# UMA ABORDAGEM HÍBRIDA PARA O GERENCIAMENTO DE DOCUMENTOS FAQ EM PORTUGUÊS

**Andre Bortolon**

Programa de Pós-Graduação em Engenharia de Produção
Universidade Federal de Santa Catarina
CTC - Campus Universitário - Trindade
Caixa Postal, 476 - Florianópolis - SC - Brasil - 88040-900
bortolon@eps.ufsc.br

**Christiane Gresse von Wangenheim**

Curso de Ciência da Computação
Universidade do Vale do Itajaí - UNIVALI
Centro de Educação Superior VII - Ciência da Computação
Rod. SC 407, Km 04 - CEP 88122-000
São José/SC - Brasil
gresse@sj.univali.br

**Marlon Domingos**

Curso de Ciência da Computação
Universidade do Vale do Itajaí - UNIVALI
Centro de Educação Superior VII - Ciência da Computação
Rod. SC 407, Km 04 - CEP 88122-000
São José/SC - Brasil
marlond@sj.univali.br

**RESUMO.** Essencial para o sucesso de sistemas FAQ é sua habilidade para gerenciar sistematicamente o conhecimento experiencial contido nos documentos FAQ. Este artigo propõe uma abordagem híbrida para o gerenciamento de documentos FAQ em Português sobre linguagens de programação. A abordagem integra vários tipos de conhecimento e provê mecanismos inteligentes para a recuperação do conhecimento, assim como a evolução contínua e o aprimoramento da base de conhecimento. O ponto forte desta abordagem é a integração de técnicas de Raciocínio Baseado em Casos e Recuperação de Informação e sua adaptação para o domínio específico da aplicação. Atualmente, a abordagem está sendo implementada e avaliada através de dois protótipos de ferramenta, uma com documentos FAQ sobre a linguagem de programação Smalltalk e outra com documentos FAQ sobre JAVA.

**Palavras-chave.** Gestão de Conhecimento, Recuperação de Informação, Raciocínio Baseado em Casos, Sistemas Baseados em Conhecimento

**Categoria**: Sessão Técnica
**Área**: Inteligência Artificial (Gestão do Conhecimento)

# UMA ABORDAGEM HÍBRIDA PARA O GERENCIAMENTO DE DOCUMENTOS FAQ EM PORTUGUÊS[1]

**RESUMO.** Essencial para o sucesso de sistemas FAQ é sua habilidade para gerenciar sistematicamente o conhecimento experiencial contido nos documentos FAQ. Este artigo propõe uma abordagem híbrida para o gerenciamento de documentos FAQ em Português sobre linguagens de programação. A abordagem integra vários tipos de conhecimento e provê mecanismos inteligentes para a recuperação do conhecimento, assim como a evolução contínua e o aprimoramento da base de conhecimento. O ponto forte desta abordagem é a integração de técnicas de Raciocínio Baseado em Casos e Recuperação de Informação e sua adaptação para o domínio específico da aplicação. Atualmente, a abordagem está sendo implementada e avaliada através de dois protótipos de ferramenta, uma com documentos FAQ sobre a linguagem de programação Smalltalk e outra com documentos FAQ sobre JAVA.

**ABSTRACT.** Essential for the success of FAQ systems is their ability to systematically manage the experiential knowledge contained in FAQ documents. In this paper, we propose a hybrid approach for the management of Portuguese FAQ documents on programming languages. The approach integrates various types of knowledge and provides intelligent mechanisms for knowledge retrieval, as well as, the continuous evolution and enhancement of the knowledge base. The principal strength of the approach lies in the integration of techniques from Case-Based Reasoning and Information Retrieval and their adaptation to the specific application domain. The approach is currently being implemented and evaluated through two prototypical tools, one on FAQ documents on the programming language Smalltalk and another one on JAVA FAQs.

