Processamento de Consultas em Linguagem Natural em Ambientes Virtuais de Aprendizagem

Antônio Luiz M. S. Cardoso
Escola de Administração - UFBA
Av. Reitor Miguel Calmon, s/n
Vale do Canela - CEP:41.100-903
+55-71-3334-8029
antoniol@ufba.br

Teresinha Fróes Burnham
Faculdade de Educação - UFBA
Av. Reitor Miguel Calmon, s/n
Vale do Canela - CEP:40.110-100
+55-71-3334-7229
tfroesb@ufba.br

ABSTRACT
This paper describes a software tool, based on natural language processing (NLP), which compares and identifies similar queries formulated in natural language (*portuguese*). This tool is applied on Educational environments in order to fetch student's queries which has been already answered. If a similar query with an answer is found, it is presented to the user. This tool has been developed applying different NLP resources and techniques, such as: Query expansion, Grammar rules of the Portuguese language, and Thesaurus Relevance.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: *Clustering, Query formulation.*

General Terms
Human Factors , Languages.

Keywords
Virtual Learning Environment.

RESUMO
Este artigo descreve uma ferramenta, baseada em processamento de linguagem natural (NLP), que compara e identifica consultas similares formuladas em linguagem natural (português). Esta ferramenta é aplicada em ambientes virtuais de aprendizagem (AVA) para buscar consultas já respondidas. Caso uma consulta similar com resposta é encontrada, esta é apresentada ao usuário. Ela foi desenvolvida utilizando diferentes técnicas de NLP, tais como: Query expansion, Regras gramaticais do português e Dicionário de sinônimos

1. INTRODUÇÃO
Indivíduos que buscam a mesma informação normalmente utilizam formas lingüísticas diferentes para formular as suas consultas. As diferenças entre consultas podem estar nos termos utilizados, na quantidade dos termos, nas variações sintáticas ou semânticas dos termos. Apesar de serem diferentes na sua formulação, muitas consultas são similares, pois buscam a mesma informação.

Este artigo descreve uma ferramenta que busca e compara diferentes consultas, formuladas em linguagem natural (português), identificando aquelas consultas que sejam similares entre si. Neste trabalho, consultas similares são aquelas que, elaboradas em linguagem natural em um mesmo contexto, buscam a mesma resposta.

A ferramenta é, então, aplicada em um Ambiente Virtual de Aprendizagem (AVA), intitulado Hospital Educacional, o qual possui uma base de perguntas e respostas sobre a temática de novas Tecnologias de Informação e Comunicação.

2. O HOSPITAL EDUCACIONAL
O Hospital Educacional (http://www.hospitaleducacional.com/) é um AVA na Web (Figura 1), que possibilita a construção e a difusão do conhecimento advindo dos atores (professores e alunos) numa sala de aula.

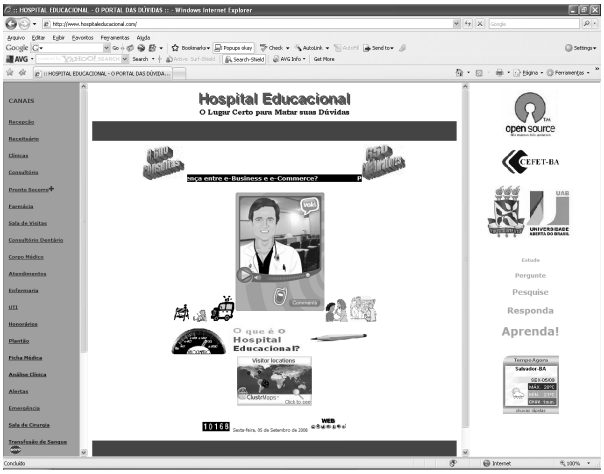

Figura 1. A tela inicial do Hospital Educacional.

