Avaliação automática de questões conceituais discursivas

Thaís L. T. dos Santos[1], Aleksandra do S. da Silva[1], Eloi L. Favero[1], Adriano D. P. Lino[1]

[1]Universidade Federal do Pará, Departamento de Engenharia Elétrica e de Computação
Belém-PA, Brasil, 66075-110
thais.tavares@gmail.com, {aleka, favero, adrianod}@ufpa.br

Abstract. This paper presents the approach for automatic evaluation of small discursive questions (from 15 to 30 words). The approach is centered in the n-grams usage to measure the similarity between texts. For conceptual questions, the method reaches acceptable results to be used in environment of e-learning: average accuracy about 82.10% Besides treating an unknown subject for Portuguese language, the paper presents two contributions: a) an approach to expand the vocabulary of the professor's answer; b) an approach centered in the normal distribution to map the scores of the answers in concepts through an adjustable scale for different classes.

Keywords: automatic evaluation, discursive conceptual questions, n-grams, similarity between texts.

Resumo. Este artigo apresenta a abordagem para avaliação automática de pequenas questões discursivas (de 15 até 30 palavras). A abordagem é centrada no uso de n-gramas para medir a similaridade entre textos. Para questões conceituais, o método alcança resultados aceitáveis para ser usado em ambiente de ensino à distância: acurácia média de 82.10%. Além de tratar um tema inédito para a língua Portuguesa, o artigo apresenta duas contribuições: a) uma abordagem para expandir o vocabulário da resposta do professor; b) uma abordagem centrada na distribuição normal para mapear os escores das respostas em conceitos numa escala ajustável para diferentes turmas.

Palavras-chave: avaliação automática, questões conceituais discursivas, n-gramas, similaridade entre textos.

1 Introdução

Com a crescente difusão do ensino à distância através da Web, a cada dia é mais relevante o estudo e o desenvolvimento de ferramentas que possam fazer avaliação automática de atividades on-line. Para a língua Inglesa, já existem diversos estudos sobre abordagens de avaliação automática de ensaios [6] e também de pequenas questões [9]. Para ensaios, já existem sistemas de avaliação automática que cumprem

o papel do segundo avaliador (o primeiro é humano) em provas, tais como a GMAT (ver [6]). Numa escala de seis pontos discretos, o avaliador automático, baseado em técnicas de regressão linear múltipla, atinge uma acurácia de até 90% (em relação à avaliação de especialistas humanos) [6]. Para a língua Portuguesa, desconhecemos trabalhos similares.

Este trabalho relata o resultado de um sistema de avaliação de questões de uma disciplina de Introdução a Banco de Dados e Programação SQL on-line, que está disponível através do ambiente virtual AmAm [5] [17]. O problema desafiador está em avaliar automaticamente pequenas questões discursivas, tais como:

***Questão**: Fale sobre um esquema de um banco de dados relacional? (20/30 palavras).*

***Resposta 01**: Um banco de dados é composto de entidades e tais entidades possuem atributos que podem assumir os valores de chave primária, chave estrangeira e chave alternativa. Esses valores são responsáveis pelo relacionamento entre as entidades, pois é através de suas chaves que eles se relacionam.*

***Resposta 02**: Um banco de dados é esquematizado a partir de tabelas que contém no seu interior registros. Tais tabelas se relacionam através de determinados campos que chamamos de chaves primárias e chaves estrangeiras.*

Nesse curso, avaliamos automaticamente questões discursivas e também consultas em SQL. Neste artigo, focalizamos a avaliação automática das questões conceituais discursivas (15 a 30 palavras) baseada num modelo de regressão linear múltipla, desenvolvido a partir de métricas centradas em *n-gramas* de palavras e de caracteres. Na pesquisa foram utilizadas 31 questões e para cada questão foram coletadas mais de **130 respostas, todas de alunos, totalizando mais de 4000 respostas**. Os resultados do estudo mostram que esta abordagem, com certas restrições, já pode ser utilizada em ambientes virtuais de ensino, pois atinge uma acurácia média de 82,10% (para todas as questões e respostas) e para algumas questões a acurácia passa dos 90%.

