EMPREGO DE RBC PARA RECUPERAÇÃO INTELIGENTE DE INFORMAÇÕES

UNIVERSIDADE FEDERAL DE SANTA CATARINA
Programa de Pós-Graduação em
Engenharia de Produção

EMPREGO DE RBC PARA RECUPERAÇÃO INTELIGENTE DE INFORMAÇÕES

Fabiano Duarte Beppler

Dissertação apresentada ao
Programa de Pós-Graduação em
Engenharia de Produção da
Universidade Federal de Santa Catarina
como requisito parcial para obtenção
do título de mestre em
Engenharia de Produção

Florianópolis
2002

Fabiano Duarte Beppler

EMPREGO DE RBC PARA RECUPERAÇÃO INTELIGENTE DE INFORMAÇÕES

Esta dissertação foi julgada e aprovada para a obtenção do título de **Mestre em Engenharia de Produção** no **Programa de Pós-Graduação em Engenharia de Produção** da Universidade Federal de Santa Catarina

Florianópolis, 28 de setembro de 2002.

_____

Prof. Edson Pacheco Paladini, Dr.
Coordenador do Curso

BANCA EXAMINADORA

_____

Roberto C.S. Pacheco, Dr.
Orientador

_____

Aran Bey Tcholakian Morales, Dr.

_____

Alejandro Martins Rodriguez, Dr.

A meus pais, Valdemar e Lenita
pelo apoio constante.
À minha Lê...

Agradecimentos

Ao meu orientador, professor Roberto Carlos dos Santos Pacheco, pelo seu acompanhamento, competência, apoio e contribuições valiosas que proporcionaram o aprimoramento deste trabalho.

Aos professores Aran Bey Tcholakian. Morales e Alejandro Martins Rodriguez, pela disponibilidade e contribuição de opiniões e apoio.

Aos amigos do Grupo Stela, por acompanhar o desenvolvimento do trabalho, compartilhando conhecimento de forma direta e indireta.

Aos meus padrinhos, Valcir Paulo Beppler e Sirley Veloso Beppler, pelo incentivo, disponibilidade e o aconchego de sua casa.

Aos meus irmãos, Kaliane, Luís, Márcio e Juliano, que contribuíram para meu desenvolvimento pessoal e profissional, pelo incentivo, críticas construtivas, apoio e grande amizade.

À Letícia Miranda de Miranda, minha Lê, que certamente soube esperar e aturar, mas acima de tudo, soube dar carinho, apoio e incentivo.

Aos meus pais, Valdemar Benjamin Beppler e Lenita Duarte Beppler, pelo exemplo de vida, sempre me apoiando, incentivando e ensinando a lutar e discernir, sem esquecer o lado prazeroso da vida.

Índice

Lista de figuras

Lista de tabelas

Lista de Abreviaturas

CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CRI	Customer Relationship Intelligence
CRM	Customer Relationship Management
IA	Inteligência Artificial
IR	Information Retrieval
MOPs	Pacotes de Organização de Memória
RBC	Raciocínio Baseado em Casos
SGBD	Sistema de Gerenciamento de Banco de Dados
SQL	Structure Query Language
START	Syntectic Analysis using Reversible Transformations
URL	Uniform Resource Locator
XML	Extensible Markup Language

Resumo

A dimensão do volume de informações disponíveis na Internet e as taxas diárias de crescimento tornam cada vez mais presentes mecanismos eficientes e eficazes de recuperação de informações. A maioria dos métodos pesquisados e aplicados tem por base o tratamento das informações disponíveis nos repositórios associados aos sites. Nesta abordagem, um elemento de conhecimento é normalmente negligenciado: a memória das interações efetuadas pelos usuários que utilizaram o site previamente a um usuário atual. A construção desta memória viabiliza o emprego de interações de busca do passado na apresentação de informações desejadas no momento das consultas. A presente dissertação propõe a construção da memória das buscas aos sites na forma de casos de consulta e a aplicação de Raciocínio Baseado em Casos para utilização destas interações passadas como subsídio em novos processos de consulta. O método proposto deu origem à ferramenta RBNet. Para demonstração de sua viabilidade, RBNet foi aplicada ao site de busca do "Diretório dos Grupos de Pesquisa no Brasil", projeto do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). RBNet permite que usuários interessados em grupos de pesquisa possam encontrar rapidamente o que desejam, quando se valem das interações semelhantes registradas na base de casos do RBNet.

Abstract

The extent of information available on the Internet and its daily growth rate make even more necessary efficient and useful methods of information retrieval. Most of the investigated and employed methods are based on available information treatment in the repositories related to the websites. In this approach, a knowledge element is often neglected: the memory of the interactions done by the users that entered the site previously to a current user. The construction of this memory makes possible the use of past search interactions presenting desired information on a new search. The present dissertation aims to propose the construction of the websites searches memory in the form of searching cases and the employment of Case-Based Reasoning to the use of these past interactions as a support to new search processes. The proposed method has originated tool named RBNet. In order to show its feasibility, RBNet was experimented on the searching engine of the “Directory of the Research Groups in Brazil”, a project from the "National Council of Scientific and Technological Development" (CNPq). RBNet allows the users, interested on research groups, to quickly find what they need, when taking advantage of similar interactions stored on RBNet's case base.

1 Introdução

Vivemos num mundo marcado pela produção vertiginosa de conhecimento e pelo progresso tecnológico, onde as transformações são muito rápidas. O conhecimento científico causou grandes inovações que têm beneficiado muito a humanidade. Pessoas com diferentes objetivos fazem pesquisas em sua área visando encontrar soluções para problemas dos domínios em que atuam, e em muitos casos, pessoas de áreas diferentes se unem para obter melhores resultados.

A medida em que se pesquisa há maior necessidade de mecanismos que possam fornecer a informação desejada, preferencialmente de maneira rápida. Um veículo que tem transformado a maneira de se pesquisar é a Internet. A Internet revolucionou metodologias de pesquisa, fornecendo grande diversidade de informação, além de aproximar as pessoas e em especial os pesquisadores (Kuramoto, 1995).

O crescimento desenfreado do número de publicações, a liberdade de organização das informações e o número de usuários fizeram da Internet uma grande fonte de informações heterogêneas. O usuário está interessado apenas em uma pequena parcela das informações disponíveis na Internet. Sendo assim, ele precisa de formas efetivas de acesso, que garantam que as informações para ele disponibilizadas estejam dentro de sua área de interesse. A grande diversidade e a quantidade de documentos disponíveis na Internet aumentaram a dificuldade em recuperar informações relevantes de modo eficiente. O gerenciamento destas informações é uma tarefa difícil e exige mecanismos de estruturação, classificação e filtragem de informações (Veneruchi, 2001). Esse crescimento súbito e sem estrutura definida tornou as ferramentas de busca na Internet uma aplicação essencial para viabilizar o acesso às informações, documentos, serviços e pessoas.

Os sites de busca estão se tornando mais robustos e recuperando informações com maior velocidade. Suas bases de dados estão abrangendo cada vez mais informações disponíveis na Internet. A utilização destes sites de busca, na grande maioria das vezes, retorna quantidade excessiva de informações, sendo que muitas destas possuem pouca ou mesmo nenhuma relevância, se comparada à informação desejada. Para Davemport (1998) a ênfase primária não está na geração e na distribuição de enormes quantidades de informação, mas no uso eficiente de uma quantia relativamente pequena.

Os sites de busca recebem milhares de visitas todos os dias, que representam usuários buscando diversos tipos de informações. A maioria dos sites de buscas procura indexar os conteúdos que dispõe e, muitas vezes, os classifica por características comuns nas informações que contém. Essa abordagem permite, por exemplo, cálculos de semelhança entre conteúdos e apresentação de resultados em ordem provável de interesse dos usuários.

Nesse trabalho, procura-se ampliar essa forma característica de tratar a busca de conteúdos na Internet (com gestão de conteúdos), pela inclusão do tratamento classificável dos critérios e parâmetros utilizados pelos usuários em um site de busca. Uma análise das consultas de todos os usuários permitiria verificar que muitos procuram os mesmos tipos de informações. Quantos *links*, resultantes de uma busca, cada usuário visita em média para suprir sua necessidade? O resultado desta informação é muito variável, pois alguns usuários nas primeiras visitas conseguem encontrar o que desejam, enquanto a maioria pode ficar até horas tentando encontrar a informação desejada, e muitas vezes não a consegue encontrar.