Ele congrega professor e alunos com a intenção de ampliar as condições de ensino e aprendizagem pela mediação das Tecnologias de Informação e Comunicação, sendo baseado em uma proposta de aprendizado colaborativo. Através dos recursos e funcionalidades criados para o Hospital Educacional, este proporciona aos alunos:

- Acesso ao material de estudo apresentado em uma sala de aula presencial;
- Espaço para trocar experiências on-line para permitir a interatividade aluno-professor e, principalmente, aluno-aluno;
- Espaço para exprimir dúvidas e questionamentos formulados através de consultas, em linguagem natural (português) sobre o material de estudo para obter orientações e saná-las; e
- Repositório de Objetos de Aprendizagem constituído por fatos, resumos e conceitos, oriundos das Orientações propostas pelos alunos, que podem ser (re)utilizados na confecção de artigos e na preparação de aulas.

Há diferentes conceitos sobre Objetos de Aprendizagem. Objetos de Aprendizagem são definidos como "qualquer recurso digital que pode ser reutilizado para apoiar à aprendizagem" [1].

O repositório de Objetos de Aprendizagem forma a Base de Conhecimentos do Hospital Educacional. Ela é construída pelos alunos de modo colaborativo com acesso aberto a qualquer usuário na Web. A Base de Conhecimentos aceita consultas formuladas em linguagem natural (português) e não em linguagem de manipulação de dados, que exigiria dos usuários conhecimentos em técnicas de programação.

Neste processo de consulta-orientação proporcionado pelo Hospital Educacional, os alunos não apenas buscam conhecimentos para si, através da formulação de consultas, mas também contribuem ao propor orientações para as consultas de seus pares, em um processo de construção colaborativa do conhecimento em que todos os participantes ganham. Ganham demandando informações e, também, propondo orientações aos seus pares.

Além disso, não apenas o conhecimento explícito, contido no material de estudo, é empregado nas orientações às consultas, mas também o conhecimento tácito é devidamente registrado e posto à disposição de todos os alunos atuais e, importante, acessado pelos futuros. Assim, o conhecimento tácito é então compartilhado e articulado através do diálogo (consultas) e da reflexão de seus pares (orientações) e o conhecimento explícito é sistematizado para recuperação futura.

Daí emerge a primeira justificativa deste trabalho: a perspectiva de registrar o conhecimento concebido pelos alunos em um determinado momento e construí-lo para os alunos das turmas seguintes. Pois, é freqüente a realidade de que o conhecimento trabalhado em uma sala de aula num determinado período escolar deva ser novamente (re)construído para os alunos das turmas vindouras, como que perdido ou mesmo ignorado o que foi realizado. Freqüentemente, os alunos das novas turmas desconhecem o que foi trabalhado nas turmas anteriores e, por outro lado, os professores muitas vezes não têm instrumentos eficientes para resgatar o passado.

Além de construir o conhecimento para frente, o ambiente possibilita aos alunos futuros uma outra perspectiva: adicionar novas informações ao conhecimento registrado na Base de Conhecimentos, complementando ou mesmo corrigindo-o.

3. A LÓGICA COMPUTACIONAL

A ferramenta utilizada no Hospital Educacional para processar as consultas e recuperar a informação (orientações) requisitada pelo usuário é estruturada em etapas lógicas seqüenciais, conforme listadas a seguir:

(1) Identificação do tipo da consulta: as consultas são classificadas por tipos baseados no pronome ou advérbio interrogativos contidos nelas, como, por exemplo, Quem, Onde/Aonde, Para que/Por quê, Como, Quanto, Quando, Qual e O que. Estes termos contêm informações importantes e exprimem a necessidade de informação específica sobre a requisição do usuário, como é afirmado em [2];

(2) Substituição de expressões similares: expressões pouco usuais são substituídas por termos mais comuns especialmente pelas siglas, mais conhecidas. Por exemplo, o termo LAN é substituído por 'Rede de Computadores';

(3) Análise Léxica da consulta: identifica e remove símbolos (dígitos, sinais de pontuação, hífens, parênteses e colchetes), além de padronizar as minúsculas/maiúsculas. Sinais de acentuação são especialmente tratados nesta etapa devido a erros gramaticais comuns na escrita;

(4) Remoção das *stop words*: *Stop word* é uma palavra que não carrega significado podendo ser ignorada em um sistema de busca computacional, conforme definido em [3]. Exemplos de stop words podem ser artigos, preposições, conjunções, verbos auxiliares, entre outras;