Essa abordagem apresenta as seguintes vantagens em relação aos sistemas com correção manual: a) *feedback* automático e imediato das questões submetidas; b) permite múltiplas respostas para cada questão, assim o aprendiz pode aprimorar sua resposta em direção à solução ótima; c) permite o instrutor identificar com mais rapidez quais estudantes apresentam dificuldades e, a partir daí, tomar medidas pedagógicas mais eficazes.

Em relação aos trabalhos prévios, o método desenvolvido apresenta duas novidades: a) uma abordagem para expandir o vocabulário da resposta do professor, que é um processo similar à expansão de *consultas*, encontrado na literatura de recuperação de informações [2], porém aqui é aplicado na expansão do vocabulário da resposta do professor; b) uma abordagem centrada na distribuição normal para mapear os escores das respostas em conceitos numa escala ajustável para diferentes turmas.

Além desta introdução, o artigo apresenta mais quatro seções: a seção 2 descreve conceitos relacionados com n-gramas; a seção 3 descreve a nossa abordagem de avaliação automática; a seção 4 discute os resultados, comparando-os com outras abordagens e a seção 5 conclui o trabalho.

2 Os n-gramas

Um n-grama é uma seqüência de n palavras, onde n geralmente é 1, 2 e 3, respectivamente uni, bi e tri-grama. Os n-gramas têm sido usados em sistemas de recuperação de informação como métricas para medir a similaridade de textos [16] [8]. Numa comparação de dois textos, os uni-gramas não avaliam se as palavras dos dois textos ocorrem na mesma ordem. Já bi e tri-gramas avaliam também se as palavras dos textos ocorrem na mesma ordem. Por exemplo, tri-gramas são usados para detectar plágio [8].

Os n-gramas podem também ser utilizados para caracteres e, neste caso, são úteis para recuperar palavras. Se compararmos a palavra *abacaxi* com a palavra *abcaxi*, elas são diferentes. No entanto, elas compartilham certa similaridade. Estas duas palavras compartilham 92% de uni, 80% de bi e 70% de tri-gramas (considerando espaços em branco no início e no fim de cada palavra, como mostrado na tabela 1). Logo, podemos afirmar que estas palavras são bem semelhantes e que a palavra correta do Português mais próxima de *abcaxi* é *abacaxi*. Neste exemplo, os n-gramas de letras servem para medir a distância entre duas palavras. Na tabela 1 são apresentados exemplos ilustrativos de n-gramas de letras.

Os n-gramas de palavras são utilizados em conjunto com duas outras técnicas: *stemming* e remoção de *stop-words* [13]. O *stemming* consiste em tirar partes de uma palavra para se obter a palavra raiz, por exemplo, "*livro*, *livros*" > "*livr*" e "*cantar*, *cantando*" > "*cant*". Assim, o *stemming* possibilita uma representação única para palavras com a mesma raiz (provavelmente com significado similar). Por outro lado, a remoção de *stop-words* simplifica a comparação de dois textos pela remoção de palavras que tem pouco significado conceitual. As *stop-words* são as palavras de classes gramaticais fechadas (artigos, preposições, pronomes, etc). Na Tabela 2 são mostrados exemplos de n-gramas de palavras, com aplicação da técnica de *stemming* e também com remoção de *stop-words*.

Tabela 1. Uni, bi e tri-gramas de caracteres de uma palavra, com a presença e ausência de espaços em branco no fim das palavras.

Palavra	Flores
1. Uni-gramas	[f, l, o, r, e, s]
2. Bi-gramas com brancos	[(32, f), (f, l), (l, o), (o, r), (r, e), (e, s), (s, 32)]
3. Tri-gramas com brancos	[(32, 32, f), (32, f, l), (f, l, o), (l, o, r), (o, r, e), (r, e, s), (e, s, 32), (s, 32, 32)]
4. Bi-gramas sem brancos	[(f, l), (l, o), (o, r), (r, e), (e, s)]
5. Tri-gramas sem brancos	[(f, l, o), (l, o, r), (o, r, e), (r, e, s)]

Tabela 2. Uni, bi e tri-gramas de palavras de uma frase, com remoção de stop-words e com o uso da técnica de stemming.

