Avaliação da Recuperação em Sistemas de RBC Estrutural e Textual: Uma Aplicação no Domínio de Help Desk

Fábio Pessôa de Sá, Marta Costa Rosatelli, Eduardo Raul Hruschka

Programa de Mestrado em Informática – Universidade Católica de Santos
R. Dr. Carvalho de Mendonca, 144 – Santos-SP – Brasil – 11070-906

fabio@iron.com.br,{rosatelli,erh}@unisantos.br

Abstract. *This paper presents an evaluation of the retrieval task in two different approaches of Case Based Reasoning (CBR): structural and textual. In order to perform the retrieval task, a real world case base of Frequently Asked Questions (FAQ) in the help-desk domain is used. The cases were modeled with the support of a domain expert. In structural CBR, the case base is structured as attribute-value pairs, and the nearest-neighbor method based on the simple matching approach is used. In textual CBR, the case base is the FAQ content, and the cases are retrieved using a vector model of the terms under interest. Empirical results shed some light on the cost versus benefits of structuring a case base against using an existing set of documents.*

Resumo. *Este artigo apresenta uma avaliação da etapa da recuperação de duas abordagens de Raciocínio Baseado em Casos (RBC): estrutural e textual. Para realizar a etapa da recuperação foi utilizada uma base de casos reais do domínio de help-desk, formada por questões freqüentes (FAQ - Frequently Asked Questions). Os casos foram modelados com o auxílio de um especialista no domínio. No RBC estrutural, a base de casos é estruturada na forma de pares atributo-valor, e o método do vizinho-mais-próximo baseado no coeficiente de casamento simples é usado. No RBC textual, a base de casos é formada pelas próprias FAQs (Frequently Asked Questions), e os casos são recuperados usando um modelo de vetor de termos de interesse. Resultados experimentais demonstram a relação custo versus benefício entre o esforço de estruturação de uma base de casos contra o uso de um conjunto de documentos existente.*

1. Introdução

O Raciocínio Baseado em Casos (RBC) consiste em uma abordagem para desenvolver sistemas baseados em conhecimento capazes de recuperar e reutilizar soluções que funcionaram em situações similares no passado (Kolodner 1993). A etapa da Recuperação é crucial no desenvolvimento de sistemas de RBC. Esta etapa se baseia numa descrição parcial do problema, que é então usada para encontrar um caso que potencialmente contém a solução para o problema - segundo uma medida de similaridade entre os dados do problema a ser resolvido e os casos da base. Abordagens diferentes de RBC têm formas diferentes de lidar com a etapa da recuperação. Este artigo apresenta uma avaliação da etapa da recuperação em duas abordagens de RBC: estrutural e textual.

Para avaliar a etapa da recuperação, utiliza-se uma base de casos real no domínio de *help-desk*. Sistemas de *help-desk* são amplamente empregados em diversas organizações, oferecendo suporte em forma de informações e/ou ações. Estas podem ser usadas pelos consumidores de determinados produtos de uma organização (Dearden e Bridge 1993). Schulz (1999) acrescenta que a utilização de aplicações de RBC para *help-desk* não oferece o conhecimento da experiência (os casos) apenas para os clientes das empresas. Ao capturar os problemas e soluções, o sistema de *help-desk* pode construir automaticamente uma "memória corporativa", de tal forma que o conhecimento gerado não ficaria restrito somente a alguns funcionários da empresa. Esse conhecimento armazenado pode servir também para consultas futuras, as quais podem ser feitas, por exemplo, por novos funcionários da empresa. Essa vantagem é bastante apropriada às áreas que estão em constante mudança e que oferecem dinamismo (e.g., área de tecnologia).

Roth-Berghofer e Iglezakis (2000) identificam que a maioria das ferramentas de suporte de *help-desk* não contribui de maneira completa na solução de problemas em todos os níveis de problemas a serem solucionados, sugerindo que diferentes problemas podem requerer diferentes organizações da base de casos e/ou diferentes métodos de recuperação de casos (e.g. ver Watson 1997; Yang *et al*. 1997; Schulz 1999; Wangenheim e Wangenheim 2003; Empolis Orenge 2006; AI-CBR 2006). Neste artigo, duas abordagens distintas (estrutural e textual) são investigadas numa aplicação real de *help-desk*.