Como forma de publicidade ou mesmo como item contábil, de empresas surgidas com a Internet, os sites procuram manter contadores com o total de usuários que acessaram suas páginas. As questões que se colocam são: "Qual a diferença entre o primeiro usuário que fez a consulta e o centésimo milésimo?", "Como o centésimo milésimo primeiro pode se beneficiar do trabalho dos cem mil anteriores?". Atualmente, a utilização de agentes ou "*cookies*" permite apresentar a usuários compras feitas por usuários que adquiriram o mesmo item do usuário atual (br-business, 2001). De forma análoga, as estratégias de buscas já utilizadas pelos usuários anteriores podem servir de subsídios às buscas dos usuários atuais. O conhecimento proporcionado pelos usuários anteriores, atualmente desperdiçado, pode servir para um novo usuário, ajudando-o a encontrar o que deseja.

Para Davemport (1998), qualquer fornecedor de informação pode agregar valor à informação ao torná-la mais acessível. Ao conduzir o usuário ao local onde se encontram os dados, melhora-se muito a possibilidade destes serem utilizados de maneira eficiente e a informação já obtida pode ser mais facilmente reutilizada.

O presente trabalho propõe a concepção do modelo e a construção de uma ferramenta que armazena as informações de cada consulta efetuada em forma de perfis de consulta e disponibiliza, através da técnica de Raciocínio Baseado em Casos (RBC), a

possibilidade de recuperar os dados das consultas semelhantes (Kolodner, 1993). Com a utilização de RBC há possibilidade de fornecer o percentual de similaridade para cada uma das consultas recuperadas, dando condições de avaliar se a informação buscada está contida em uma das consultas recuperadas. Deste modo, o trabalho realizado por usuários em um site que disponibiliza informações transforma-se em "casos de consulta" para os próximos usuários. O resultado é a possibilidade de que os novos usuários encontrem muito mais facilmente o que procuram, com base nas buscas semelhantes, realizadas anteriormente.

1.1 Justificativa

Segundo Filhoa (2002), as pessoas derivam conhecimento a partir de informações de diversas formas: por comparação, pela experimentação, por conexão com outros conhecimentos e através das outras pessoas. As atividades de criação de conhecimento têm lugar com e entre os seres humanos. O conhecimento é transmitido por pessoas e para pessoas, através de meios estruturados como vídeos, livros, documentos, páginas web, e etc. Além disso, as pessoas obtêm conhecimento daquelas que já o têm, através de aprendizado interpessoal e compartilhamento de experiências e idéias.

A Internet, maior repositório de informações disponíveis no planeta, é também um poderoso instrumento para disseminação do conhecimento, colaboração, comunicação e interatividade (IDG, 2002). A busca de informações na Internet é, muitas vezes, um problema complexo que demanda tempo e conhecimento. Como a quantidade e a diversidade de informações disponíveis na Internet é grande, o principal problema se tornou em como levar somente a informação desejada a quem à procura.

Nesse contexto, ferramentas, métodos e algoritmos que venham facilitar a busca pela informação desejada são de grande relevância, principalmente quando a quantidade e a necessidade de informação crescem diariamente. A Internet, segundo Glória 10 (2001), é utilizada por mais de 340 milhões de pessoas hoje, com previsão para 700 milhões em 2005, todos desejosos cada vez mais por informação. Segundo Zakon (2002), no final de 2000 havia aproximadamente 25 milhões de sites WWW, no final de 2001 este número ultrapassou 36 milhões e atualmente só a ferramenta de busca Google (www.google.com) já efetua busca em mais de 2 bilhões de páginas.

Um dos métodos mais empregados para busca de informação na Internet é a utilização de sites de busca. As consultas a estes sites retornam uma relação de sites com provável relevância com a pesquisa realizada. Um conhecimento precioso que costuma ser desperdiçado na utilização de sites de busca está nos parâmetros das buscas e nas ações seguintes dos usuários. Este conhecimento pode ser armazenado e utilizado para auxiliar as buscas de outros usuários. A inteligência artificial, em especial por meio da técnica de Raciocínio Baseado em Casos, provê mecanismos que auxiliam a resolução de um problema através da resolução de problemas passados. Neste sentido, esta técnica pode ser utilizada em uma ferramenta que transforme as informações das buscas já efetuadas em conhecimento, utilizando este para auxiliar nas buscas de novos usuários.

1.2 Objetivo geral

Este trabalho tem como objetivo principal a concepção e desenvolvimento de uma ferramenta que, por meio de Raciocínio Baseado em Casos, construa, memorize e reutilize perfis de consultas, facilitando a usuários o encontro de informações, baseando-se nas semelhanças entre as diversas seções de buscas, realizadas pelos demais usuários.

1.2.1 Objetivos específicos

1. Estabelecer fundamentação teórica e conceituação acerca do tema em estudo, por exemplo, Raciocínio Baseado em Casos e Recuperação de Informação;
2. Utilização de conceitos de Recuperação de Informação juntamente com a técnica de Raciocínio Baseado em Casos para construção de uma ferramenta que facilite o ato de encontrar informações;
3. Implementar um protótipo, utilizando a base de dados do *Diretório dos Grupos de Pesquisa no Brasil*, desenvolvida pelo CNPq, para validar a ferramenta proposta.

1.3 Metodologia

O presente trabalho fundamenta-se em quatro etapas: (a) estudo sobre recuperação de informação e algumas de suas técnicas; (b) estudo da técnica de Raciocínio Baseado em Casos; (c) apresentação do modelo proposto; e (d) aplicação do modelo proposto utilizando

a base de dados do Diretório dos Grupos de Pesquisa no Brasil. A Figura 1.1 mostra uma visão geral da estrutura do trabalho.

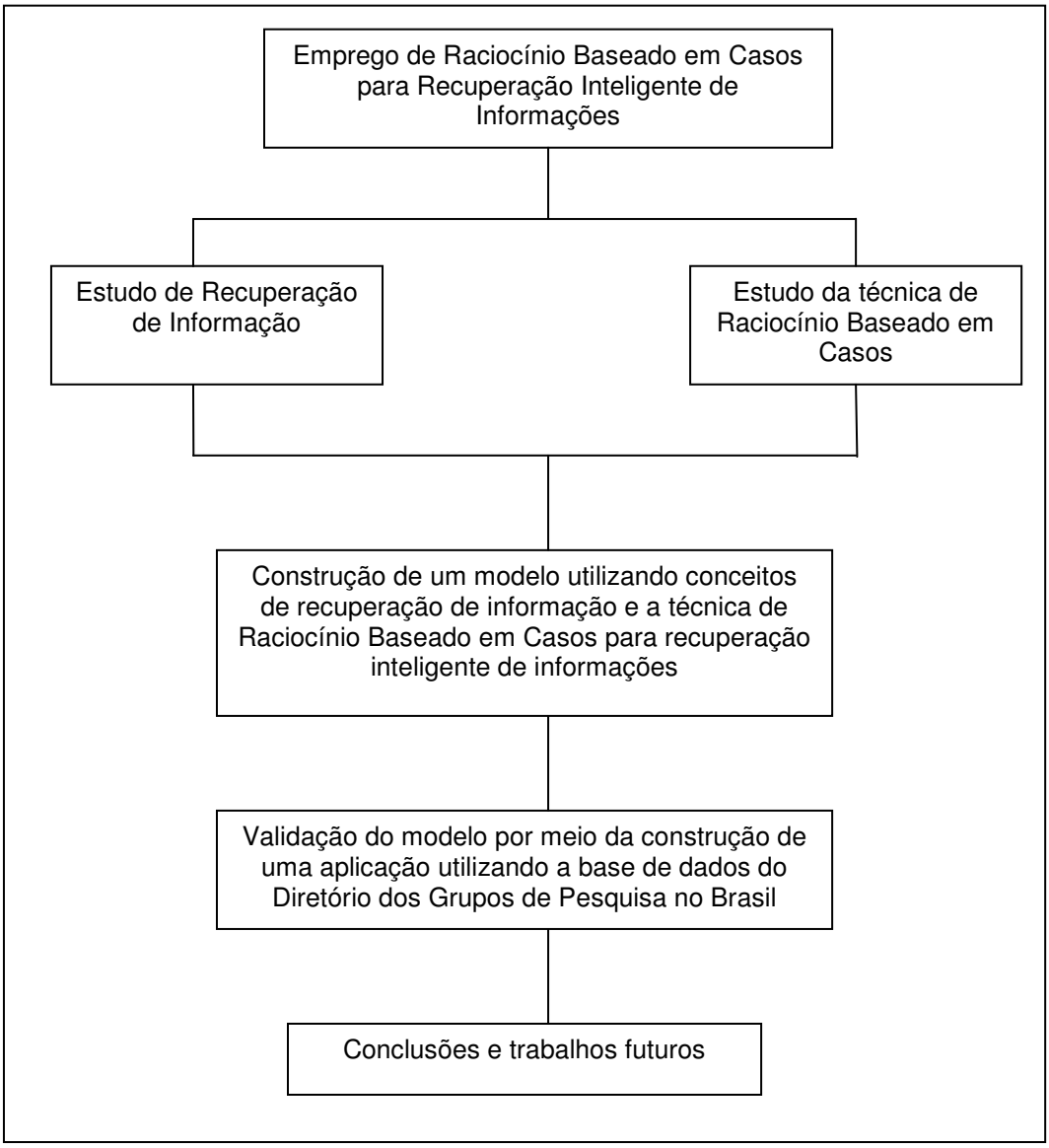


Figura 1.1 - Estrutura do trabalho

A primeira etapa está divida em quatro pontos principais:

- introdução e uma breve visão histórica da utilização de recuperação de informação;
- apresentação de alguns modelos utilizados em recuperação de informação. Os modelos apresentados são: booleano, difuso, vetorial, probabilístico e o modelo utilizando linguagem natural;

- a utilização da técnica *text mining* que além de poder ser utilizada para recuperar informação, extrai conhecimento, sendo esse o seu principal objetivo;
- a utilização da Recuperação de Informação em banco de dados relacionais, detalhando as diferenças entre a Recuperação de Informação clássica e uma abordagem de Recuperação de Informação necessariamente formatada para a utilização em banco de dados relacionais;
- o emprego de algumas técnicas de recuperação de informação na Internet, detalhando alguns de seus aspectos.

A segunda etapa está divida, também, em quatro pontos principais:

- definição do que é Raciocínio Baseado em Casos e um breve histórico de como se originou esta técnica;
- apresentação detalhada da arquitetura de Raciocínio Baseado em Casos, mostrando cada uma das etapas componentes;
- estudo da utilização de Raciocínio Baseado em Casos em recuperação de informação, detalhando as características comuns e as características que devem ser adaptadas;
- a utilização de Raciocínio Baseado em Casos em aplicações na Internet, mostrando a viabilidade e suas vantagens. Esta abordagem é especificada através de exemplos reais.

A terceira parte do trabalho, que descreve o modelo proposto, está divida em três pontos principais:

- uma breve introdução do modelo proposto, oferecendo uma visão geral da viabilidade em utilizá-lo;
- descrição detalhada da arquitetura proposta para a ferramenta;
- apresentação de cada módulo que compõe a arquitetura: os módulos de entrada e saída, o cálculo do grau de similaridade, a base de casos e o perfil das consultas;

A última etapa mostra a utilização do modelo proposto na construção de uma aplicação utilizando a base de dados do *Diretório dos Grupos de Pesquisa no Brasil* e está divida em três pontos principais (CNPq, 2001):

- introdução à aplicação desenvolvida utilizando a base de dados do Diretório dos Grupos de Pesquisa no Brasil, retratando a composição desta base de dados;
- apresentação da aplicação da ferramenta por meio da descrição detalhada de cada uma das fases de implantação e utilização, tais como a formação da base de casos, os módulos de entrada e saída e os cálculos do grau de similaridade;
- apresentação dos resultados descrevendo o protótipo desenvolvido e apresentação de exemplos com a finalidade de mostrar alguns resultados da aplicação do modelo proposto.

1.4 Estrutura do trabalho

Este trabalho compreende cinco capítulos, além do capítulo atual. Estes capítulos visam demonstrar um modelo para recuperação de informações inteligente utilizando a técnica de Raciocínio Baseado em Casos. A Figura 1.2 mostra a estrutura do trabalho, segundo sua divisão em capítulos.

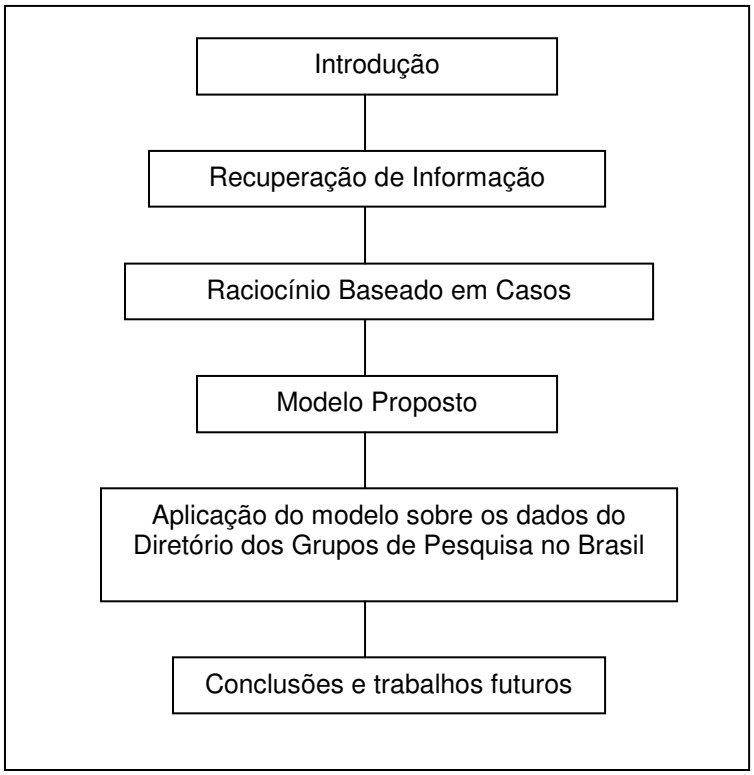


Figura 1.2 - Estrutura do trabalho

O capítulo dois aborda uma introdução à Recuperação de Informação, descreve a arquitetura dos principais modelos utilizados, sua aplicação em conjunto com banco de dados relacionais e resumidamente apresenta a sua utilização na Internet.

O capítulo três descreve a técnica de Raciocínio Baseado em Casos (RBC). Especifica cada uma das fases componentes da arquitetura de um sistema que utiliza esta técnica. Mostra também, sua utilização na recuperação de informação e aborda RBC em aplicações na Internet.

O capítulo quatro apresenta o modelo proposto para recuperação inteligente de informações. Descreve a arquitetura e cada módulo utilizado, como o cálculo do grau de similaridade, a definição e composição da base de casos e a definição dos perfis de consultas.

O quinto capítulo demonstra o modelo proposto por meio do desenvolvimento de uma aplicação utilizando a base de dados referente ao Diretório dos Grupos de Pesquisa no Brasil, projeto do CNPq. Este capítulo descreve o que é o Diretório de Grupos de Pesquisa, como é composta a base de dados utilizada e mostra os resultados da aplicação do modelo, detalhando cada uma das fases.

No sexto e último capítulo são discutidas as conclusões do trabalho e apresentadas recomendações para trabalhos futuros.

1.5 Registro de propriedade intelectual

Um dos principais resultados do presente trabalho foi a concepção e desenvolvimento de um instrumento de indexação e busca inteligente para sites na Internet que tenha disponibilidade de consultas aos seus usuários. Este instrumento denomina-se “RBNet” cuja descrição e registro nacional está sendo providenciado junto ao Instituto Nacional de Propriedade Intelectual.

À referência, consulta ou replicação dos conceitos descritos nesta dissertação aplicam-se além dos preceitos de metodologia científica (referência e citação), os pressupostos do registro de propriedade intelectual.

1.6 Delimitações da pesquisa

A presente pesquisa inclui os seguintes elementos delimitadores:

- Emprego exclusivo de RBC como técnica de indexação e recuperação de consultas semelhantes;
- Abordagem de configuração manual para utilização em sites específicos;
- Necessidade de ajustes para utilização de outros modelos de recuperação de informação, por exemplo, modelo vetorial e modelo utilizando linguagem natural;
- Delimitação adicional: o tipo de conteúdo dos sites deve ser baseado em termos de busca.

2 Recuperação de informação

2.1 Introdução

Durante os últimos anos, um volume crescente de informações tem sido registrado em várias bases de dados, nos mais diversos domínios do conhecimento e sob diversas formas (numéricas, textuais, imagens, etc.). Considerando que os recursos informacionais estão cada vez mais acessíveis aos usuários finais, incluindo os pessoais, o principal problema é saber como acessar tais recursos de forma fácil e precisa. É neste sentido que se faz necessário a utilização de sistemas de recuperação de informação, pois informação é o ingrediente básico para tomada de decisão (Kuramoto, 1995).