Frase	As flores amarelas que perfumam os campos
6. Uni	[as, flores, amarelas, que, perfumam, os, campos]
7. Bi	[(as, flores), (flores, amarelas), (amarelas, que), (que, perfumam), (perfumam, os), (os, campos)]

8. Tri	[(as, flores, amarelas), (flores, amarelas, que), (amarelas, que, perfumam), (que, perfumam, os), (perfumam, os, campos)]
9. Uni + sem SW + stemming	[flor, amarel, perfum, camp]
10. Bi + sem SW + stemming	[(flor, amarel), (amarel, perfum), (perfum, camp)]
11 Tri + sem SW + stemming	[(flor, amarel, perfum), (amarel, perfum, camp)]

3 A abordagem de avaliação usando n-gramas

O estudo tem como base uma disciplina de Introdução a Banco de Dados e Programação SQL via Web. Desde o ano de 2005, a disciplina já foi testada em diferentes cursos de graduação (Bacharelado em Ciência da Computação, Sistemas de Informação) e também em cursos de Especialização. Para testar e aprimorar nossa abordagem de avaliação automática, foi coletado uma base com mais de 4000 respostas para as 31 questões discursivas. Desta base, um subconjunto de respostas foi manualmente avaliado por especialistas com um escore de 00 a 10. Cada escore foi também mapeado para um conceito: INS (Insuficiente) <05, REG (Regular) <07, BOM (Bom) <09, EXC (Excelente) <10.

Foram realizados três tipos de experimentos para validar a proposta:

- (a) medir a correlação entre as 11 métricas de n-gramas coletas das respostas e a nota dada pelos especialistas. A correlação é usada porque ela serve para medir o grau de relacionamento entre duas variáveis: resulta num valor entre (–1) e 1; próximo a (–1) para variáveis inversamente correlacionadas, próximo a 1 para variáveis diretamente correlacionadas e próximo de zero para variáveis independentes.
- (b) medir a acurácia de acertos entre os conceitos correspondentes a nota do especialista e diferentes modelos de regressão linear múltipla, criados a partir das melhores métricas; para cada questão foi criado um modelo de regressão linear múltipla diferente;
- (c) idem (b), porém, considerando acertos quando a divergência do valor é menor que um ponto na escala de 00 a 10.

Dentro deste contexto foram desenvolvidos quatro principais estudos: seleção das métricas; criação dos modelos de regressão; expansão do vocabulário; método para predizer o conceito.

Seleção das métricas (n-gramas): O primeiro estudo foi a análise das onze métricas, visando a escolha das melhores métricas para se criar modelos de regressão linear múltipla. Primeiro medimos a correlação entre as métricas coletadas nas respostas dos estudantes contra a nota dada pelos especialistas. Dentre as onze métricas, as que se mostraram mais significativas foram os tri-gramas dos caracteres (C: Tri), uni e bi-gramas de palavras (W: Uni, W: Bi) e uni e bi-gramas de palavras combinados com a técnica de *stemming* (W: UniS e W: BiS). Em todas as métricas de palavras foi aplicada a remoção de *stop-words*. Na Tabela 3 é mostrada a correlação

destas métricas com a nota do sistema, criada a partir de um modelo de regressão linear múltipla.

Criação dos modelos de regressão: A partir das cinco métricas selecionadas, foram criadas duas classes de modelos de regressão linear múltipla: considerando todas as questões juntas e depois considerando cada questão separada. Os modelos com menores erros (que se ajustaram mais) foram para questões individuais. Na Tabela 4 são mostrados alguns resultados de regressão linear múltipla.

Neste estudo, percebemos que a regressão linear tem dificuldade de prever a nota para questões com pouco texto e também para as que realmente possuem um escore baixo dado pelo especialista. Portanto, rodamos novamente os experimentos, mas removendo as notas inferiores a 04. Na Tabela 5, que roda o mesmo experimento da Tabela 4, é mostrada uma melhora de 3.65% quando se eliminam as respostas com escore menor que 04. Como conclusão, os modelos são mais assertivos para predizer o que está correto. Por outro lado, se o modelo é utilizado para predizer o conceito, então este problema das notas baixas não é significativo, pois toda nota abaixo de 05 é INS (Insuficiente).

Expansão do vocabulário: Após a seleção das métricas e da criação dos modelos de regressão linear múltipla, foram feitos estudos visando aprimorar ainda mais a acurácia. Inspirada na idéia da expansão de consultas via técnica de filtragem colaborativa [10], foi incorporada a etapa de expansão de vocabulário. Ela consiste em comparar a resposta do instrutor com todas as respostas dos estudantes e selecionar as cinco (valor que melhor se ajustou) melhores respostas. Juntando-se o conteúdo destas cinco respostas, monta-se uma resposta modelo com o vocabulário expandido. Na Tabela 4 é mostrado o resultado da expansão do vocabulário, na qual ocorre um melhora significativa de em média 16.27%.