(5) Separação de nomes próprios: os nomes próprios (pessoas, países, estados, cidades ou organizações) não passam por nenhum tratamento morfológico, pois determinam entidades únicas/exclusivas. Deste modo, eles são identificados e separados dos processamentos subseqüentes até serem efetivamente processados;

(6) Análise Gramatical: os termos das consultas são identificados gramaticalmente. Nesta etapa, são reconhecidas e processadas até 220 regras gramaticais da língua portuguesa, incluindo tempos verbais (passado, presente, futuro), plural/singular, feminino/ masculino, aumentativo/ diminutivo, advérbios, entre outras regras;

(7) Expansão dos termos das consultas (*query expansion*): Em [4], é afirmado que "*Query expansion* é uma ferramenta essencial para recuperação da informação que interativamente recomenda novos termos relacionados a uma particular consulta". É um processo incremental de transformar uma consulta em outra, com novos termos. Os termos das consultas são expandidos utilizando uma lista de sinônimos com aproximadamente 40.000 entradas e 500.000 sinônimos. Esta lista é constantemente revista e ampliada com palavras de idéias afins para torná-la sempre mais coloquial e universal nas áreas de conhecimento da solução, aproximando-a de um verdadeiro Tesauro;

(8) Adição de nomes próprios: nesta etapa, os nomes próprios são adicionados íntegros à consulta para serem processados pela solução;

(9) Seleção da área de conhecimento: através da área de interesse definida pelo usuário, a solução seleciona as consultas relacionadas à requisição do usuário;

(10) Identificação da similaridade: nesta última etapa, através do grau de similaridade definido pelo usuário, a solução calcula matematicamente a similaridade entre as consultas recuperando aquelas que possuem orientações armazenadas na Base de Conhecimentos, similares à consulta do usuário.

4. A BASE DE CONHECIMENTOS

A Base de Conhecimentos do Hospital Educacional armazena dois conjuntos fundamentais:

1. Todas as consultas formuladas pelos alunos; e

2. As orientações, propostas pelo professor, bolsistas e alunos, que passaram pelo processo de validação/aceitação.

Atualmente, a Base de Conhecimentos suporta as seguintes áreas de conhecimento:

- **Redes de Computadores**, incluindo os seguintes tópicos: Conceitos, Classificação, Protocolos, Modelo ISO/OSI, Internet, Cabeamento, Tecnologia *wireless*, Topologia, Dispositivos de conexão (*Hub*, *Switch*, *Gateway*), etc...;
- **Segurança Digital**, incluindo: Conceitos, Vírus, Spyware, Firewall, Antivírus, Hacker, Cracker, etc...;
- **Banco de Dados**, incluindo: Sistema Gerenciador de Banco de Dados, Tipos de Banco de Dados, Modelo Relacional (Tabelas, Registros, Atributos), Relacionamento, Tipos de Chaves (Primária, Secundária, Estrangeira), etc...;
- **Engenharia de Software**, incluindo: Linguagens de programação, Ferramentas, Certificações, CMM, etc...;
- **Sistemas Operacionais**, incluindo: Conceitos, Classificação, Tipos de Processamento, MS-Windows, Linux, etc...;
- **Sistemas de Informação**, incluindo: Conceitos, Evolução, Tipos de Sistemas, Software Livre, Categorias de Profissionais (Analista de Sistemas, CIO), CRM, SCM, ERP, *Business Intelligence*, Data Mining, eCommerce, eBusiness, MS-Project, etc...; e
- **Outras**, incluindo todas as orientações que não se encaixam nas opções anteriores, mas que tenham cunho em Tecnologia da Informação e Conhecimento.

Estas áreas de conhecimento fazem parte do conteúdo das disciplinas de Administração de Sistemas de Informação e Sistemas de Informação Gerencial da Escola de Administração da UFBA.

Todavia, estas áreas de conhecimento, disponíveis no Hospital Educacional, não são exclusivas. Ou seja, outras áreas podem ser adicionadas à Base de Conhecimentos a depender apenas de um professor que monitore as consultas e valide as orientações nessa outra área de conhecimento a ser adicionada.

