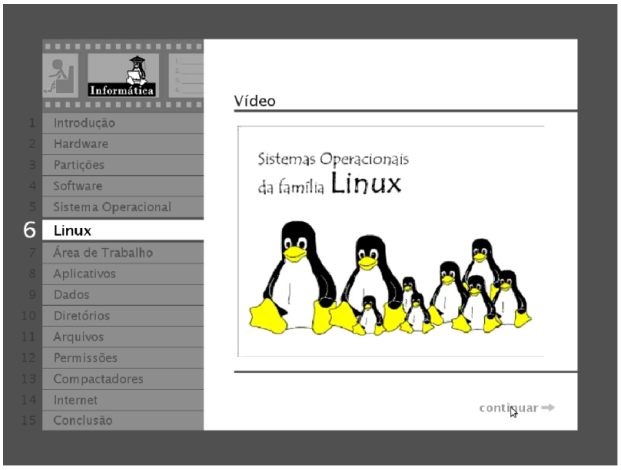

Figura 3. Interface gráfica do STI na vídeo-aula sobre Linux.

5. Resultados

5.1. Análise Estatística Descritiva

Em Sistemas Tutores Inteligentes, a retenção de conhecimento evidencia a capacidade de ensinar de tais sistemas. A Tabela 2 mostra as medidas das notas iniciais e finais, o ganho normalizado (ver Equação 5) e o tempo total da atividade (tutoria com tempo livre).

$$GanhoNormalizado = \frac{NotaFinal - NotaInicial}{NotaMaxima - NotaInicial} \quad (5)$$

Tanto na amostra geral como nos grupos, a nota final apresentou valores maiores que a inicial, sendo o ganho na retenção de conhecimento igual a 45,2%. Os dois tipos de controle, inteligente e livre, possuem comportamentos semelhantes quanto ao desempenho e ao tempo total de tutoria (com tempo livre). Em média, a atividade durou 2h31min no controle inteligente e 2h34min no controle livre. O desvio padrão aponta para heterogeneidade na amostra, justificada pela dificuldades da pesquisa com humanos.

Tabela 2. Estatística descritiva do desempenho e do tempo total (em segundos).

Estatísticas	Média			Desvio Padrão		
Medida \ Amostra	Inteligente	Livre	Geral	Inteligente	Livre	Geral
Nota Inicial	4,13	3,88	4,00	1,55	1,52	1,53
Nota Final	6,71	6,56	6,64	1,95	1,75	1,84
Ganho Normalizado	45,2%	45,2%	45,2%	29,4%	23,4%	26,4%
Tempo Total (s)	9089	9235	9162	1529	1704	1608

No questionário sobre a percepção da experiência, os alunos indicaram a vídeo-aula como etapa mais interessante do curso. No controle inteligente, as outras etapas (exercício, sugestão prática, tempo livre e exercício de revisão) apresentaram distribuição mais uniforme que no controle livre (ver o gráfico da Figura 4).

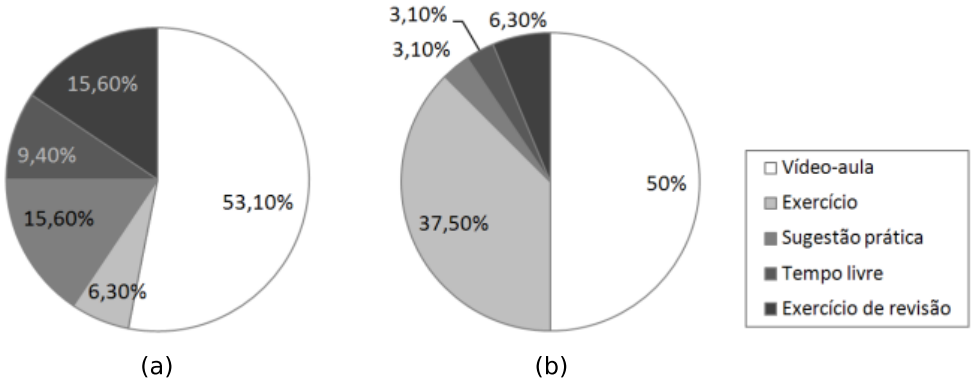

Figura 4. Etapa mais interessante do curso para o grupo (a) inteligente e (b) livre.

A maioria dos alunos considerou o tempo livre suficiente, sendo apontado por 65,6% dos estudantes no grupo experimental e 78%, no grupo controle. De forma unânime, os estudantes gostaram da atividade, sendo que as opções "Gostei." e "Gostei demais." receberam 46,9% e 53,1% das respostas no controle inteligente, respectivamente, e 40,6% e 59,4%, no livre.

5.2. Análise Estatística Inferencial

A análise inferencial generaliza as conclusões a partir de testes estatísticos e estimações. Nesta seção, as provas de hipóteses verificam a presença de diferenças significativas entre as variáveis observadas nos grupos experimental e controle.

O Teste t-*Student* pareado comparou a nota inicial com a nota final dos participantes. A hipótese nula (H_0) foi rejeitada ($\alpha_{efetivo} < 0,01$) tanto na amostra geral como nos grupos, ou seja, houve ganhos significativos na retenção de conhecimento, evidência de que os sistemas tutores realmente atingiram sua meta.

Para analisar o desempenho e o tempo total entre os grupos, aplicou-se o Teste t com amostras independentes. Baseado nos valores do nível de significância efetivo (valor-p) da Tabela 3, a prova de hipótese afirma equivalência entre os grupos, H_0 verdadeira.

Os valores observados sobre a percepção da experiência de tutoria foram insuficientes para a prova de Qui-Quadrado, pois observaram-se frequências inferiores ao ne-

Tabela 3. Estatística inferencial do desempenho e do tempo total.