Tabela 3. Contribuição de cada métrica na nota prevista (valores de correlação); contra a resposta do instrutor.

Contra Instrutor	C: Tri	W: Uni	UniS	Bi	BiS	Regressão
Questão 066	0,254814	0,102660	0,196876	-0,040510	0,024639	0,446829
Questão 113	0,582236	0,691850	0,689498	0,553952	0,644100	0,817986
Questão 114	0,416820	0,365143	0,352880	0,191470	0,173047	0,543401
Questão 115	0,615585	0,786815	0,862058	0,250899	0,038817	0,895979
Questão 116	0,053783	0,065565	0,252883	-0,100954	-0,040400	0,443629
Questão 117	0,377184	0,240498	0,457063	0,005899	0,092630	0,584051
Médias	0,383404	0,375422	0,468543	0,143459	0,155472	0,621979

Tabela 4. Contribuição de cada métrica na nota prevista (valores de correlação); contra o conjunto das 5 melhores respostas.

Top 5 Respostas	C: Tri 2	W: Uni 2	UniS 2	Bi 2	BiS 2	Regressão
Questão 066	0,368874	0,153755	0,082465	-0,080747	-0,087746	0,674556
Questão 113	0,716434	0,701096	0,730149	0,444258	0,478063	0,806015
Questão 114	0,612310	0,545979	0,403868	0,246340	0,221834	0,727820
Questão 115	0,621268	0,633288	0,752045	0,518003	0,485122	0,828927

Questão 116	0,578184	0,430774	0,474335	0,138097	0,147357	0,698167
Questão 117	0,446381	0,320383	0,570573	0,159834	0,238803	0,603615
Médias	0,557242	0,464213	0,502239	0,237631	0,247239	0,723183

Tabela 5. Contribuição de cada métrica na nota prevista (valores de correlação); foram eliminadas as respostas com notas inferiores a 4.0.

Sem notas <4.0	**C: Tri 2**	**W: Uni 2**	**UniS 2**	**Bi 2**	**BiS 2**	**Regressão**
Questão 066	0,179595	-0,046321	-0,172922	-0,275552	-0,272066	0,704260
Questão 113	0,321215	-0,271277	-0,304199	-0,310265	0,298845	0,821690
Questão 114	0,124117	0,022603	-0,265743	-0,236027	-0,366826	0,782990
Questão 115	0,526607	0,541897	0,694359	0,414180	0,390536	0,824782
Questão 116	0,271353	0,057341	0,035628	-0,327921	-0,411294	0,822094
Questão 117	-0,277465	-0,372908	-0,003472	-0,100401	0,013670	0,541845
Médias	0,190904	-0,011444	-0,002725	-0,139331	-0,057856	0,749610

Predizendo o conceito: Observamos que se o mapeamento de escore para conceito for o padrão para todas as turmas, algumas turmas como as de Especialização mostram um grande aumento de *notas baixas* (quando comparadas com turmas dos alunos do curso de Bacharelado em Ciência da Computação – que são melhor selecionados pelo vestibular e se dedicam mais para o estudo). Para não penalizar os estudantes da Especialização, vimos ser necessário um ajuste nos limites das faixas que fazem o mapeamento de escore para conceito. Outra alternativa seria fazer uma modelo de regressão para cada categoria de turma: bacharelado, especialização, etc. Porém, nesse caso, o modelo de regressão perde a generalidade.

Inspirados em técnicas de normalização de dados e de classificação da área de mineração de dados [19], propomos o uso de uma curva normal para predizer os conceitos de cada questão por turma. A idéia é coletar as repostas de uma turma, rodar o modelo de regressão e depois o instrutor decide sobre os limites do mapeamento para conceitos. Existem duas possibilidades para esta decisão:

1. utilizando-se uma base de no mínimo 30 respostas para uma turma, com as respostas já pré-avaliadas em termos de conceito; neste caso, o sistema identifica a partir das 30 respostas os limites para cada faixa de valores correspondente a um conceito; ou
2. utilizando-se uma interface, como a da Figura 2, onde o instrutor move os controles deslizantes decidindo a percentagem em cada faixa de conceitos sobre a curva normal.