**Palavras-chave.** Gestão de Conhecimento, Recuperação de Informação, Raciocínio Baseado em Casos, Sistemas Baseados em Conhecimento

## 1 Introdução

Para implementar efetiva e eficientemente sistemas de software, os profissionais devem ter um conhecimento amplo e detalhado sobre as linguagens de programação. Entretanto, como o domínio de software é caracterizado por rápidos avanços tecnológicos, um desafio é ter ferramentas de suporte ao aprendizado de novas linguagens de programação. Neste contexto, o conhecimento tácito descreve experiências concretas obtidas por indivíduos. Por exemplo, o método de tentativa e erro, tem-se mostrado um importante fator que contribui para um processo de aprendizado efetivo. Uma possível forma de representar e transmitir este tipo de conhecimento são listas de *Frequently Asked Questions* (FAQ - Perguntas Freqüentemente Realizadas), as quais expressam uma questão e a resposta dada por um especialista. As FAQs explicitamente capturam *know-how* e estratégias de solução para auxílio na procura de uma solução adequada para o problema atual.

Neste contexto, o objetivo deste trabalho é desenvolver um sistema de software para o gerenciamento de documentos FAQ sobre questões relacionadas a uma linguagem de programação. O sistema deve, principalmente, armazenar documentos FAQ, recuperando um documento relevante para uma determinada questão e suportar a evolução contínua da base de conhecimento. Isto, no contexto do domínio da aplicação, requer técnicas para a representação de casos contendo informação textual sobre a questão e sua respectiva resposta em uma forma estruturada. De maneira a habilitar a recuperação inteligente de documentos FAQ, são necessários mecanismos baseados em similaridade que permitam a recuperação de documentos com questões similares, mas não necessariamente idênticas. Além disto, o sistema deve estar

[1] Pesquisa parcialmente patrocinada pelo Programa Probic/UNIVALI.

apto a tratar com consultas e documentos em linguagem natural. Isto inclui mecanismos para correção ortográfica e normalização de palavras, assim como a extração automática de termos relevantes. Na presente aplicação, a principal linguagem utilizada é o Português. Entretanto, as consultas podem ser expressadas juntamente com jargões em Inglês do domínio da aplicação (ex.: "O que é *class*?"), os quais devem ser tratados separadamente. De maneira a permitir a evolução contínua da base de conhecimento, a aquisição e integração de novos documentos FAQ deve ser suportada. Isto inclui mecanismos para a indexação semi-automática de novos documentos FAQ, assim como o incremento contínuo e a atualização do conhecimento geral do domínio.

Entretanto, a maioria dos repositórios FAQ existentes não provêm um acesso eficiente. Em geral, existem duas abordagens: visualização e busca. Abordagens de visualização oferecem uma coleção organizada que pode ser explorada pelo usuário, como por exemplo em *newsgroups*, mas que em geral demoram para encontrar uma resposta para a questão apresentada. A segunda abordagem é a procura utilizando técnicas de Recuperação de Informação (RI) ou procura por palavra-chave, como por exemplo os mecanismos de busca da Internet. Entretanto, estas abordagens freqüentemente têm resultados de baixa precisão e relevância. Recentemente, vários sistemas foram desenvolvidos utilizando técnicas de Raciocínio Baseado em Casos (RBC) [AP94], que suportam a recuperação baseada em similaridade de casos e a evolução incremental da base de conhecimento. Entretanto, o foco das abordagens existentes é o manuseio de documentos em Inglês. Hoje, não existe uma abordagem específica para recuperar documentos em Português, o qual é morfologicamente muito mais complexo.

Este artigo apresenta uma abordagem híbrida que pode tratar este tipo de problema. Ela integra técnicas de RBC e RI adaptadas ao domínio específico da aplicação.

## 2 O FAQSystem

O objetivo do FAQSystem é proporcionar uma ferramenta que, para uma determinada consulta formulada em Português, ajude a encontrar documentos FAQ relacionados, armazenados em uma base de casos, de maneira a ajudar os profissionais a resolver problemas e questões que aparecem durante a programação de um sistema de software. Como entrada do sistema, o usuário formula uma questão em Português e os termos relevantes são extraídos automaticamente da consulta baseados em vocabulários do domínio. Baseado nos termos extraídos da consulta dada, a base de casos que representa os documentos FAQ, é inquirida. São identificados documentos relevantes baseados no conjunto de índices referenciando o conteúdo da FAQ. Pela comparação dos casos com a consulta, usando uma medida de similaridade, uma ordem parcial é induzida entre os casos da base. Em uma primeira tentativa, o caso mais similar é sugerido ao usuário. Se este caso não corresponde satisfatoriamente a questão apresentada, o usuário pode explorar os próximos dez casos mais similares. Se o sistema falhar em apresentar algum caso suficientemente similar ou todos os casos recuperados não corresponderem à questão satisfatoriamente, o usuário pode requisitar o suporte de um especialista. Neste caso, o

especialista é informado sobre a questão e solicitado que forneça uma resposta. Assim que a resposta estiver disponível, ela é remetida para o usuário e, combinada com a consulta feita por ele, é capturada como um novo caso para a base de conhecimento. O novo caso é automaticamente indexado usando vocabulários do domínio. Os resultados deste processo são revisados pelo especialista e, se necessário, melhorados.