As demais seções desse artigo estão organizadas da seguinte forma: A Seção 2 apresenta o sistema de RBC estrutural, enquanto a Seção 3 descreve o sistema de RBC textual. Na Seção 4 é descrito o desenvolvimento das aplicações. A Seção 5 reporta os resultados da avaliação experimental, executada numa base de casos reais, dos sistemas de RBC propostos neste artigo. Finalmente, a Seção 6 apresenta as conclusões e algumas sugestões para trabalhos futuros.

2. Sistema de RBC Estrutural

O RBC estrutural é caracterizado pela representação dos casos estruturados na forma de pares atributos-valor. Ou seja, a modelagem do caso, os atributos e seus valores são criados de modo a representar o problema e a solução do caso. Nesses sistemas a similaridade dos casos é computada com respeito à estrutura e ao conteúdo dos casos (Roth-Berghofer e Iglezakis 2000).

Para exemplificar a modelagem de um caso em um sistema de RBC estrutural são utilizados neste trabalho problemas e soluções relacionadas à interface do Windows 95. Esses problemas e soluções podem estar, por exemplo, documentados em uma *FAQ*. A idéia de uma *FAQ* é armazenar opiniões de um grupo relacionadas a uma questão comum e deixar as respostas disponíveis em algum meio de comunicação (Burke *et al*. 1997).

Tabela 1. FAQ sobre a interface do Windows 95

Como removo todos os arquivos de minha lista Documentos?	Para remover os arquivos do Menu Documentos, faça o seguinte: 1. Dê um clique no botão Iniciar, escolha Configurações e depois Barra de Tarefas. 2. Dê um clique na guia Programas do Menu Iniciar e selecione o botão Limpar na seção Menu Documentos.

Utilizando-se o exemplo descrito acima, as informações são modeladas de maneira estruturada e resumida através de pares atributo-valor, conforme exposto na Tabela 2.2. A categoria de modelagem da base de casos utilizada neste trabalho é a homogênea, em função do grupo de atributos escolhidos para modelagem dos casos poder representar toda a base de casos.

Tabela 2. Atributos e valores de um caso

Atributos	**Valores**
O problema está relacionado ao menu Iniciar?	Sim
O problema se refere a barra de tarefas?	Sim
O problema ocorre na área de trabalho?	Não
O problema está relacionado com os ícones ou atalhos?	Não

No presente trabalho, os atributos utilizados no RBC estrutural possuem valores binários pertencentes ao conjunto {0,1}. Estes são considerados atributos simétricos, pois a não existência de um valor automaticamente implica em seu oposto. Nesta perspectiva, escolheu-se o bem conhecido coeficiente de casamento simples (Wangenheim e Wangenheim 2003) para avaliar a similaridade entre os casos. Mais formalmente, assumamos que um caso **c** possa ser descrito por um conjunto de *n* atributos. Pode-se usar a notação vetorial para representar um determinado caso, i.e., **c** = $[c_1, c_2, ..., c_n]$. Assim, o coeficiente de casamento simples (CCS) entre um caso **q** = $[q_1, q_2, ..., q_n]$ e um caso **c**=$[c_1, c_2, ..., c_n]$ é dado por:

$$CCS = \frac{\sum_{i=1}^{n} s_i}{n} \tag{1}$$

Onde:

$$s_i = \begin{cases} 1 & \text{se } q_i = c_i \\ 0 & \text{se } q_i \neq c_i \end{cases} \tag{2}$$

O método do vizinho-mais-próximo baseado no coeficiente de casamento simples é muito utilizado em ferramentas comerciais de diagnóstico como, por exemplo, o sistema *CBR-Works* (Schulz 1999). Este método apresenta simplicidade e boa

cobertura na recuperação dos casos além ser adequado à base de casos utilizada neste trabalho.