Segundo Baeza-Yates e Ribero-Neto (1999), a recuperação, representação, armazenamento, organização e acesso são os processos de gestão da manipulação da informação. Os elementos de representação e armazenamento da informação devem estar dispostos de tal forma a fornecer ao usuário um fácil acesso, isto é, a recuperação da informação desejada.

Na década de 1940 o problema de armazenamento e recuperação de informações atraiu demasiada atenção, pois a quantidade de informação aumentava gradativamente, mas a velocidade de acesso à informação desejada estava cada vez mais lenta, necessitando de maior demanda de trabalho de pessoas. Esta lentidão ao acesso da informação causa, entre outras coisas, a perda de informação relevante, prejudicando a tomada de decisão (Rijsbergen, 2001).

A princípio é muito simples armazenar e recuperar informação. Supõe-se que há documentos em um determinado local e uma pessoa formula uma pergunta. A resposta que satisfaz esta pergunta é um conjunto de documentos. A pessoa analisa este conjunto de documentos, retira os documentos relevantes e descarta os não relevantes. Este relato expressa uma recuperação perfeita de informação. Esta solução é obviamente impraticável, afinal nenhum usuário tem tempo suficiente para suprir sua necessidade lendo todo o conjunto de documentos recuperados, principalmente quando esta quantidade de documentos eleva-se a dezenas de milhares de itens.

Com o advento dos computadores, uma das metas que se viabilizou, passou a ser a construção de sistemas inteligentes para fornecer informações de maneira rápida e objetiva. Em bibliotecas, por exemplo, o elevado volume de informações disponíveis em papel pôde ser administrado de melhor forma com o uso de computadores. Estes podem ser aplicados para gestão da relação de livros e a própria administração da biblioteca. Contudo, o problema da recuperação efetiva de informações continuava sem solução (Rijsbergen, 2001).

A idéia de recuperar informação é simples e objetiva: armazenar toda a informação relevante e recuperar somente a parcela desejada. A questão central está na forma de armazenar esta informação. Logo se notou que o texto em linguagem natural de um documento não só causa problemas no armazenamento, como torna impossível a interpretação do tipo de informação disponível no formato digital, sem a utilização de técnicas avançadas de extração do conhecimento. O principal problema está em como interpretar a informação, isto é, o computador deve saber que tipo de informação contém cada documento, para então, poder recuperar somente os documentos que resolvem uma determinada necessidade, a informação relevante. Isto é possível para um ser humano, que consegue estabelecer a relevância de um determinado documento para uma determinada questão. Para um computador fazer isso é necessário construir um modelo de decisões relevantes que possam ser determinadas de maneira dinâmica.

2.2 Modelos de recuperação

Todas as estratégias e modelos de recuperação de informação são baseados na comparação de uma determinada *query*[1] com documentos armazenados. As distinções feitas entre os diferentes modelos de recuperação de informação podem, algumas vezes, serem entendidas como a procura de uma linguagem de busca que esteja mais adequada à informação que está armazenada. Por exemplo, uma linguagem de busca que permita declarações de termos de busca pode ser expressa por meio de combinações lógicas, necessitando uma verificação para escolher qual modelo de recuperação é mais indicado.

[1] Query – termo em inglês para consulta, aqui utilizado no contexto de consulta a banco de dados.

Segundo Korfhage (1997), há vários modelos de recuperação de informação e suas respectivas estruturas. Os modelos mais freqüentes utilizados são: booleano, difuso, vetorial, probabilístico e modelos que utilizam a linguagem natural.

2.2.1 Modelo booleano

O modelo booleano é um dos mais simples de implementar, sendo utilizado pela maioria dos sistemas comerciais. Um modelo padrão de busca por uma lista de termos é considerado um modelo booleano. A busca booleana é baseada nos conceitos de lógica ou álgebra booleana, onde os termos são ligados por meio de conectores lógicos. As consultas são especificadas como expressões booleanas, formadas através da ligação destes conectores lógicos padrão: AND, OR e NOT (Korfhage, 1997).

O modelo booleano representa os documentos por meio de um conjunto de termos indexados, onde cada índice é uma variável booleana. O índice é considerado verdadeiro para um determinado documento se ocorre neste documento. O valor do status da recuperação é uma medida de similaridade entre a expressão da consulta e o documento armazenado. No modelo booleano o valor do status de recuperação é 1 (um) se a expressão de busca for considerada verdadeira, senão, o valor do status de recuperação é 0 (zero). Todos os documentos cujo valor de status de recuperação é igual a 1 são considerados relevantes para a consulta.

No modelo booleano a *query* que usa o conector AND requer que ambos os termos ligados por este conector estejam presentes no documento. Por exemplo, a expressão "*redes AND neurais*" significa que os documentos recuperados serão somente aqueles que contiverem as duas palavras, "*redes"* e "*neurais*", não influenciando a ordem e as posições destas palavras no documento. Já o conector OR requer que ao menos um dos termos esteja presente no documento. No exemplo anterior, substituindo o conector AND pelo conector OR a expressão torna-se "*redes OR neurais*", significando que os documentos que possuem a palavra "*redes*" ou a palavra "*neurais*" serão recuperados. O conector NOT solicita que um termo específico não esteja presente no documento. Por exemplo, a expressão "*redes NOT neurais*" representa uma *query* em que os documentos recuperados possuem a palavra "*redes*" mas não possuem a palavra "*neurais*".

Com estes conectores há possibilidade de fazer consultas mais complexas, o que provavelmente tornará a recuperação mais demorada. Por meio do uso destes conectores juntamente com a utilização de parênteses para especificar um termo de busca, o usuário pode restringir ou elevar a quantidade de documentos recuperados. Por exemplo, um usuário poderá especificar três termos de busca, sendo que os documentos resultantes deverão conter pelos menos dois termos, como mostra a expressão abaixo.

(redes AND neurais) OR (redes AND conexionistas) OR (neurais AND conexionistas)

Outro detalhe que deve ser observado é a ordem de precedência dos termos com a utilização de conectores lógicos. A ordem de precedência será aplicada primeiro ao conector NOT, depois ao conector AND e por último ao conector OR. Esta ordem leva em consideração a disposição destes conectores e de seus termos entre parênteses. O seguinte exemplo mostra como será interpretada a expressão dentro de um modelo lógico.

redes OR neurais AND conexionistas = *redes OR (neurais AND conexionistas)*

redes OR neurais AND conexionistas ≠ *(redes OR neurais) AND conexionistas*

Neste modelo não é possível ordenar a importância dos documentos de saída, pois o valor do status de recuperação será sempre o mesmo. Por exemplo, se um usuário monta uma expressão de busca contendo dez termos ligados pelo operador AND, o documento que tiver somente nove destes termos não será recuperado.

Buscas utilizando este modelo podem ser feitas sobre arquivos texto ou sobre arquivos estruturados, como em banco de dados relacional, por exemplo. Em banco de dados relacional, o SGBD (Sistema de Gerenciamento de Banco de Dados) constitui-se de uma coleção de programas que permitem criar e manter um banco de dados. Por meio deste, por exemplo, é possível recuperar os dados de um banco de dados sendo que a linguagem básica utilizada é o SQL (*Structure Query Language*). Esta linguagem está baseada no modelo de busca booleano (Filho, 2002). A seguir, apresenta-se um exemplo de *query* para recuperar dados de uma base de dados relacional. O objetivo da *query* apresentada é buscar nomes de pessoas que nasceram no ano 1950 na cidade de Florianópolis.

select nome_pessoa from tabela_pessoa where

ano_nascimento = 1950 AND local_nascimento = "Florianópolis"

A relevância no *feedback* do usuário é freqüentemente utilizada em sistemas de recuperação de informação e serve para melhorar os resultados da recuperação. Normalmente, o usuário é questionado no sentido de indicar a relevância ou irrelevância de alguns documentos recuperados. Considerando que os documentos recuperados não são ordenados, a relevância dos documentos, neste modelo, torna-se difícil.

2.2.2 Modelo difuso

Criada em 1965 por Loft A. Zadeh (1965), a Teoria de Conjuntos Difusos redefine o conceito clássico de pertinência de um elemento a um conjunto, por meio da redefinição do conceito de função de pertinência do intervalo fechado e discreto [0, 1] € I para o intervalo contínuo [0, 1] € IR. Assim, um conjunto difuso é {x € Ā | μ_x(A) € [0, 1] € IR}.