Estas duas opções decidem os valores E, B, R usados no algoritmo de mapeamento dado abaixo:

```
se      nota>media+E*desvio então EXC;
senão se nota>media-B*desvio então BOM;
senão se nota>media-R*desvio então REG
senão INS
```

Por exemplo, os valores *default* podem ser E=1, B=0.5, R=1 (ver Figura 1 – curva normal com média igual a 7.8, desvio padrão igual a 1.3). No eixo horizontal do gráfico estão presentes os valores obtidos pela média e desvio padrão, que serão as condições para os conceitos quando forem comparados com a nota. Já no eixo vertical do gráfico estão presentes os valores da porcentagem de alunos na turma que tiraram cada um dos conceitos. Esta técnica se mostrou flexível para acomodar o problema de diferentes turmas de alunos possuírem limiares diferentes para os conceitos.

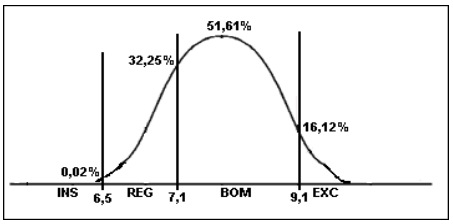

Fig. 1. Curva normal com média 7,8.

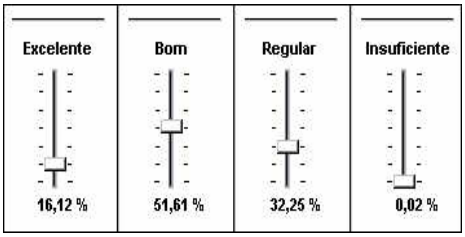

Fig. 2. Controles deslizantes de percentagem dos conceitos.

Quando se opta pelo sistema inferir os limites, o problema é criar uma base de pelo menos 30 respostas por questão para aquela categoria de turma. Uma solução é propor as questões como atividade de exercícios para os estudantes. Durante os exercícios, elas são avaliadas com os limiares de outra turma já existente e nos próximos exercícios ou avaliações o limiar é ajustado. Por exemplo, as mesmas questões dos exercícios podem ser reutilizadas numa prova, fazendo-se um sorteio para cada estudante.

4 Resultados e trabalhos relacionados

Os resultados de nosso estudo podem ser comparados com os resultados de alguns trabalhos relacionados, conforme é mostrado na Tabela 6. Para efeito de comparação, são utilizadas duas medidas:

- correlação de 0.70 entre a resposta do sistema e de especialistas (em nosso caso, obtida com um modelo de regressão múltipla para cada questão); e
- acurácia de 82.10% numa escala de 00-10, considerando casos com diferença menor que 1.

Nos trabalhos relacionados, a acurácia é superior a de nossa proposta, mas a escala deles é de 0-6 enquanto que a nossa é de 0-10. Nos dois casos, considera-se correta a predição quando a diferença do valor estimado contra o valor dado pelo especialista é de até um ponto.

Tabela 6. Alguns sistemas existentes com os resultados da acurácia ou correlação (Acur: a acurácia; Corr: a correlação).

Sistema	Técnica	Resultado	Aplicação
BLEU [12]	n-gramas	Corr 0.95	Nove diferentes dados *benchmark* de exames reais de definições obtidas do *Google Glossary*.
Automark [11]	EI	Corr 0.93	Teste de conhecimento de Java dos estudantes de engenharia do 1º ano da Universidade de Brunel
Auto-marking [18]	Híbrida	Acur 0.88	Exame GCSE* de Biologia
CamelTC [15]	Híbrida	Acur 0.85	Análise qualitativa de questões de Física
C-rater [3]	PLN	Acur 0.80	Compreensão de leitura e exercícios de álgebra
E-rater [3]	Híbrida	Acur 0.87	GMAT*, escrita para não-nativos da língua Inglesa
Esta proposta	n-gramas	Acur 0.82 Corr 0.70	Questões conceituais discursivas sobre Banco de Dados
**GCSE: General Certificate of Secondary Education; GMAT: Graduate Management Admissions Test;*			

Os trabalhos relacionados utilizam diferentes técnicas, tais como (ver [6]): Extração de Informação (EI), Processamento de Linguagem Natural (PLN), Classificação de Textos [4], Análise Semântica Latente [7], etc. Muitos sistemas utilizam abordagens híbridas, combinando duas ou mais das técnicas mencionadas. A abordagem de **Extração de Informação (EI)** consiste em identificar no texto nomes de entidades e relacionamento entre entidades, eventualmente preenchendo *templates*. Assume-se que aqui não é necessária uma análise sintática completa do texto, mas apenas uma heurística para identificar nomes, verbos, etc. A abordagem de **Processamento de Linguagem Natural (PLN)** implica na existência de ferramentas

que fazem uma completa análise sintática do texto, correção morfossintática, identificação de relações de retórica, etc.