Os principais aspectos da abordagem, representação do conhecimento, recuperação e evolução contínua são descritos em detalhes nas seções subseqüentes.

### 2.1 Representação do Conhecimento

A informação e o conhecimento nos documentos FAQ são representadas na forma de casos. Para representar os documentos FAQ de maneira mais acessível, a descrição textual é mapeada em uma representação estruturada, a qual consiste do texto da questão, o texto da resposta, um conjunto de índices e o tipo da questão, conforme apresentado na figura 1.

| Caso 007 | |
|---|---|
| Pergunta | Como ordenar uma coleção? |
| Resposta | Enviando a mensagem sort para esta coleção |
| Índices | ordenar, coleção |
| Tipo | modo (tipo 3) |

Figura 1. Exemplo de representação do caso

Os índices indicam termos do texto da questão que são relevantes para a recuperação de casos úteis no domínio da aplicação. A classificação dos casos por tipo de questão expressa a necessidade de diferentes tipos de respostas. Por exemplo, a questão “O que é uma coleção?” requer uma resposta diferente da questão “Como ordenar uma coleção?”, embora parte dos índices (“coleção”) sejam os mesmos. Seis categorias de questões foram definidas, englobando questões sobre definição, quantidade, modo, utilidade, exemplo e genéricas [Bor01].

Além do conhecimento representado em casos, o conhecimento geral do domínio é representado de maneira a dar suporte à extração automática de texto, correção ortográfica e recuperação baseada em similaridade. Vários tipos de conhecimento geral do domínio estão incluídos:

**Vocabulário específico ao domínio**, o qual define expressões indicativas em um domínio específico (ex.: “classe”, “executar”, “rápido”) permitindo a extração automática de informações. Dois tipos de vocabulários foram separados, um com nomes de classes (ex.: “Collection”, “OrderedCollection”) e um com nomes de métodos freqüentemente usados (ex.: “addAll”, “initialize”).

**Dicionário Inglês-Português específico ao domínio,** representa termos de jargões em Inglês e sua tradução (ex.: “class” –> “classe”) permitindo a tradução automática para o Português.

***Thesaurus* específico ao domínio**, o qual representa relações entre termos específicos ao domínio (ex.: “OrderedCollection” –> “Collection”) permitindo a recuperação de casos similares.

**Regras de normalização**, as quais servem para realizar a remoção ou alteração de sufixos, mudanças de gênero da palavra e para as conjugações de verbos irregulares.

**Vocabulário geral de Português**, o qual representa um vocabulário geral da Língua Portuguesa, incluindo mais de 20.000 palavras baseado em [KK99].

***Stop list***, que inclui cerca de 200 palavras, tais como "aqui", "acima", "para", as quais são extremamente comuns no idioma e que são excluídas da indexação.

### 2.2 Processo de Recuperação

De maneira a permitir a recuperação de uma resposta útil para a questão do usuário, um método de recuperação baseado em similaridade foi definido, incluindo os seguintes passos.

#### 2.2.1 Formulação da consulta

A consulta é descrita pelo usuário pela formulação de uma questão em linguagem natural em Português, a qual pode também incluir certos termos específicos em Inglês (ex: "O que é *class*?").

#### 2.2.2 Interpretação da consulta

A consulta é automaticamente analisada e transformada em uma representação interna. O objetivo deste passo é extrair todos os termos relevantes da questão como índices e classificar o tipo da questão, seguindo os passos:

**Separação em *tokens***: a separação em *tokens* intenciona a divisão do texto da consulta (ex.: "Como ordenar uma *OrderedCollection*?") em *strings* de caracteres (ex.: {"como", "ordenar", "uma", "*OrderedCollection*"}), denominada como *QueryStringList*.