3. Sistema de RBC Textual

Nos sistemas de RBC textuais, documentos relevantes são recuperados a partir de uma coleção de documentos como resposta a uma questão colocada por algum usuário desse sistema (Lenz *et al*. 1998). Nesse contexto, um documento não é necessariamente um texto único, mas pode consistir de vários componentes, como por exemplo, uma coleção de perguntas freqüentes (*FAQ*). A base de casos no sistema de RBC textual usada neste trabalho é representada através de uma FAQ relacionada ao domínio de *help-desk*.

O modelo de recuperação de casos textuais aqui adotado é baseado no modelo de vetor. Neste, cada documento é representado por um vetor, que é uma lista de termos com seus respectivos pesos associados, os quais descrevem o valor do termo para um documento. Um termo é definido como uma palavra ou frase utilizada em um documento. O peso do termo pode ser quantificado por meio da freqüência bruta da ocorrência do mesmo dentro de um documento (Baeza-Yates e Ribeiro-Neto 1999). Tal freqüência é usualmente denominada fator *tf* (*freqüência do termo*). Além disso, o peso do termo pode também ser quantificado através da medida inversa da freqüência do termo em relação a uma determinada coleção de documentos. Este fator costuma ser denominado como *freqüência inversa de documentos - idf*. A motivação para utilização do fator *idf* é que termos que aparecem em muitos documentos não são muito úteis para distinguir um documento relevante de um documento não relevante. No presente trabalho, o peso *w* de cada termo é calculado levando-se em conta tanto *tf* quanto *idf*. Tal abordagem é amplamente utilizada na prática (Baeza-Yates e Ribeiro-Neto 1999) e pode ser formalmente definida por meio da equação (3).

$$w = tf \cdot idf \tag{3}$$

Na equação (3), a freqüência normalizada, *tf*, de um termo em um documento é dada por:

$$tf = \frac{freq}{\max freq} \tag{4}$$

Onde *freq* é a freqüência de um determinado termo no documento e *max freq* é o número máximo de vezes que um termo aparece no texto do documento, calculado sobre todos os termos possíveis. O termo *idf* por sua vez, leva em conta o número total de documentos da base de casos, *N*, e o número de documentos em que determinado termo aparece (*n*), sendo calculado pela equação (5).

$idf = \log \frac{N}{n}$	(5)

Representar documentos por meio de vetores dos pesos dos termos torna possível comparar dois documentos usando o método de similaridade baseado na medida do co-seno entre dois vetores em um espaço *n* dimensional, onde *n* é o número

de termos utilizados na coleção inteira de vetores (Wilson e Bradshaw 1999). Mais especificamente, o grau de similaridade de um documento, representado por um vetor **d**=[d_1, d_2, ..., d_n], com a consulta (*caso-problema)*, representada por um vetor **q**=[q_1, q_2, ..., q_n] é avaliado por meio de *sim*(**d,q**) ∈ [0,1] (Baeza-Yates e Ribeiro-Neto 1999) descrito na equação (6).

$$sim(\mathbf{d},\mathbf{q}) = \frac{\mathbf{d}\cdot\mathbf{q}}{|\mathbf{d}|\cdot|\mathbf{q}|} = \frac{\sum_{i=1}^{n}(d_i.q_i)}{\sqrt{\sum_{i=1}^{n} d_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}} \qquad (6)$$

4. Desenvolvimento das Aplicações

Duas aplicações de RBC - estrutural e textual - foram desenvolvidas com o objetivo de avaliar a etapa de recuperação utilizando uma mesma base de casos. A base de casos contém 70 problemas e soluções sobre o domínio do sistema operacional Microsoft Windows. Em particular, os casos abordam problemas com a interface do Windows. Essa base é puramente textual, segue o padrão de uma FAQ, e é utilizada pela empresa *Tree Tools Informática* para testes com suas ferramentas de desenvolvimento.

4.1. RBC Estrutural

Para desenvolver a aplicação de RBC estrutural foi utilizada a *shell* de desenvolvimento de aplicações *CBR-Works Professional* versão 4.0. Esse sistema apresenta um módulo denominado *Concept Manager* que serve como um editor para modelar o domínio da aplicação. Neste módulo são determinados os atributos e seus tipos, sendo também possível especificar as medidas de similaridade locais dos atributos (CBR-Works 2001).