Estatística \ Medida	Nota Inicial	Nota Final	Ganho Normalizado	Tempo Total
α efetivo	0,52	0,38	0,50	0,36

cessário. Quanto ao gosto pela atividade, com dados suficientes, o teste de Qui-Quadrado evidenciou α efetivo igual a $0,801$, levando à não rejeição de H_0.

5.3. Discussão

No grupo experimental, os estudantes perceberam melhor o tempo livre como componente da estratégia de ensino, conforme observações feitas durante a coleta dos dados e confirmadas pela análise descritiva. O sistema tutor livre permitiu ao aluno controlar a duração da pausa, desviando, contudo, parte de sua atenção para o processo de escolha. Assim, o grupo controle considerou o tempo livre e as etapas imediatamente antes (sugestão prática) e depois (exercício de revisão) apenas como entretenimento.

Quanto à técnica de Aprendizagem por Reforço, a instabilidade do ambiente (interação do aluno) penalizou a aprendizagem computacional. O aluno mudava bruscamente de opinião em relação à duração das pausas conforme a tutoria evoluiu. Apenas em dois casos, o sistema proposto convergiu (desvio padrão da política maior que $0,4$).

No trabalho de [Martins et al. 2007], a Aprendizagem por Reforço foi empregada na seleção de conteúdos "estáticos" para os aprendizes no cenário tradicional de uma única sessão, obtendo ganho normalizado médio de 52,61%. O presente trabalho aliou o uso de pequenas vídeo-aulas e a inserção de tempo livre, atingindo ganhos equivalentes não apenas para o sistema inteligente, mas inclusive para o tutor livre.

6. Conclusão

Esta pesquisa introduz a inserção de pausas em Sistemas Tutores Inteligentes, ressaltando seu forte potencial como componente na estratégia de ensino, aliada ao uso de pequenas vídeo-aulas. As condições do delineamento foram insuficientes para comprovação da hipótese experimental, ou seja, a possível influência do tipo de controle do tempo livre sobre o desempenho do aluno.

Em trabalhos futuros, sugere-se a investigação sobre a eficácia do tempo livre na retenção de conhecimento, comparando STIs com e sem pausas. Em relação aos módulos do curso, novos experimentos poderiam usar a sequência das etapas fundamentada na Aprendizagem Baseada em Problema [Jacinto e de Oliveira 2007]. Sugestões práticas seriam substituídas por problemas, verdadeiros pontos iniciais dos módulos. Para o controle inteligente do tempo livre, a Aprendizagem por Reforço poderia ser modelada com base em comportamento não-verbal, como ansiedade física e expressões faciais [Rodrigues e Carvalho 2005] [Sarrafzadeh et al. 2008].

Referências

Catania, A. C. (1999). *Aprendizagem: Comportamento, Linguagem e Cognição*. Ed. Artes Médicas Sul, Porto Alegre, 4a edição.

Geiger, B. (1996). A time to learn, a time to play: Premack's principle applied in the classroom. *American Secondary Education*, 25:2–6.

Holland, J. G. e Skinner, B. F. (1975). *A Análise do Comportamento*. Ed. da Universidade de São Paulo, São Paulo.

Jacinto, A. S. e de Oliveira, J. M. P. (2007). Uma investigação de STI que emprega a PBL de forma individual. *Anais do XVIII Simpósio Brasileiro de Informática na Educação*, páginas 432–440.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285.

Leung, E. W. C. and Li, Q. (2007). An experimental study of a personalized learning environment through open-source software tools. *IEEE Transactions on Education*, 50(4):331–337.

Martins, W., Afonseca, U. R., Nalini, L. E. G., e Gomes, V. M. (2007). Tutoriais inteligentes baseados em aprendizado por reforço: Concepção, implementação e avaliação empírica. *Anais do XVIII Simpósio Brasileiro de Informática na Educação*, páginas 617–626.

Martins, W., Meireles, V., de Melo, F. R., and Nalini, L. E. G. (2004). A novel hybrid intelligent tutoring system and its use of psychological profiles and learning styles. *Lecture Notes on Computer Science*, 3220:830–832.

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education*, 10:98–129.

Osborne, J. G. (1969). Free-time as a reinforcer in the management of classroom behavior. *Journal of Applied Behavior Analysis*, 2(2):113–118.

Premack, D. (1959). Toward empirical behavior laws: I. Positive reinforcement. *Psychological Review*, 66:219–233.

Rodrigues, L. M. L. e Carvalho, M. (2005). STI-I: Sistemas tutoriais inteligentes que integram cognição, emoção e motivação. *Revista Brasileira de Informática na Educação*, 13(1):20–34.

Saeb, S., Weber, C., and Triesch, J. (2009). Goal-directed learning of features and forward models. *Neural Networks*, 22:586–592.

Sarrafzadeh, A., Alexander, S., Dadgostar, F., Fan, C., and Bigdeli, A. (2008). "How do you know that I don't understand?" A look at the future of intelligent tutoring systems. *Computers in Human Behavior*, 24:1342–1363.

Skinner, B. F. (1972). *Tecnologia do Ensino*. Ed. da Universidade de São Paulo, São Paulo.

Skinner, B. F. (2003). *Ciência e Comportamento Humano*. Martins Fontes, São Paulo, 11a edição.

Subrahmanyam, K., Greenfield, P., Kraut, R., and Gross, E. (2001). The impact of computer use on children's and adolescents' development. *Applied Developmental Psychology*, 22:7–30.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning*. Bradford Book/MIT Press, Cambridge, Massachusetts and London, England.