Ao generalizar a função característica dos conjuntos clássicos, a Teoria dos Conjuntos Difusos fez com que todas as áreas que fundamentaram seus desenvolvimentos em teoria clássica de conjuntos pudessem ser generalizadas para uma teoria difusa. Esse é o caso da busca difusa, que generaliza a busca booleana.

O modelo difuso de recuperação de informação é baseado na teoria dos conjuntos difusos. Os operadores lógicos apropriados são redefinidos para incluir partes de um conjunto, onde o processamento da consulta do usuário ocorre, de forma similar ao utilizado no modelo booleano (Gudvada *et al*., 1997).

A lógica difusa pode ser definida como a lógica que fundamenta sistemas de raciocínio definidos como aproximados, ao invés de exatos. Isto difere de sistemas lógicos tradicionais em suas características e detalhes. Na lógica difusa o raciocínio exato corresponde a um caso limite do raciocínio aproximado, sendo interpretado como um processo de composição difusa. O valor verdade de uma proposição pode ser um subconjunto difuso de qualquer conjunto parcialmente ordenado, ao contrário dos sistemas lógicos binários, onde o valor verdade só pode assumir dois valores: verdadeiro ou falso (Takemura, 2001).

A justificativa para recuperação difusa de informação está no fato de que o sistema e freqüentemente o usuário não podem informar que um dado documento fornece toda a informação necessária. Esta incerteza está modelada na avaliação difusa do documento com respeito à busca. Desta forma a recuperação difusa de informação é relatada como um sistema clássico de recuperação de informação que calcula algumas medidas de relevância para um documento e apresenta os documentos recuperados por ordem de grau de relevância. O conceito de recuperação difusa de informação permite avaliar ambos, documento e *query* (Korfhage, 1997).

No modelo difuso o usuário pode especificar um peso, relevância, para cada termo da busca. Este peso é utilizado no cálculo do grau de relevância, que é o processo de comparação dos termos de busca com todos os documentos presentes na base de dados. A utilização de pesos para cada termo pode ser prejudicial para a busca, porque a má utilização destes pode representar a recuperação de conjuntos de documentos sem relevância (Wives, 2000). Por exemplo, considerando a expressão de busca "*(computação) 0.4 (inteligência artificial) 0.8"* significa que os documentos que possuem o termo "*inteligência artificial*" são mais relevantes que os documentos que possuem o termo de busca "*computação*", caso apareçam separados, e nos documentos que aparecem juntos a relevância é maior.

Na recuperação difusa de informação há, também, a possibilidade de usar operadores difusos, que podem facilmente exercer uma melhor comparação dos documentos com uma determinada expressão de busca (Wives, 2000). Estes operadores podem ser qualitativos ou quantitativos. Os operadores qualitativos incluem descritores numéricos, tais como, *baixo*, *alto* e *grande*, e descritores não numéricos (lingüísticos) como *bonito* e *colorido*. Os operadores quantitativos incluem palavras como *pouco* e *muito*. O uso destes operadores em uma *query* faz com que os documentos resultantes ao final da busca possam ser classificados como mais relevantes ou menos relevantes, por exemplo. O problema é tornar estes operadores parte de funções associadas com a recuperação difusa. Em muitos casos, o processo de comparação e a verificação da relevância do documento podem ser combinados com a utilização de outras técnicas, como o processo de comparação utilizado no modelo booleano (Korfhage, 1997).

Um modelo que utiliza a busca por meio de *queries* difusas é apresentado por Ribeiro e Moreira (2002) no trabalho sobre um sistema inteligente que busca informações

das 500 maiores empresas não financeiras de Portugal. Este trabalho tem o objetivo de obter respostas específicas sobre as características dessas empresas. O usuário insere uma *query*, em forma de linguagem natural, seguindo algumas regras e dispondo de alguns operadores difusos, como, por exemplo, *muito*, *pouco* e *produtivo*. Ou, caso seja um usuário experiente, pode usar opções de configurações avançadas. O sistema interpreta a *query* (são as informações inseridas pelo usuário, seja por meio de linguagem natural ou por seleção de opções avançadas) através de um *parser*[1], disponibilizando as informações de maneira estruturada que são utilizadas pelo mecanismo que contém as funções difusas. Esse mecanismo interpreta essas informações e processa a "defusificação", gerando regras que são aplicadas sobre uma base de dados relacional, objetivando recuperar as informações que satisfaçam a *query* previamente inserida. Se, por exemplo, digita-se a seguinte *query*: "*Muitas empresas no centro de Portugal tem lucro em vendas?*", o sistema gera a resposta afirmando que *"muitas empresas tem lucro*" e apresenta o percentual de acerto, que nesse caso é 100%. Como parte da resposta, ainda são apresentadas algumas informações resultantes das equações componentes do módulo difuso, indicando o porquê da resposta. Como complemento de resposta do exemplo acima, são apresentados alguns indicadores que justificam a resposta.

2.2.3 Modelo vetorial

No modelo vetorial cada documento é representado por um vetor ou por uma lista de termos ordenados, ao invés de um conjunto de termos como utiliza o modelo booleano. A diferença entre o modelo booleano e o modelo vetorial está na representação do termo individual e nos métodos de determinação da similaridade entre o documento e a *query*.

Na recuperação do modelo booleano, a similaridade entre um documento e a *query* é baseada na presença exclusivamente dos termos da *query* no documento. No modelo vetorial o cálculo de similaridade para recuperação de informação utiliza um cálculo mais sofisticado. Os métodos de avaliação podem ser baseados num vetor de zeros e uns, sendo que se o termo da *query*, que ocupa uma posição no vetor, não estiver no documento, seu valor será 0, e, caso contrário, se o termo estiver presente no documento o valor será 1. Outra opção a ser utilizada é por meio de um vetor de pesos, onde cada componente do

[1] Parser: programa que subdivide uma entrada (por exemplo, uma frase texto) para que um outro possa atuar sobre ela. Analisador gramatical.

vetor terá um peso associado. Este vetor conterá todos os termos, que podem ser encontrados nos documentos a serem recuperados, juntamente com o valor do seu peso.

O modelo vetorial que utiliza um vetor de pesos tem como característica a esparsidade. A grande maioria dos vetores que representam um documento terá o valor dos pesos igual a zero. Dessa forma, embora a utilização do vetor seja um modelo adequado na teoria matemática para a avaliação da similaridade e a recuperação, na prática uma representação compacta é mais usada, consistindo somente nos pesos dos termos presentes no documento. A chave do sucesso na utilização desse modelo é manter compatibilidade dimensional, isto é, o sistema precisa ser desenvolvido para assegurar que a comparação de dois documentos (ou entre um documento e uma *query*) seja baseada na comparação de termos iguais para cada documento (Korfhage, 1997).

Associar pesos aos termos de um documento contidos em um vetor é um processo complexo. Uma maneira é considerar a freqüência de ocorrência do termo no documento, pois supostamente o termo que mais aparece no documento poderá ser considerado mais importante. Esta maneira não deverá ser adotada para a construção do vetor contendo os termos de busca resultante da *query*. Outra forma é oferecer ao usuário a possibilidade de indicar, através de valores contidos em um intervalo, os pesos relativos a cada termo da *query*. Mesmo utilizando este método o usuário poderá encontrar problemas com a definição dos valores dos pesos, porque a indicação de pesos irá influir diretamente no cálculo de similaridade e um termo com peso sub-dimensionado ou super dimensionado afetará diretamente a recuperação, podendo resultar documentos sem importância (Korfhage, 1997).

A recuperação de informação utilizando funções de similaridade levanta uma questão: quais documentos deverão ser considerados e quais documentos deverão ser rejeitados, deduzindo que cada documento armazenado irá receber um grau de similaridade? Os documentos poderão ser ordenados pelo grau de similaridade, desta forma, os documentos com grau de similaridade maior serão recuperados primeiro. Quanto ao total de documentos recuperados, poderá ser definida uma quantidade máxima ou um grau de similaridade mínimo, restringindo a recuperação de documentos com grau de similaridade igual ou superior ao estipulado.