**Classificação do tipo da questão**: a classificação do tipo da questão determina a categoria da questão. A classificação é baseada nos pronomes ou advérbios interrogativos utilizados no início da questão. Por exemplo, uma questão iniciada com "Como" é classificada como uma questão de modo, a qual expressa o modo ou maneira que algum assunto é tratado ou feito.

**Extração de termos em Inglês específicos ao domínio**: a extração automática de termos em Inglês misturado com o texto em Português é feita através do uso de vocabulário dos nomes de classes e métodos e do dicionário Inglês-Português específico ao domínio. A medida que na aplicação específica os termos em Inglês não são derivados ou declinados, não é necessária a normalização. Entretanto, uma característica específica dos nomes de classes e métodos é que várias palavras são concatenadas em um termo, como em "*OrderedCollection*". Desta forma, um erro de digitação freqüente é a separação destes termos (ex.: "*Ordered Collection*"). Relativo a estas características específicas, a extração dos termos em Inglês é feita por um processo iterativo baseado em uma concatenação de termos e uma verificação do termo resultante no vocabulário de nomes de classes, no vocabulário de nomes de métodos ou no dicionário.

**Correção ortográfica**: o objetivo é corrigir erros de ortografia no texto em Português, por exemplo ausência, excesso ou troca de caracteres, inversão de letras ou ausência de acentos. A correção ortográfica é baseada na comparação de uma palavra de *QueryStringList* com os termos do vocabulário geral de Português. Se a palavra estiver no vocabulário ou existir alguma palavra

com similaridade maior que 60%, a palavra de *QueryStringList* é substituída por ela.

**Normalização**: a normalização de termos possibilita a comparação de termos em variantes morfológicas diferentes. Na aplicação específica, foi observada a necessidade de normalização relativas à declinação de substantivos, adjetivos e verbos. Através do processo de normalizaçao, cada palavra é convertida para sua forma normal seguindo a Nomenclatura Gramatical Brasileira, por exemplo substantivos para a forma masculina singular e verbos para a forma infinitiva. Isto é executado por processo iterativo baseado em uma redução de palavras através de regras e uma verificação do termo resultante no vocabulário geral de Português. As regras basicamente desfazem as regras ortográficas para adição de afixos, cobrem a geração de plurais e outras declinações, como as desinências verbais.

**Extração de termos relevantes em Português**: após serem corrigidos e normalizados, os termos em Português relevantes para a recuperação dos documentos FAQ são extraídos de *QueryStringList*. A extração é feita pela comparação de cada palavra de *QueryStringList* como os termos definidos no vocabulário específico do domínio. Se a palavra é encontrada no vocabulário é adicionada como um índice, caso contrário é ignorada.

### 2.2.3 Cálculo da Similaridade

Em relação aos índices e ao tipo de questão da consulta, para todos os casos da base é calculado um valor de similaridade usando medidas de similaridade em diferentes níveis.

**Similaridade Global**: a similaridade global de um caso $c_k$ da base com a consulta $q$ é calculada por:

$$sim(q, c_k) = \frac{\sum_{i,j=1}^{n} simLoc(q_i, c_{kj}) + simTipo(q, c_k)}{n+1}$$

onde $simLoc(q_i, c_{kj})$ é a similaridade local, $simTipo(q, c_k)$ é a similaridade do tipo da questão e $n$ é o número total de índices da consulta $q$. Qualquer caso com uma similaridade superior a 40% é considerado como um candidato em potencial para uma resposta para a consulta.

**Similaridade Local**: a determinação da similaridade entre a consulta e um caso na base é melhor aprimorada através da integração do *thesaurus*. Isto permite a consideração de valores similares, mas não necessariamente iguais. Por exemplo, se a consulta for “Como executar uma imagem?”, casos representando questões similares, como “Como rodar uma imagem?”, caracterizados através de termos similares, tais como “executar” e “rodar”, deveriam também ser considerados como candidatos potenciais. A similaridade local do i-ésimo índice $q_i$ da consulta $q$ é calculada pela comparação do índice $q_i$ com cada índice do caso $c_k$ da base de casos, considerando também termos similares baseados no *thesaurus* ($s_i$ é um conjunto de termos similares a um determinado termo $q_i$):