Espaço para melhorar: Com base nos resultado da Tabela 6, nossa abordagem se mostra satisfatória, alcançando acurácia de 82.10% e correlação de 0.70 que está próxima destes sistemas. Com objetivo de melhorar nossa abordagem, estamos trabalhando em três frentes: a) no desenvolvimento de um etiquetador para incluir também a informação sintática como uma métrica; b) no aprimoramento do algoritmo de *stemming*, especializando-o para nossa aplicação; c) na coleta e teste de outras métricas em direção a uma abordagem híbrida que melhore a acurácia.

Críticas: Dentro da taxonomia de Bloom [1] sobre avaliação, existem dois grandes grupos de questões: abertas e fechadas. Hoje a grande parte das provas com correção automática é centrada em questões fechadas. Trabalhos, como este, mostram caminhos promissores para tratar de forma automática as questões abertas, porém não podemos esquecer os pontos fracos das abordagens de avaliação automática de questões discursivas. Nossa abordagem é mais bem utilizada para "classificar" as resposta numa escala de conceitos onde as notas mais baixas são todas de uma mesma categoria, por exemplo, qualquer valor abaixo de 05 é insuficiente. Além disso, a abordagem se comparada com métodos que utilizam PLN, ela é fraca para tratar questões não conceituais. A grande vantagem da abordagem é sua generalidade e o baixo custo em treinar a avaliação de cada questão: basta avaliar um conjunto significativo de respostas, tipicamente 40 respostas.

Num processo de avaliação (exercício ou prova) é necessário compensar estes pontos fracos da avaliação das questões abertas, utilizando outros tipos de questões. Podem-se utilizar questões fechadas (tais como falso/verdadeiro e múltipla escolha) e também questões semi-abertas, tais como o desenho de mapas conceituais a partir da escolha de um conjunto de conceitos e frases de ligação. Neste caso, inclui-se na lista de conceitos (e frases de ligação) um percentual de ruído (conceitos que são errados) para testar se o aprendiz sabe optar pelos conceitos corretos [14]. Assim, com uma criteriosa combinação de questões abertas e fechadas pode-se construir um conteúdo para um curso via Web, onde a avaliação é totalmente automática.

Um resultado secundário foi a criação da base de questões e respostas (com mais de 4000). Esta base é essencial para aprimoramento dos métodos de avaliação automática. Serve, por exemplo, para *benchmarking* de diferentes abordagens de avaliadores automáticos.

5 Conclusão

O trabalho apresentou uma abordagem para avaliar automaticamente pequenas questões conceituais discursivas. A abordagem é centrada no uso de métricas coletadas de n-gramas e de um modelo de regressão linear múltipla para cada questão. O modelo proposto atinge uma acurácia de 82.10% e uma correlação de 0.70, próximas dos sistemas encontrados na literatura.

Foi proposta uma abordagem de expandir o vocabulário da resposta do instrutor, selecionando as melhores respostas de uma coleção de respostas. Esta expansão do vocabulário melhora a correlação em média de 16.27%. Também propomos o uso de

uma abordagem centrada na distribuição normal para mapear os escores das respostas em conceitos numa escala ajustável para diferentes turmas.

O sistema se mostrou deficiente para tratar respostas com pouco texto, de crítica pessoal e com frases negativas. Na seção de resultados discutimos alternativas para este problema. Por outro lado, a abordagem dá bons resultados para avaliar textos do tipo descritivo conceitual. Em especial, no contexto de ambientes virtuais de ensino, um sistema de avaliação automática tem duas grandes vantagens: avaliação imediata (*feedback*) e possibilidade dos estudantes responderem a mesma questão várias vezes, sem sobrecarregar o instrutor com o trabalho de corrigir a questão.