5. A INTELIGÊNCIA DA SOLUÇÃO

A "inteligência" da solução é incremental, não intrínseca. Isto porque a solução somente é capaz de fornecer uma Orientação a uma Consulta de um aluno caso uma Consulta similar já tenha sido respondida anteriormente e a Orientação esteja armazenada na Base de Conhecimentos.

Caso não haja uma Orientação, o aluno espera até que outro colega proponha uma válida. Esta nova Orientação incrementa a "inteligência" da solução (a sua habilidade em atender a uma nova requisição similar).

A nova requisição, a ser respondida, não necessariamente deve ter os mesmos termos para ser respondida, como ocorre em um FAQ. A nova requisição precisa apenas conter termos sinônimos a uma Consulta previamente respondida.

Deve-se ressaltar que o ponto da inovação tecnológica não é apenas devido aos recursos de Processamento de Linguagem Natural construídos na solução. Mas, sobretudo na aplicação da solução juntamente com as práticas pedagógicas elaboradas neste trabalho que possibilitam a interação colaborativa entre os alunos para a construção e difusão do conhecimento.

5.1 Um Exemplo

Para clarificar o funcionamento da ferramenta, a seguir é apresentada uma seqüência de três figuras que representa momentos distintos a demonstrar como ocorre o processamento das Consultas no ambiente. As Consultas e a Orientação, deste exemplo, são dados reais extraídos da Base de Conhecimentos.

Na Figura 2, representando um primeiro momento, o aluno W. Troelsen formulou uma Consulta em 28 de nov. de 2006 no Hospital Educacional. Nota-se que a Consulta foi formulada em português coloquial (linguagem natural), inclusive contendo um erro gramatical (ausência do verbo).

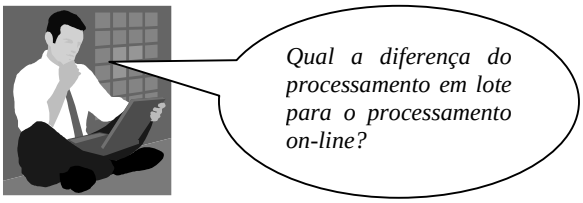

Figura 2. Aluno formula uma consulta no Hospital Educacional.

Para essa Consulta, o Ambiente não propôs automaticamente nenhuma Orientação, pois não havia uma Consulta similar a ela com Orientação armazenada na Base de Conhecimentos. O aluno teve que aguardar algum colega propor uma Orientação válida à sua Consulta.

A Figura 3, representando um segundo momento, descreve uma Orientação proposta pela aluna M. MacCulloch, em 10 de março de 2007, para a Consulta. A Orientação foi encaminhada ao professor para ser validada. A Orientação foi considerada correta e, portanto, armazenada na Base de Conhecimentos.

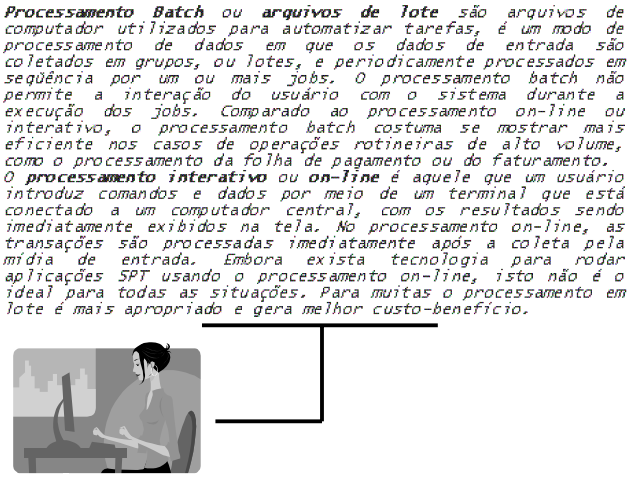

Figura 3. Aluna propõe Orientação à Consulta.

Na Figura 4, descrevendo um terceiro instante, o professor ou bolsista valida a Orientação. Ela é, então, armazenada na Base de Conhecimentos e disponível para responder novas Consultas similares a ela.