$$simLocj(q_i, c_{kj}) = \begin{cases} 1.0 se(\exists(x \in c_{kj})) : x = q_i \\ 0.9 se((((\neg\exists(x \in c_{kj}))) : x = q_i i) \wedge ((\exists(y \in s_i))) : y = q_i) \\ 0 se((((\neg\exists(x \in c_{kj}))) : x = q_i) \wedge ((\neg\exists(y \in s_i))) : y = q_i) \end{cases}$$

onde $c_{kj}$ é o j-ésimo índice do caso $c_k$. A similaridade local $simLoc(q_i, c_{kj})$ do i-ésimo índice $q_i$ da consulta $q$ e o caso $c_k$ é determinada através do máximo $simLoc_j(q_i, c_{kj})$ para todos os índices $c_{kj}$.

**Similaridade de Tipo**: a similaridade do tipo da questão $simTipo(q, c_k)$ é determinada pela comparação do tipo da questão $tipo_q$ da consulta $q$ com o tipo $c_k$ de um caso $c_k$ da base de casos:

$$simTipo(q, c_k) = \begin{cases} 1 se tipo(q) = tipo(c_k) \\ 0 se tipo(q) \neq tipo(c_k) \end{cases}$$

**Similaridade de Desempate**: utilizando a medida de similaridade global descrita anteriormente, uma ordenação parcial é induzida entre os casos da base. Entretanto, vários casos podem ter o mesmo valor máximo de similaridade para uma determinada consulta. Entretanto, como o objetivo primário do FAQSystem é recuperar uma única resposta, a medida de similaridade de desempate $simDes(q, c_k)$ serve nestas situações para tentar refinar o cálculo da similaridade através de um tipo diferente de normalização:

$$sim(q, c_k) = \frac{\sum_{i,j=1}^{n} simLoc(q_i, c_{kj}) + simTipo(q, c_k)}{\frac{n+m}{2} + 1}$$

onde $n$ é o número total de índices da consulta $q$ e $m$ é o número total de índices do caso $c_k$.

### 2.2.4 Proposta de resposta

O caso mais similar, se tiver associado a um valor de similaridade maior do que um determinado limite (40%), é retornado para o usuário como a resposta em potencial para sua questão. Se não satisfeito, o usuário pode também requisitar os dez casos mais similares seguintes recuperados.

### 2.2.5 Resposta manual

Se não satisfeito com os casos retornados pelo processo de recuperação ou se nenhum caso suficientemente similar for encontrado, o usuário pode requisitar a resposta manual de sua questão. Desta forma, a questão do usuário é armazenada na base de casos (marcada como ainda sem resposta) e um especialista do domínio da aplicação é informado via *e-mail* que um usuário está requisitando sua resposta. Assim que a resposta estiver disponível, o usuário é comunicado também via *e-mail*.

## 2.3 Evolução contínua do FAQSystem

### 2.3.1 Aquisição de casos

Toda a vez que uma consulta for manualmente respondida por um especialista, um novo caso é adquirido, permitindo a evolução e a atualização contínua da base de casos. De maneira a facilitar a integração do novo caso à base de casos existente, o processo de indexação é feito automaticamente, extraindo-se os termos relevantes usando as técnicas descritas na Seção 2.2.2. O novo caso criado é revisado pelo especialista do domínio que pode adicionar, modificar ou remover índices associados baseados no texto da questão do documento FAQ. No caso de novos termos, que ainda não estão incluídos no conhecimento geral do domínio, tornarem-se relevantes para a descrição do caso, os vocabulários, dicionário e *thesaurus* serão conseqüentemente

atualizados.

### 2.3.2 Evolução do vocabulário do domínio

Os vocabulários específicos ao domínio devem ser atualizados sempre que novos termos tornem-se disponíveis. Para identificar termos com poder descritivo na coleção de casos, novos casos adquiridos são pré-processados, utilizando técnicas que incorporam a freqüência de termos e a freqüência inversa de documentos [SM83]. Pesos são determinados para cada termo na questão que não foi reconhecido pela indexação automática. Usando uma *stop list*, quaisquer palavras semanticamente não relevantes são filtradas. Os termos restantes são ordenados através da utilização da técnica de freqüência inversa de documentos, onde o peso do termo *k* no documento *i* é representado por: $weigth_{ik} = tf_{ik} \cdot (\log_2 n - \log_2 df_k + 1)$

onde $tf_{ik}$ é a freqüência do termo *k* no documento *i*, $df_k$ o número de documentos nos quais *k* ocorre e *n* o número total de documentos na base de casos. Este peso induz a uma ordenação parcial entre os novos termos, direcionando a investigação manual pelo especialista do domínio.