A consulta ao especialista da empresa embasou a definição dos atributos para representar a base de casos utilizada no trabalho. Os tipos de atributos são definidos como binários (*Boolean*) conforme descrito na Seção 2, e somente o atributo denominado *Solução* não é marcado como *previsor* - por não ser levado em conta durante o processo de recuperação dos casos. Os atributos escolhidos cobrem toda a base de casos, caracterizando uma base de casos homogênea. Neste trabalho, assume-se que todos os atributos são igualmente importantes, de tal forma que pesos unitários são atribuídos para todos os atributos.

No módulo *Concept Manager* define-se ainda a medida de similaridade a ser utilizada ao recuperar os casos. Neste trabalho, a medida de similaridade utilizada é a denominada "padrão" no *CBR-Works*. Dessa forma, o sistema de RBC estrutural modelado utiliza a abordagem do coeficiente de casamento simples em função dos tipos de valores dos atributos.

4.2. RBC Textual

Para desenvolver a aplicação de RBC textual foi implementado um algoritmo em *VB.Net* que trabalha com um método de pesagem de termos em conjunto com a medida

de distância do co-seno, além de utilizar a técnica de pré-processamento de textos de eliminação de *stopwords*.

A base de casos utilizada durante a implementação da recuperação do RBC textual é a mesma base do RBC estrutural, ou seja, a FAQ "Interface com o usuário" do Windows 95. Por ser uma base de casos existente e de conteúdo textual, a base de casos foi importada diretamente para o sistema e convertida em um arquivo do tipo XML. Foi utilizada uma *stoplist* que também foi convertida em um arquivo do tipo XML para ser usada no sistema.

O sistema apresenta dois módulos denominados *Cadastros* e *Busca*. No módulo *Cadastro* é possível: 1) incluir ou excluir FAQ's; 2) inserir e editar os casos das FAQ's (permitindo excluir casos duplicados, por exemplo); e 3) inserir e editar a *stoplist*. No módulo *Busca* a recuperação do caso é feita conforme a consulta realizada. Após a consulta, esse módulo apresenta em uma ordem pré-determinada o número de identificação do caso, os casos mais similares e o valor da medida de similaridade do caso em relação à consulta.

Não houve participação do especialista no domínio para a implementação da recuperação no RBC textual, uma vez que a mesma base que já havia sido categorizada para o RBC estrutural, além de não haver necessidade de estruturação dos casos por meio de atributos nesse tipo de sistema de RBC.

5. Avaliação Experimental

Wallis e Thom (1996) colocam que a avaliação de sistemas de RBC envolve duas medidas na qualidade da solução, denominadas *precisão* (*precision*) e *sensitividade* (*recall*). A *precisão* é a razão entre número de casos relevantes encontrados e o número total de casos recuperados, enquanto a *sensitividade* é o quociente entre o número de casos relevantes encontrados e a quantidade de casos relevantes presentes na base de casos. Estas duas medidas são usadas neste trabalho, sendo intuitivo observar que, quanto maiores tais medidas, melhor será o processo de recuperação.

Inicialmente, o especialista no domínio examinou a base de casos, observando a parte do caso denominado problema e indicando o(s) caso(s) cujas soluções fosse(m) relevante(s) para a solução daquele problema - além da solução contida no caso examinado. Este procedimento foi realizado pelo especialista em todos os casos da base. Por se tratar de um domínio bem delimitado, 55 dos 70 casos da base tinham ao menos um caso que apresentava uma solução relevante. Em seguida foram feitos testes com a recuperação dos casos nos dois sistemas (estrutural e textual, descritos nas Seções 2 e 3 respectivamente).