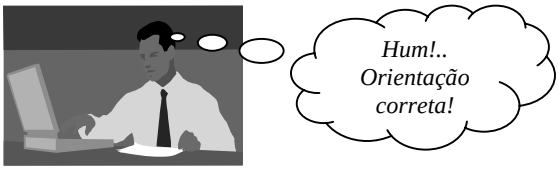

Figura 4. Professor valida Orientação.

As Orientações são 'oficialmente' validadas pelo professor ou bolsista. Futuramente, planeja-se uma participação efetiva dos alunos no processo de validação de orientação.

5.1.1 Outras consultas

A Tabela 1, a seguir, apresenta uma série de Consultas formuladas pelos alunos em diferentes datas. Estas Consultas foram consideradas similares, pelo ambiente, com aquela Consulta apresentada no primeiro momento. Assim, todas as consultas foram respondidas aos alunos *automaticamente* pela solução com a mesma Orientação proposta no segundo momento. Note que as Consultas são similares, porém não iguais.

Tabela 1. Consultas similares

Consulta	Autor	Data
Qual a diferença estrutural entre o processamento em lote e o processamento on-line?	M. Pereira	11/03/07
Qual a diferença entre processamento em lote e o processamento online?	M. Moura	09/04/07
Em que o processamento em lote se diferencia do processamento on line?	A. Barbosa	14/04/07
Qual a diferença entre processamento de dados em lote e on-line?	N. Andrade	07/04/08
Qual a diferença entre processamento em lote e online?	L. Santos	18/04/08

6. RESULTADOS E CONSIDERAÇÕES FINAIS

O uso da tecnologia amplia e potencializa situações nas quais professores e alunos, e estes entre si, pesquisem, discutam, se relacionem e permitem a construção e difusão de novos conhecimentos [5].

Alguns dados têm sido coletados desde 2005-2 e são apresentados a seguir confirmando esta proposição. Como o semestre 2008-2 apenas iniciou-se, os valores relativos deste semestre ainda são preliminares.

A Figura 5 apresenta o número total (8.629) de consultas formuladas no Hospital Educacional, classificado por semestre letivo.

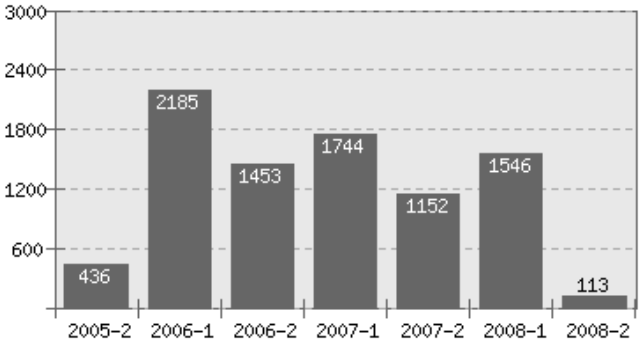

Figura 5. Total de Consultas formuladas por semestre.

Na Figura 6, é apresentada a quantidade de Orientações (5.684) armazenadas na Base de Conhecimentos do Hospital Educacional. Somente as Orientações validadas/aceitas pelo professor são armazenadas.

Na Figura 6, as Orientações estão classificadas pelas áreas suportadas pelo ambiente, na ordem da esquerda para direita: Redes, Segurança Digital, Banco de Dados, Engenharia de Software, Sistemas Operacionais, Sistemas de Informação e Outras.

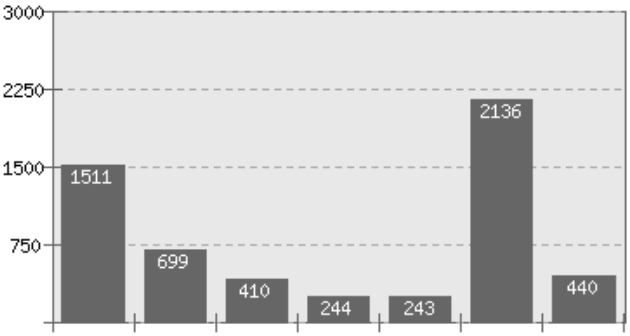

Figura 6. Total de Orientações por área.