## 2.4 Evolução do *thesaurus* específico ao domínio

Toda a vez que um novo termo *k* é adicionado aos vocabulários específicos ao domínio, é necessário atualizar também o *thesaurus*. Para isto, um pré-processamento é feito com base em co-ocorrência estatística de palavras [SM83], onde os coeficientes de similaridade são obtidos entre pares de termos distintos baseados em coincidências nas associações de termos para os documentos da coleção. Desta forma, os documentos da base são representados por uma matriz como mostrado na Tabela 1 baseado no modelo de espaço vetorial, onde as linhas da matriz representam os vetores individuais dos documentos e as colunas identificam as associações de termos aos documentos.

| | T1 | T2 | ... | Tk | ... | Tm |
|---|---|---|---|---|---|---|
| D1 | tf11 | tf12 | ... | tf1k | ... | tf1m |
| ... | ... | ... | ... | ... | ... | ... |
| Dn | tfn1 | tfn2 | ... | tfnk | ... | tfnm |

Tabela 1. Matriz de associação de termos

Assim, a similaridade entre o novo termo *k* e qualquer termo *l* pode ser medida baseada nos respectivos pares de colunas da matriz:

$$sim(TERM_k, TERM_l) = \frac{\sum_{i=1}^{n} tf_{ik} \cdot tf_{il}}{\sum_{i=1}^{n} tf_{ik}^{\,2} + \sum_{i=1}^{n} tf_{il}^{\,2} - \sum_{i=1}^{n} tf_{ik} \cdot tf_{il}}$$

dados os vetores de termos na forma de $TERM_k = (tf_{1k}, ..., )$ onde $tf_{ik}$ indica a freqüência de $TERM_k$ no documento *i*, assumindo *n* documentos na base. Como resultado, é computado um vetor de associação $term_k$-term $T_k$, expressando a similaridade do termo *k* com cada termo *l* do vocabulário específico ao domínio através de $sim(TERM_k, TERM_l)$. Em relação da ordem parcial criada pelo valor de similaridade o especialista revisa a seleção pré-processada e, se adequado,

adiciona novas associações ao thesaurus específico ao domínio.

### 2.5 Aperfeiçoamento contínuo através de *feedback*

Como o conteúdo dos documentos e os termos relevantes poderem modificar-se com o tempo, o aperfeiçoamento contínuo do conhecimento do domínio e da medida de similaridade para a recuperação necessita ser suportada durante todo o ciclo de vida de um sistema FAQ. Isto deve ser feito pela análise da performance do sistema e possíveis modificações no contexto da aplicação pelo engenheiro de conhecimento, por exemplo com base em *log files* ou *feedback* do usuário.

## 3 Implementação

Baseado na abordagem apresentada, dois protótipos estão sendo desenvolvidos atualmente, um para o gerenciamento de documentos FAQ sobre a linguagem de programação Smalltalk [Bor01], cuja interface pode ser vista na figura 2 e outro sobre a linguagem de programação JAVA [DG00].