Um exemplo do uso de recuperação de informação com o modelo vetorial é o sistema de busca de informações na Internet apresentado por Yuwono *et al* (2001). O

sistema possui um *Web Robot* para adicionar e manter atualizadas as informações em uma base de dados indexada, utilizada para pesquisa por meio de palavras-chaves. O *Web Robot* extrai as informações (palavras-chaves) de um documento e monta um vetor, armazenando-o em uma base de dados, base de vetores. Ao inserir as informações para gerar uma *query*, é disparado o processo de busca sobre a base de vetores. Da *query* é gerado um vetor que será comparado com os vetores da base de vetores. O grau de similaridade é medido por uma função que utiliza as informações do vetor da *query* e do vetor que representa cada documento. Cada componente de um vetor corresponde ao peso de uma palavra ou termo em uma *query* ou documento. O valor deste peso é gerado por uma função de freqüência. Terminada a comparação, o resultado é a relação de documentos que tiverem grau de similaridade igual ou superior ao definido pelo usuário, ordenados pelo maior grau de similaridade.

2.2.4 Modelo probabilístico

Quando se procura informação deseja-se obter somente documentos relevantes, descartando os documentos que não são relevantes. Para considerar um documento relevante entre um conjunto de documentos, há necessidade de conhecer estes documentos. Mas na recuperação de informação quem irá decidir se os documentos serão relevantes ou não é o próprio usuário. A recuperação de informação utilizando o modelo probabilístico tem o objetivo de determinar quais documentos podem ser relevantes para uma determinada *query*.

A probabilidade de relevância neste modelo oferece uma noção estatística e não semântica. Porém, o grau de relevância calculado com análises estatísticas tende a ser muito similar ao cálculo utilizado no semântico. Este modelo é similar ao modelo difuso, mas impõe a necessidade de funções de avaliação usadas para probabilidades. Especificamente, em um modelo probabilístico, o conjunto de documentos retornados de uma *query*, supostamente consiste em documentos que satisfaçam esta *query* com uma probabilidade mais alta que a probabilidade especificada. A principal diferença deste modelo para o modelo difuso é que há necessidade de algumas regras de probabilidades serem satisfeitas. Por exemplo, há determinados termos que possuem regras específicas (Rijsbergen, 2001a).

A freqüência de termos no documento pode ser usada para estimar a probabilidade dos termos de uma *query* em relação a um documento, determinando se o documento satisfaz ou não a *query*. Ainda sobre este aspecto há a possibilidade de armazenar os termos mais utilizados na recuperação de um conjunto de documentos e utilizá-los no cálculo da probabilidade (Korfhage, 1997).

O sistema INQUERY, desenvolvido pelo *Center for Intelligent Information Retrieval* (http://ciir.cs.umass.edu/index.html) da universidade de Massachusetts, usa o modelo probabilístico para recuperação de informação baseado em redes *Bayseanas*. No modelo probabilístico a relevância estimada de um documento em relação a uma *query* é uma função de probabilidade, na qual cada um dos vários termos de um documento ocorre no mínimo num documento relevante, mas não em documentos irrelevantes. Uma rede *Bayseana* é um modelo gráfico que codifica relacionamentos probabilísticos entre variáveis de interesse, por isso a vantagem de usá-la nesse tipo de modelo de recuperação de informação. O sistema INQUERY constrói dois tipos de redes, uma rede de documentos e a outra rede de *query*. A rede de documentos é estática para uma dada coleção. Os nós representam documentos que estão conectados aos nós que representam os termos. A rede de *query* é construída pela conexão de termos da *query* aos nós que representam como estes termos devem ser combinados com documentos relevantes. Para fazer a recuperação o sistema conecta estas duas redes, uma a outra, e calcula a probabilidade condicional de qual informação necessita existir em um documento. O sistema então ordena os documentos por esta probabilidade (Miller *et al,* 2002).

2.2.5 Modelo utilizando linguagem natural

Mesmo não havendo nenhuma restrição quanto à formalidade de *queries* nos outros modelos apresentados, o usuário se sente mais à vontade utilizando uma descrição em linguagem natural do seu problema. Muitos sistemas de recuperação de informação usam um módulo intermediário que interpreta o problema descrito em linguagem natural e o transforma em uma *query* para consulta.

Uma *query* em linguagem natural é muito fácil de formular, contudo é imprecisa, inexata e freqüentemente não respeita regras gramaticais. O usuário pode formular uma pergunta em que a informação necessária para utilizar na recuperação não está descrita explicitamente. Com o aumento crescente do número de usuários, o interesse por

desenvolvimento de sistemas que suportem a recuperação de informação por meio de linguagem natural está aumentando (Korfhage, 1997).

Para que um sistema computacional interprete uma sentença em linguagem natural, é necessário manter informações morfológicas, sintáticas e semânticas armazenadas em um dicionário, juntamente com as palavras que um sistema compreende. As etapas do processo de linguagem natural, segundo Oliveira (2001), são:

- **análise morfológica:** identifica palavras ou expressões isoladas em uma sentença, sendo este processo auxiliado por delimitadores (pontuação e espaços em branco). As palavras identificadas são classificadas por seu tipo de uso ou, em linguagem natural, categoria gramatical;
- **análise sintática:** através da gramática da linguagem a ser analisada e das informações do analisador morfológico, o analisador sintático procura construir árvores de derivação para cada sentença, mostrando como as palavras estão relacionadas entre si;
- **análise semântica:** significados compõem as estruturas criadas pelo analisador sintático. A questão da representação do significado apresenta diversas dificuldades. Pode-se mencionar a questão dos significados associados aos morfemas componentes de uma palavra (mercado, hipermercado), a questão da ambigüidade (tomar, em "tomar de alguém", em "tomar um banho" ou em "tomar suco"), ou diferenciação entre significado e sentido ("casa", "minha casa"). O analisador semântico analisa e verifica as estruturas das palavras que foram reagrupadas pelo analisador sintático, uma vez que o analisador morfológico permite identificar estas palavras individualmente;
- **pragmática:** à medida em que se avança é necessário fazer uma interpretação do todo e não mais analisar o significado de suas partes, do ponto de vista léxico e gramatical.

No processamento em linguagem natural, mesmo com a aplicação de suas etapas, a clareza da informação ainda pode estar implícita, sendo identificada através de deduções. Por exemplo, se considerar a seguinte pergunta "*Você sabe a hora?*", não significa que a resposta deva ser *sim* ou *não*. Mesmo com a aplicação destas regras é difícil determinar o

que se quer buscar, pois uma *query* apresenta muito pouca informação, tornando a extração da informação relevante para a busca, muitas vezes, subjetiva.

O sistema START (*SynTactic Analysis using Reversible Transformations*) consiste em dois módulos que compartilham a mesma gramática. O módulo de compreensão analisa o texto em inglês e produz uma base de conhecimento que incorpora a informação existente num texto. Dado um segmento apropriado da base de conhecimento, o módulo de geração produz sentenças em inglês. Um usuário pode recuperar a informação armazenada na base de conhecimento formulando perguntas em inglês. Dada uma sentença, em inglês, contendo várias cláusulas relativas, contradições, etc, o sistema START divide em pequenas unidades, chamada sentenças *kernel* (essas sentenças normalmente contêm um verbo). Separadamente analisa cada sentença *kernel*, organizando todos os elementos em uma árvore, construindo um conjunto de estruturas representacionais. Estas estruturas são feitas com o número de campos correspondendo a vários parâmetros sintáticos de uma sentença. Os três parâmetros mais importantes, assunto da sentença, objeto e a relação entre eles são selecionados porque representam uma regra especial na indexação. Estes parâmetros são explicitamente representados em uma rede de diferenciação para recuperação eficiente. Como resultado, todas as sentenças analisadas são indexadas como uma expressão ternária (*ternary expressions*). Outros parâmetros (adjetivos, nomes possessivos, etc) são usados para criar novas expressões ternárias, na qual preposições e várias palavras especiais podem servir como relações (Katz, 2002).

2.3 Recuperação de informação e banco de dados relacionais

Na visão clássica das duas áreas, recuperação de informação e banco de dados relacionais, a recuperação de informação não necessita dos dados formatados, pois a origem de sua informação está na forma de textos, enquanto que para os bancos de dados há a necessidade de dados formatados. Para formatar e utilizar os dados em banco de dados é necessário modelos de dados e linguagens de busca. As áreas de pesquisa estão focando principalmente a recuperação de textos, métodos de análises de textos, peso de termos e modelos de recuperação baseados na inferência de incerteza. Devido aos diferentes ambientes de aplicações, o campo de banco de dados tem desenvolvido métodos para trabalhar com integridade, segurança, concorrência e recuperação de dados. A grande

maioria das aplicações que utilizam banco de dados relacional é voltada para área comercial com recuperação dos dados baseada no modelo booleano (Fuhr, 2001).