A Figura 7 apresenta a quantidade de Consultas que foram respondidas automaticamente pela solução por semestre. A redução apresentada nos últimos semestres está relacionada com um menor número de Consultas formuladas nos respectivos períodos.

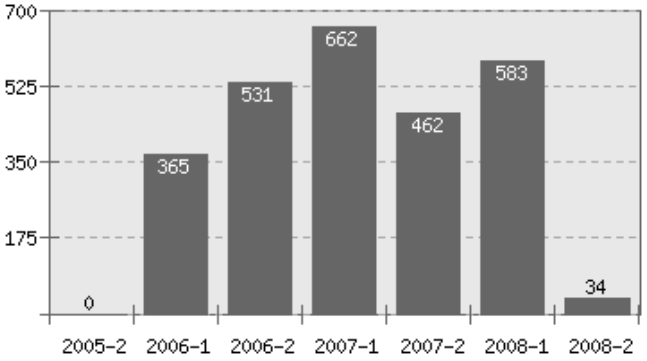

Figura 7. Orientações automáticas por semestre.

A Figura 8 informa a quantidade de Orientações (5.018) propostas pelos alunos. Eles começaram a propor Orientações somente a partir de 2006-2. Pelo volume de contribuições, confirma-se a existência de uma relação colaborativa entre eles.

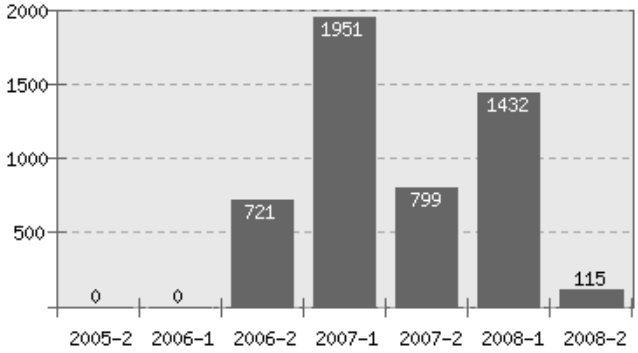

Figura 8. Orientações propostas pelos alunos por semestre.

Não apenas a quantidade de Orientações propostas pelos alunos é relevante, mas a sua qualidade é também expressiva. A Figura 9 apresenta a percentagem de Orientações propostas que foram consideradas corretas em relação ao total de Orientações propostas.

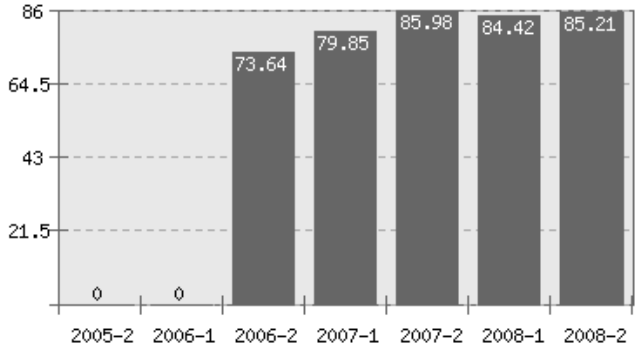

Figura 9. Qualidade das Orientações propostas pelos alunos por semestre.

É importante ressaltar que a cada nova Orientação armazenada na Base de Conhecimentos, a solução aumenta a sua capacidade em responder novas Consultas automaticamente. A Figura 10 relata esta capacidade crescente.

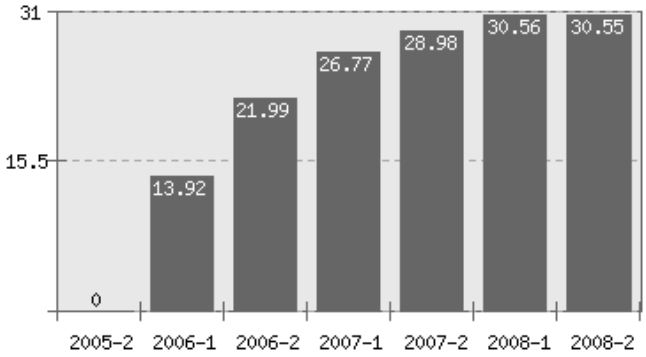

Figura 10. Índice de respostas automáticas por semestre.