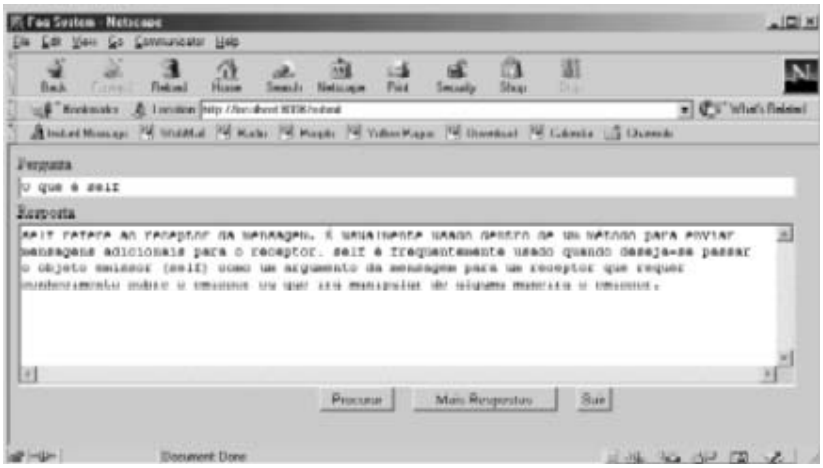


Figura 2. Exemplo de interface do sistema FAQ@Smalltalk

As ferramentas possuem ma arquitetura cliente-servidor consistindo de três camadas lógicas: apresentação, aplicação e armazenagem de dados. O conhecimento, incluindo os documentos FAQ assim como o conhecimento do domínio, é armazenado em um sistema de arquivos. A camada de aplicação provê ferramentas de suporte para a recuperação de documentos FAQ, o processo de resposta manual, a aquisição de novos casos e o aperfeiçoamento do conhecimento do domínio. O acesso é realizado através de *browsers* e sistemas de *e-mail* via Internet. A ferramenta foi desenvolvida independente de plataforma em Smalltalk em VisualWorks 5i.1.

## 4 Avaliação

Adaptando as técnicas de avaliação de sistemas de RI e RBC para os sistemas FAQ, a abordagem apresentada foi avaliada de acordo com os critérios da velocidade de recuperação (o tempo requerido para realizar a recuperação) e *recall* (percentual de questões para as quais o sistema retornou a resposta correta, se existir). Para determinar a contribuição dos vários

aprimoramentos feitos, foi realizado um estudo através da remoção do sistema dos componentes mais sofisticados:

1. Sistema completo: todos os componentes do sistema presentes;
2. Sem correção ortográfica: o sistema sem o módulo de correção ortográfica;
3. Sem normalização: o sistema do teste anterior sem o módulo de normalização;
4. Sem extração de informação: o sistema sem os vocabulários do conhecimento do domínio;
5. Sem similaridade local: o sistema sem a medida de similaridade local;
6. Sem similaridade global: o sistema sem medida de similaridade (casamento perfeito);

Os testes com 40 consultas foram realizados usando uma base de casos com 200 casos em um Pentium III 800 Mhz com 128 MB de memória RAM.

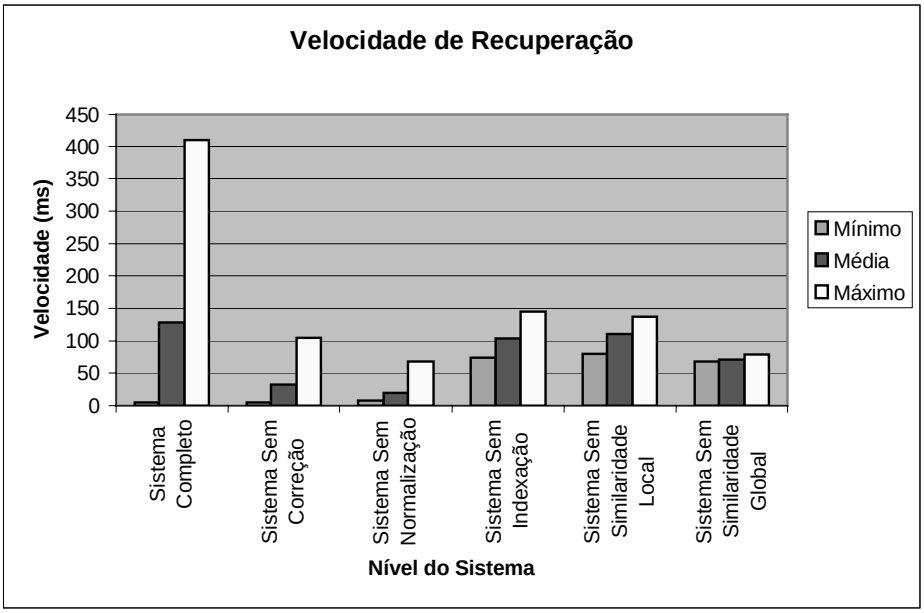


Figura 3. Gráfico com o tempo de recuperação do sistema

**Velocidade de Recuperação**. Em geral, a velocidade de recuperação é extremamente rápida, com uma média de 129 ms por consulta para o sistema completo. Como mostrado na figura 3, a velocidade de recuperação não aumenta de acordo com a complexidade do sistema. Por exemplo, foi observada uma redução do tempo de recuperação quando incluídos mecanismos de extração de informação. Isto pode ser explicado devido ao fato dos testes 4, 5 e 6 processarem mais índices (ex.: incluindo termos irrelevantes como artigos). Como resultado, a integração dos mecanismos de extração de informação reduziu significativamente o tempo de recuperação, deixando-o quase o mesmo de um sistema baseado em comparação perfeita. Também foi verificado que dependendo da necessidade de correção ortográfica e normalização, a velocidade pode variar significativamente.