A independência de dados lógicos e físicos é fundamental para o conceito de banco de dados. Esta independência direciona a três níveis de organização: física, lógica e de nível externo. Infelizmente, a independência de dados é um conceito totalmente desconhecido em recuperação de informação, pois as tarefas são divididas em indexação e recuperação. Implementar independência de dados lógicos em recuperação de informação significa que uma *query* pode conter condições sem relação com um processamento através de um índice, havendo métodos para validar esta condição em banco de dados. A formulação da *query* para recuperação de informação deverá ser independente da presença de um índice (Fuhr, 2001).

Rijsbergen (2001a) mostra um exemplo, apresentado na Figura 2.1, de documento para um típico modelo utilizando recuperação de informação acoplado a um banco de dados relacional. A primeira coluna, do lado do texto em linguagem natural, contém dados de fatos de diferentes ordens (exemplo: datas, nomes e instituição), que são os dados formatados. Além da recuperação de informação contida nos textos, os usuários também podem querer buscar por dados formatados, havendo a necessidade de busca por informação por meio da forma tradicional dos bancos de dados relacionais. Neste caso há necessidade de prever as duas situações:

- aumentando a recuperação de documentos, os usuários podem querer extrair informações sobre diferentes tipos de objetos que ocorrem no conjunto de documentos recuperados, isto é, quais termos indexados ocorrem mais freqüentemente no conjunto de documentos recuperados? Ou de quais autores são os documentos?
- para algumas aplicações, há necessidade de extrair outras informações do banco de dados, que consiste na resposta de objetos que não são documentos. Por exemplo, um pesquisador procura por colegas de trabalho ou pesquisadores que atuam na mesma área, como saber a instituição onde atuam estes pesquisadores? O que um determinado pesquisador fez em uma determinada área?

L2	ANSWER 1 OF 111
AN	87(01):411 NTIS Order Number: AD-A172 502/7/XAD
TI	Controlling Inference. (Doctoral thesis)
AU	Smith, David E.
CS	Stanford Univ., CA. Dept. of Computer Science
NC	Contract: N00014-81-K-0004
NR	AD-A172 502/7/XAD; STAN-CS-86-1107 197p. NTIS Prices: MF A01 Availability: Microfiche copies only.
PD	860400
LA	English CY United States
OS	GRA&18701
AB	Effective control of inference is a fundamental problem in Artificla Intelligence. Unguided inference leads to a combinatorial explosion of facts or subgoals for even simple domains. To overcome this problem, …
CC	95F Bionics and artificial intelligence
CT	* Artificial intelligence; data bases; decision making; global; information exchange; problem solving; efficiency *Inferece; control; theses expert systems
UT	NTISDODXA

Figura 2.1 - Modelo de documento em uma base de dados bibliográfica

Um modelo utilizando um banco de dados relacional, que possibilite a recuperação de dados utilizando a linguagem de consulta do banco de dados relacional e uma técnica de recuperação de informação oferece ao usuário uma ferramenta de busca poderosa, dando a oportunidade de busca por informação, não somente através dos textos dos documentos utilizando os conceitos de recuperação de informação, mas possibilitando a utilização como uma ferramenta de extração de dados formatados. Esta junção pode ser feita por meio de um modelo de banco de dados relacional bem projetado e uma linguagem de busca que possa prover tais características.

2.4 Recuperação de informações na Internet

O crescimento desenfreado do número de publicações, a liberdade de organização das informações e o número de usuários fizeram da Internet uma grande fonte de informações heterogêneas disponíveis aos mais variados tipos de interesses e de necessidades. O gerenciamento desta grande quantidade de informações é uma tarefa difícil e exigirá mecanismos de estruturação, classificação e filtragem de informações. O usuário está interessado apenas em uma pequena parcela das informações disponíveis na Internet. Sendo assim, há necessidade de formas efetivas de acesso que garantam a

recuperação de informações relativas somente à sua área de interesse (Gudivada et at., 1997).

A recuperação de informações na Internet é uma das atividades mais utilizadas pelos usuários. Contudo, segundo Huang (1999), comparada com esta recuperação de informação, a chamada recuperação de informação clássica, os sistemas de recuperação de informações na Internet requerem atenção especial sobre alguns aspectos:

- **volume de dados:** a quantidade de informação é muito grande;
- **Internet é dinâmica:** as informações na Internet estão em constantes alterações;
- **informação heterogênea:** a Internet contém grande diversidade de documentos;
- **variedade de linguagens:** os documentos podem estar em diferentes tipos de linguagens, por exemplo, inglês ou português;
- **duplicação:** a quantidade de informação duplicada é grande;
- **alta integração:** cada documento pode ter *links* para outros documentos, e estes para outros;
- **comportamento específico:** há uma estimativa, ainda segundo Huang (1999), de que 85% dos usuários, que utilizam sites de busca não olham a segunda tela com os resultados da busca.

Atualmente, as buscas na Internet são feitas de duas formas básicas. A primeira ocorre manualmente, através do conhecimento prévio de um determinado endereço URL (*Uniform Resource Locator*). A segunda é feita por meio de palavras-chave em servidores que possuem documentos indexados, através de sites de busca, como o *Altavista* (www.altavista.com), *Google* (*www.google.com*), *Excite* (www.excite.com) e outros.

Esses mecanismos de buscas utilizados atualmente na Internet são compostos de duas partes: um *web robot* que percorre os documentos cadastrando as partes interessantes em um banco de dados com índices específicos e uma ferramenta de consulta que permite que os usuários, por meio de um navegador, formulem consultas sobre os índices destes bancos de dados. Mesmo com o apoio destes mecanismos de busca, o crescimento do número de documentos na Internet faz com que as consultas simples gerem resultados

quantitativamente volumosos, contendo muitos *links* irrelevantes, repetidos ou, muitas vezes, inacessíveis.

Embora a utilização destes instrumentos de busca agilize o processo de recuperação de informações na Internet, ainda existem outros dois problemas: (a) a sobrecarga de informação, que corresponde ao fato do usuário receber mais informações do que é capaz de processar e assimilar individualmente; e (b) a relação de uma mesma consulta várias vezes por usuários diferentes, gerando resultados praticamente iguais. Estas informações devem ser analisadas e incorporadas às técnicas de recuperação de informações na Internet, viabilizando um sistema com capacidade personalizada para recuperar somente informações desejadas.

Atualmente, sistemas para filtragem de informações tentam monitorar o interesse dos usuários sugerindo-lhes novas informações coletadas na Internet, além de aproveitar a capacidade de análise e avaliação humana para recomendar informações a outros usuários com áreas de interesse comum, criando grupos de usuários com interesses afins (Veneruchi, 2001).

Pode-se classificar o processo de filtragem em três categorias: (a) **filtragem personalizada**, onde as informações que não satisfazem os interesses do usuário são eliminadas; (b) **filtragem colaborativa**, que visa à criação de ferramentas que apóiem o usuário na busca de informações, tendo como uma das suas características a identificação de usuários comuns e o compartilhamento de conhecimentos, experiências e informações de forma transparente entre os mesmos; (c) **sistemas de recomendação de informação**, representando o processo através do qual informações podem ser sugeridas a um ou mais usuários baseados nas suas necessidades de informação individual ou coletiva, podendo combinar filtragem personalizada ou colaborativa (Veneruchi, 2001).

Dado o volume crescente de informações publicadas e distribuídas na Internet, os usuários que dedicam a maior parte do tempo à realização de trabalhos correspondentes às suas áreas de interesse, podem se valer dos sistemas de recomendação para que, de forma bastante automática, procurem e recomendem informações em concordância com as necessidades individuais ou coletivas dos usuários.

Como ocorre na recuperação de informação usando linguagens de consulta, um grande problema no desenvolvimento de ferramentas de filtragem é o tratamento de

informações não estruturadas. Isto representa uma grande dificuldade, pois o sistema deve criar um modelo de documento para determinar o tipo de conteúdo do mesmo.

2.5 Resumo esquemático

A revisão dos cinco modelos de Recuperação de Informação, técnicas de *Text Mining*, recuperação de informações em banco de dados e recuperação de informações disponíveis na Internet permite a construção de um resumo esquemático, contido na Tabela 2.1, das principais questões e variantes do termo de recuperação de informação.