Referências

1. Bloom, B. S., et al. Taxonomia dos objetivos educacionais. *Globo*, Porto Alegre (1974).
2. Brandberg, G. Query Expansion using Collaborative Filtering Algorithm. *Uppsala Master's Thesis in Computing Science, Uppsala University*, ISSN 1401-5749 (2001).
3. Burstein, J., Leacock, C. and Swartz, R. Automated evaluation of essays and short answers. In: *Proceedings of the 5th International Computer Assisted Assessment Conference* (July, 2001).
4. Burstein J. et al. Automated Scoring Using A Hybrid Feature Identification Technique. *Proc. Ann. Meeting Association of Computational Linguistics*. Montreal, Canada, 1998; www.ets.org/research/erater.html (current Nov. 2000).
5. Harb, M. P. A. A., Brito, S. R., Silva, A. S., Favero, E. L., Tavares, O. L., Francês, C. R. L. AmAm: ambiente de aprendizagem multiparadigmático. In: *Simpósio Brasileiro de Informática na Educação*, (Novembro, 2003). Rio de Janeiro: NCE-IM-UFRJ. p. 223-232.
6. Hearst, M.A. The debate on automated essay grading. In: *IEEE Intelligent Systems* (IEEE CS Press, 2000). v. 15. p. 22-37.
7. Landauer, T., and Dumais, S. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. In: *Psychological Rev*, (1997) v. 104, p. 211-240.
8. Lyon, C., Barrett, R., and Malcolm, J. Plagiarism Is Easy, But Also Easy To Detect. *Plagiary: Cross Disciplinary Studies in Plagiarism, Fabrication, and Falsification*, 1 (5): 1-10 [temporary pagination for advance online copies of articles]. ISSN 1559-3096 (2006).
9. Marin, D.P. Automatic evaluation of users' short essays by using statistical and shallow natural language processing techniques: Advanced Studies Diploma Work. *Escuela Politécnica Superior, Universidad Autónoma de Madrid (UMA)*. Madrid, España (2004).
10. Massa, P. and Avesani, P. Trust-aware Collaborative Filtering for Recommender Systems. International Conference on Cooperative Information Systems. Larnaca, Cyprus (2004).
11. Mitchell, T., Russell, T., Broomhead, P., Aldridge, N. Towards robust computerised marking of free-text responses. In: *Proceedings of the 6th International Computer Assisted Assessment Conference (CAA)* (July, 2002)
12. Perez, D., Alfonseca, E., Rodriguez, P. Application of the Bleu method for evaluating free-text answers in an e-learning environment. In: *Proceedings of LREC-2004*, (Lisbon, 2004).
13. Richardson, W.C., Avondolio, D., Vitale, J., Len, P., Smith, K.T. Professional Portal Development with Open Source Tools: Java™ Portlet API, Lucene, James, Slide. *Wiley Publishing, Inc*. Indianapolis, Indiana, USA (2004).
14. Rocha, F.E.L. et. al. A new approach to meaningful learning assessment using concept maps: Ontologies and genetic algorithms. In: *Primer Congreso Internacional Sobre Mapas Conceptuales*, Pamplona (2004).

15. Rosé, C., Roque, A., Bhembe, D., VanLehn, K. A hybrid text classification approach for analysis of student essays. *Build Educational Applications Using Natural Language Processing*, p. 68-75 (2003).
16. Salton, G., Wong, A. and Yang, C. A vector space model for automatic indexing. In: *Communications of the ACM*, (1975) v. 18, n. 11, p. 613–620.
17. Silva, A.S.; Brito, S.R.; Favero, E.L.; Hernández-Domínguez, A. Tavares, O.L.; Francês, C.R.L. Uma arquitetura para desenvolvimento de ambientes interativos de aprendizagem baseado em agentes, componentes e framework. In: *Simpósio Brasileiro de Informática na Educação*. 14., (Novembro, 2003), NCE-UFRJ.
18. Sukkarieh, J., Pulmand, S., Raikes, N. Auto-marking: using computational linguistics to score short, free text responses. In: *Proceedings of the 29th Annual Conference of the International Association for Educational Assessment*, Manchester (U.K, 2003).
19. Tan, P., Steinbach, M., Kumar, V. Introduction to Data Mining. *Addison-Wesley* (2005). ISBN : 0321321367.