Para testar a etapa de recuperação (estrutural e textual), utilizou-se um método estatístico de validação cruzada conhecido como "*deixar um fora"* (Witten e Frank 2005). Neste método, simula-se a extração de cada um dos casos da base e, para cada um destes, executa-se o procedimento de recuperação. Para cada recuperação realizada, calcula-se a precisão e a sensitividade. Ao final do processo, obtêm-se valores médios

para precisão e sensitividade. Estes são então reportados para comparar os métodos usados no presente trabalho.

Os cálculos da precisão e da sensitividade levam em conta quais casos foram considerados relevantes pelos sistemas relativamente aos casos que o especialista indicou como relevantes. A fim de tornar a avaliação mais abrangente, foram considerados três cenários diferentes. As diferenças entre os cenários considerados levam em conta o número de casos que cada sistema considera relevante, em função dos casos indicados pelo especialista. Mais precisamente, os cenários utilizados na avaliação podem ser descritos como:

a) o sistema considera como relevantes os cinco melhores casos recuperados (de acordo com cada um dos sistemas avaliados – i.e., estrutural e textual);

b) o sistema considera relevante o melhor caso (maior valor de similaridade) e eventualmente os empates dos valores de similaridade, considerando-se no máximo cinco casos;

c) o sistema considera como casos relevantes casos na mesma quantidade que o especialista indicou como relevantes e eventualmente os empates – incluindo-se no máximo cinco casos.

No cenário (b) a recuperação no RBC textual considerou somente o melhor caso, pois essa recuperação não retornou em nenhum experimento medidas de similaridades idênticas. Analogamente, no cenário (c) a recuperação no RBC textual considerou relevante a mesma quantidade de casos considerados pelo especialista.

O número-limite de casos considerados relevantes pelo sistema (cinco) leva em conta o dia-a-dia dos operadores de *help-desk*, que normalmente localizam a solução nos primeiros casos que o sistema recupera. Além disso, tal limiar (cinco casos) foi a maior quantidade de casos que o especialista indicou como relevantes para os casos da base em questão.

O resultado geral das medidas de qualidade usadas na avaliação das abordagens de RBC estrutural e textual, utilizando os cenários propostos é apresentado na Tabela 5, onde μ_p e μ_s representam respectivamente as médias de precisão e sensitividade, enquanto que σ_p e σ_s correspondem às suas respectivas medidas de desvio-padrão.

Tabela 5. Resultados da avaliação da recuperação

Cenários	**(a)**		**(b)**		**(c)**	
RBC estrutural	μ_p=0,21 σ_p=0,19	μ_s=0,64 σ_s=0,44	μ_p=0,43 σ_p=0,40	μ_s=0,54 σ_s=0,44	μ_p=0,34 σ_p=0,33	μ_s=0,56 σ_s=0,44
RBC textual	μ_p=0,17 σ_p=0,16	μ_s=0,55 σ_s=0,46	μ_p=0,27 σ_p=0,45	μ_s=0,16 σ_s=0,31	μ_p=0,21 σ_p=0,36	μ_s=0,21 σ_s=0,36

A Tabela 5 mostra que resultados melhores foram obtidos pelo sistema de RBC estrutural. Acredita-se que isso se deve, em parte, à participação do especialista durante a modelagem dos casos no RBC estrutural, que leva em conta o problema e a solução para transformar o conhecimento contido no caso em uma estrutura formada pelos pares de atributo-valor. Em outras palavras, conquanto a abordagem estrutural seja mais dispendiosa do que a abordagem textual, e também mais dependente de um especialista no domínio, observa-se que a estruturação dos casos pode fornecer melhores índices de recuperação. A necessidade de estruturar os casos e preencher todos os seus atributos no RBC estrutural faz com que o RBC estrutural seja, num primeiro momento, mais trabalhoso que o RBC textual, onde esse trabalho inicial é praticamente inexistente. Esse trabalho inicial, entretanto, aparentemente favorece uma melhor recuperação.

6. Conclusão e Trabalhos Futuros

Neste artigo, avaliou-se a etapa da recuperação de abordagens de RBC estrutural e textual numa base de casos reais no domínio de *help-desk*. Resultados empíricos mostram que a abordagem estrutural tem melhores resultados. Entretanto, tal abordagem tende a ser menos passível de automatização do que abordagem textual.