***Recall.*** O *recall* obtido da abordagem foi muito alto, com cerca de 83% de consultas respondidas corretamente pelo sistema completo. Metade das questões não respondidas corretamente não eram cobertas pela base de casos. Isto significa que o sistema retornou uma resposta em uma

situação onde não deveria ter retornado alguma. Comparando os diferentes componentes do sistema, o grande incremento do *recall* foi obtido através da integração da medida de similaridade local e do *thesaurus* específico ao domínio. como pode ser observado na figura 4.
Esta avaliação mostrou que a abordagem híbrida desenvolvida permite resultados significantemente melhores em comparação com outros sistemas que utilizam estas técnicas separadamente.

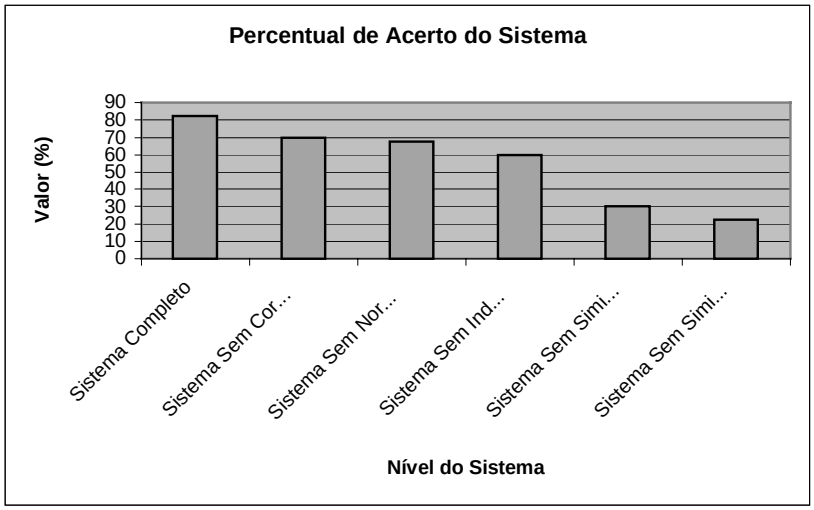


Figura 4. Gráfico com os resultados do *recall*

## 5 Conclusão

Comparado a outras abordagens existentes, este trabalho traz uma contribuição com relação, especialmente, à extração automática de informação de consultas ou documentos FAQ na Língua Portuguesa e ao suporte semi-automático para a evolução contínua do conhecimento do domínio. Os protótipos implementados estão atualmente sendo usados no grupo de pesquisa The Cyclops Project da UFSC e no curso de Ciência da Computação da UNIVALI. Baseado no *feedback* desta utilização na prática, pretende-se direcionar a pesquisa para a ampliação da ferramenta para outras áreas, assim como a evolução e generalização das técnicas aplicadas.

**Agradecimentos**
Os autores querem agradecer a todos os membros do The Cyclops Project que participaram do estudo inicial e na aplicação do sistema FAQ@Smalltalk e à Maria Marta Leite pela revisão deste artigo.

**Referências Bibliográficas**

[AP94] AAMODT, A., PLAZA, E. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications, 17(1), 1994.

[Bor01] BORTOLON, A. Desenvolvimento e Implementação de uma Abordagem Híbrida para o Gerenciamento de Documentos FAQ em Português. Dissertação de Mestrado, Engenharia de Produção, Universidade Federal de Santa Catarina, 2001.

[DG00] DOMINGOS, M., WANGENHEIM, C. Gresse von. Desenvolvimento de um sistema FAQ@JAVA. Seminário de pesquisa, Universidade do Vale do Itajaí - CES VII, 2000.

[KK99] KUENNING, G. H., KARPISCHEK, R. U. International Ispell Version 3.1.20.

[SM83] SALTON, G, MCGILL, M. J. Introduction to Modern Information Retrieval. New York: McGraw Hill, 1983