A Figura 11 apresenta a participação do professor na construção da Base de Conhecimentos. Repare que esta contribuição é declinante em detrimento da participação dos alunos. Isto demonstra uma relação de colaboração entre os alunos e que eles também podem (e devem) contribuir fortemente em um ambiente educacional.

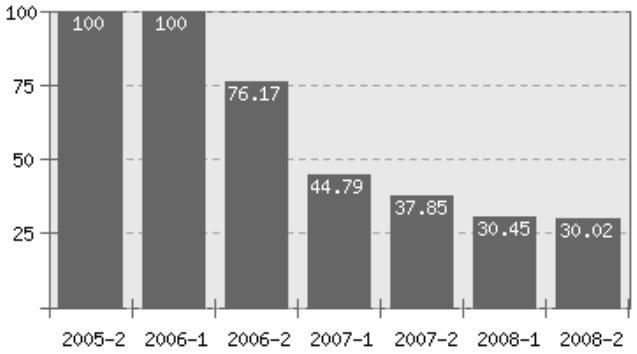

Figura 11. Participação do professor.

É importante ressaltar que a Base de Conhecimentos não armazena verdades absolutas, nunca foi este o propósito. Por isso, o ambiente possui ferramentas que permitem aos alunos corrigirem Orientações que não estejam corretas ou completas. A Figura 12 relata a quantidade de correções propostas pelos alunos às Orientações e aceitas pelo professor.

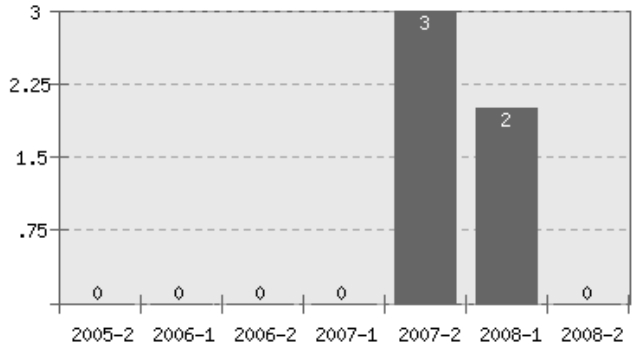

Figura 12. Correções à Base de Conhecimentos.

È notória a baixa quantidade de correções, apenas 5, em relação à totalidade de Orientações armazenadas (5.684) na Base de Conhecimentos. Isto é algo que está sendo trabalhado e discutido junto aos alunos. Entre as diversas razões possíveis para este

fenômeno, pode ser que o aluno espera e aceita toda orientação "oficial" como correta sem questioná-la.

Diversas outras estatísticas podem ser encontradas no ambiente do Hospital Educacional em http://www.hospitaleducacional.com/estatisticas.html.

7. REFERÊNCIAS

[1] Wiley, D. A. 2000, "Connecting Learning Objects to Instructional Theory: A Definition, a Metaphor, and a Taxonomy. The Instructional Use of Learning Objects -- Online Version", Open Publication License. http://www.reusability.org/read/chapters/wiley.doc

[2] Wen, J., Nie, J. e Zhang, H. 2002. "Query Clustering Using User Logs". ACM Transactions on Information Systems (TOIS). vol. 20, no. 1, pp. 59-81, 2002 DOI= http://doi.acm.org/10.1145/503104.503108

[3] Yates, R. e Neto, B. 1999, Modern Information Retrieval, New York: Addison-Wesley, 1ª edition.

[4] Vélez, B. et al. 1997, "Fast and Effective Query Refinement", In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in (Philadelphia PA, USA, July 02 - 05, 1997). SIGIR ´97. vol. 31, Issue SI, pp.06-15. DOI= http://doi.acm.org/10.1145/258525.258528

[5] Varella, G. 2002. "Aprendizagem Colaborativa em Ambientes Virtuais de Aprendizagem: A Experiência Inédita da PUCPR". Revista Diálogo Educacional, Curitiba, PR, Brazil, v.3 (maio/ago 2002), n. 6, p. 11-27. DOI= http://www2.pucpr.br/reol/index.php/DIALOGO?dd1=684&dd99=view