Quanto à trabalhos futuros, é possível melhorar os resultados da avaliação da recuperação do RBC estrutural aumentando o número de atributos relevantes para os casos. Isso pode levar o sistema a recuperar um número maior de casos relevantes relativamente aos casos indicados pelo especialista. Dessa maneira, é possível encontrar uma maior quantidade de casos correlacionados entre si. Por outro lado, essa possibilidade aumenta o tempo de preenchimento dos atributos dos casos quando da consulta.

Uma melhoria a ser considerada na recuperação do RBC textual seria a utilização de um dicionário de vocábulos e/ou um dicionário de sinônimos pelo sistema de RBC. Além disso, vale ressaltar que as consultas no sistema de RBC textual foram comparadas somente com as perguntas contidas na FAQ, desconsiderando o campo relativo à *solução do caso*, pois a pergunta é, em princípio, a parte mais relevante para determinar a melhor resposta ao usuário. Essa consideração foi levada em conta para a implementação da recuperação no RBC textual, sendo prática comum em outros trabalhos da área. Entretanto, Burke *et al*. (1997) colocam que é possível melhorar a avaliação da recuperação no RBC textual por meio do uso adicional do campo de *solução do caso*. Essa abordagem sugere um trabalho futuro promissor.

Agradecimentos

Os autores agradecem o suporte financeiro da FAPESP e do CNPq para a realização desse trabalho. Os autores também agradecem a colaboração da *Tree Tools Informática* por disponibilizar uma base de casos real, da Ilog Tecnologia por disponibilizar a *stoplist* e da *Empolis* por ceder a *shell CBR Works.*

Referências

AI-CBR: Information Resources to Case-Based Reasoning Academics and Commercial Developers. Dept. of Computer Science at the University of Auckland, New Zealand. http://www.ai-cbr.org, 2006.

Baeza-Yates, R., Ribeiro Neto, B. (1999), Modern Information Retrieval, ACM Press, New York.

Burke, R., Hammond, K., Kulyukin, V., Lytinen, S., Tomuro, N. & Schoenber, G. (1997) "Question Answering from Frequently Asked Question Files". AI Magazine, 18(2), 57-66.

CBR-Works (2001). CBR-Works 4 - Reference Manual**,** Empolis Knowledge Management GmbH, Revision 2.3.

Dearden, M. A. & Bridge, G. D. (1993) "Choosing a Knowledge Based System to Support a Help Desk". The Knowledge Engineering Review 8(3), 201 – 222.

Empolis Orenge. Technology White Papers. http://www.ovitas.com/PDF/orengeWhite paper.pdf, 2006.

Kolodner, J. L. (1993), Case-Based Reasoning, Morgan Kaufmann, San Mateo.

Lenz, M., Hübner, A. & Kunze, M. (1998) Question Answering with Textual CBR. In. *Proc. of the Int. Conf. on Flexible Query Answering Systems*.

Roth-Berghofer, T. & Iglezakis, I. (2000). Developing an integrating multilevel help-desk support system. In. *Proc. of the 8th German Workshop on Case Based Reasoning*, pages 145-155

Schulz, S. (1999). CBR-Works: A state-of-the-art shell for case-based application building. In *Proc. of the 7th German Workshop on Case-Based Reasoning*.

Wallis, P. & Thom, J. A. (1996) "Relevance Judgements for Assessing Recall". Information Processing & Management 32 (3), 273-286.

Watson, I. (1997), Applying Case-Based Reasoning: Techniques for Enterprise Systems, Morgan Kaufmann, San Mateo.

Wilson, D. & Bradshaw, S. (1999). CBR textuality. In *Proc. of the 4th UK Case-Based Reasoning Workshop*.

Witten, I. & Frank, E. (2005), Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Mateo.

Yang, Q., Kim, E. & Racine, K. (1997). CaseAdvisor: Supporting interactive problem solving and case base maintenance for help desk applications. In *Proc. of the Int. Joint Conference on Artificial Intelligence*, *Workshop on Practical Applications of CBR*.