Modelo	Forma de busca	Característica
Booleano	Expressões lógicas	▪ simples de implementar ▪ não permite ordenação dos resultados ▪ não permite linguagem natural ▪ buscas sobre arquivos textos ou estruturado
Difuso	Expressões lógicas difusas	▪ dificuldade de implementar ▪ permite ordenação dos resultados ▪ não permite linguagem natural ▪ indicação de peso para cada termo a ser buscado ▪ operadores difusos
Vetorial	Vetores de termos	▪ dificuldade de implementar ▪ permite ordenação dos resultados ▪ indicação de pesos para os termos de busca ▪ restrição de resultados por grau de similaridade
Probabilístico	Termos de busca	▪ dificuldade de implementar ▪ noção estatística para probabilidade de relevância ▪ permite ordenação de resultados ▪ necessidade de regras probabilísticas satisfeitas ▪ regras específicas para termos
Linguagem Natural	Linguagem natural	▪ dificuldade de implementar ▪ facilidade para fazer buscas ▪ regras diferentes para cada língua
Text Mining	Resultados formatados	▪ buscas sobre arquivos textos ▪ dificuldade de implementar ▪ recuperação de conhecimento de textos ▪ utilização de técnicas de *data mining*
Banco de dados	Expressões lógicas	▪ fácil de implementar ▪ necessidade de dados formatados ▪ formalização para consultas utilizando lógica booleana ▪ busca sobre dados estruturados ▪ não permite ordenação do resultado

Tabela 2.1 Resumo esquemático de Recuperação de Informação

Como se pode perceber, o tema de Recuperação de Informações prevê a análise de modelo e lógicas de recuperação (com impacto na forma e na flexibilidade com que os usuários farão suas consultas) e a origem e formato dos dados que se deseja recuperar (com impacto tanto na viabilidade do modelo de recuperação como, principalmente, no método

utilizado). A Figura 2.2 ilustra de forma esquemática o problema da recuperação de informação.

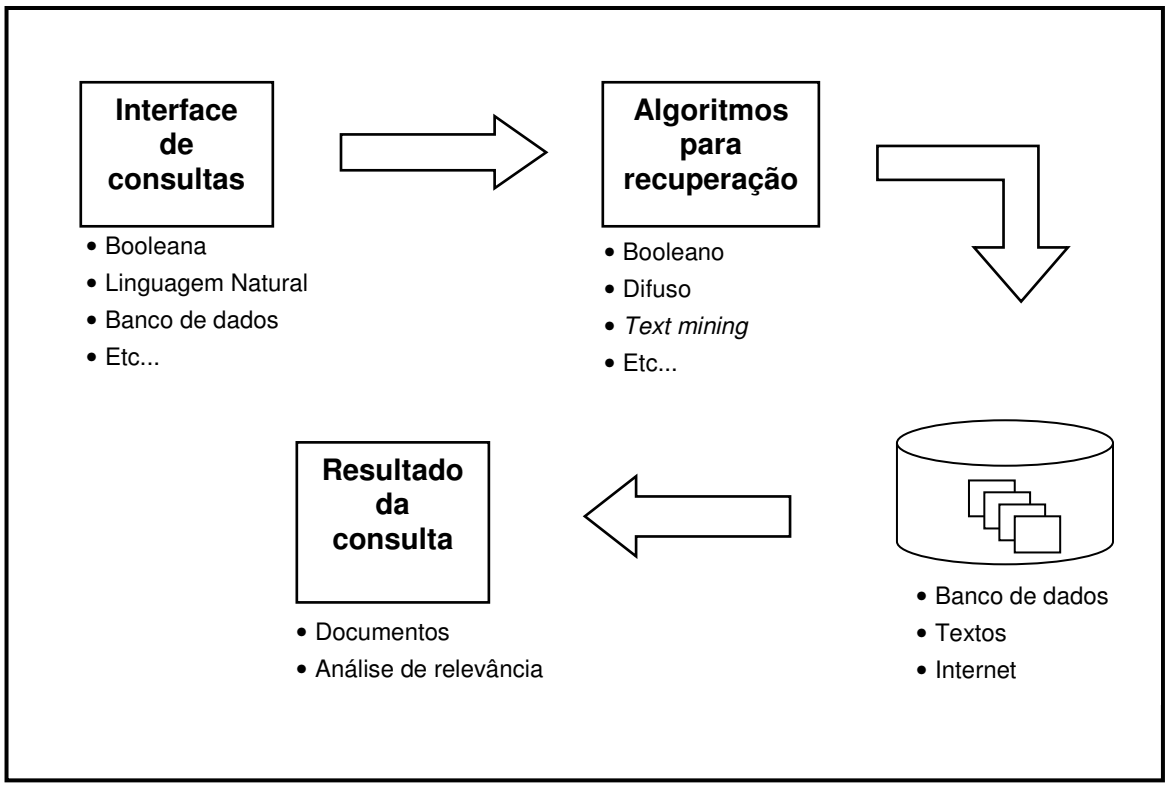


Figura 2.2 Esquema de recuperação de informação

2.6 Considerações finais

Este capítulo abordou conceitos e técnicas de recuperação de informação. O estudo apresenta um breve comentário histórico da recuperação de informação e principalmente os motivos que culminaram na criação desta técnica.

Foram apresentadas também, arquiteturas utilizadas para recuperação de informação, discutindo-se o modelo conceitual de cada uma e indicando onde sua utilização melhor se aplica. Outro aspecto discutido refere-se à utilização de recuperação de informação em conjunto com banco de dados, citando uma arquitetura para utilização em conjunto de conceitos particulares de recuperação de informações e banco de dados relacional.

A Internet como fonte de informação torna-se um ambiente rico em possibilidades para desenvolvimento de aplicações utilizando técnicas de recuperação de informação,

uma vez que as máquinas de busca atuais não conseguem fornecer as informações de maneira rápida e precisa.

A presente dissertação pode ser contextualizada na área de recuperação de informações como a proposição de ferramenta de filtragem colaborativa (Veneruchi, 2000). A cooperação entre usuários ocorre de forma indireta, por meio da disponibilização de consultas de usuários anteriores ao usuário atual, armazenadas, recuperadas e processadas como "casos de consultas", comparadas ao caso do usuário atual. Para tal, a ferramenta proposta tem base na técnica de IA Raciocínio Baseado em Casos (RBC), abordada no próximo capítulo.

3 Raciocínio Baseado em Casos (RBC)

3.1 Introdução

Raciocínio Baseado em Casos (RBC) é uma área de conhecimento da Inteligência Artificial (IA) que emula uma das formas de raciocínio humano. É uma metodologia de resolução de problemas que, em alguns aspectos, diferencia-se de outras técnicas de IA. Ao invés de conter somente um conhecimento geral do domínio do problema ou fazer associações de relacionamento generalizados entre descrição e conclusões de um problema, a técnica de RBC utiliza o conhecimento específico de uma experiência passada, um problema concreto, para resolver um problema atual. Um novo problema é resolvido através da busca por um caso similar passado e a solução poderá ser adaptada para este novo problema. Outra diferença entre RBC e outras técnicas de IA, está na resolução de um problema que inclui processo de armazenamento do mesmo numa base de casos para posteriormente ser utilizado na solução de futuros problemas (Watson, 2000).

A técnica de RBC teve origem em uma pesquisa na área da ciência cognitiva por Roger Schank e Abelson em 1977. Neste trabalho, os autores afirmam que o conhecimento é armazenado através de *scripts*. Estes *scripts* descrevem informações sobre situações que ocorrem com seres humanos, por exemplo, ir a um museu, jantar, almoçar. Porém, experimentos com fatos mostraram que somente os *scripts* não representavam todo o conhecimento armazenado na memória. Era comum pessoas misturarem fatos parecidos, como consulta médica e consulta dentária, no momento da recuperação da informação (Shank, 1982).

Em 1980, Schank fez um trabalho sobre memória dinâmica, de como a lembrança é utilizada para situações padrões de resolução e aprendizagem de problemas. A memória dinâmica usa uma estrutura hierárquica denominada de Pacotes de Organização de Memória (MOP's), que agrupam um conjunto de casos com características similares. Nesta estrutura, os casos são caracterizados pelos episódios aos quais estão associados e seus atributos não são apenas nomes próprios, mas atributos das abstrações que juntas modelam um contexto do caso (Aamdot, 1994).

As experiências passadas são as entidades denominadas “casos”. Segundo Kolodner (1993), um caso é a abstração de uma experiência, que deve estar descrita em termos de seu conteúdo e contexto. O caso pode assumir diferentes formas de representação. O exemplo mais simples de um caso é uma experiência descrita por meio de atributos valorados.

3.2 O ciclo de RBC

Segundo Aamodt e Plaza (1994), o ciclo proposto para o processo de RBC, mostrado na Figura 3.1, é composto por quatro tarefas principais: recuperar, reutilizar, revisar e reter.

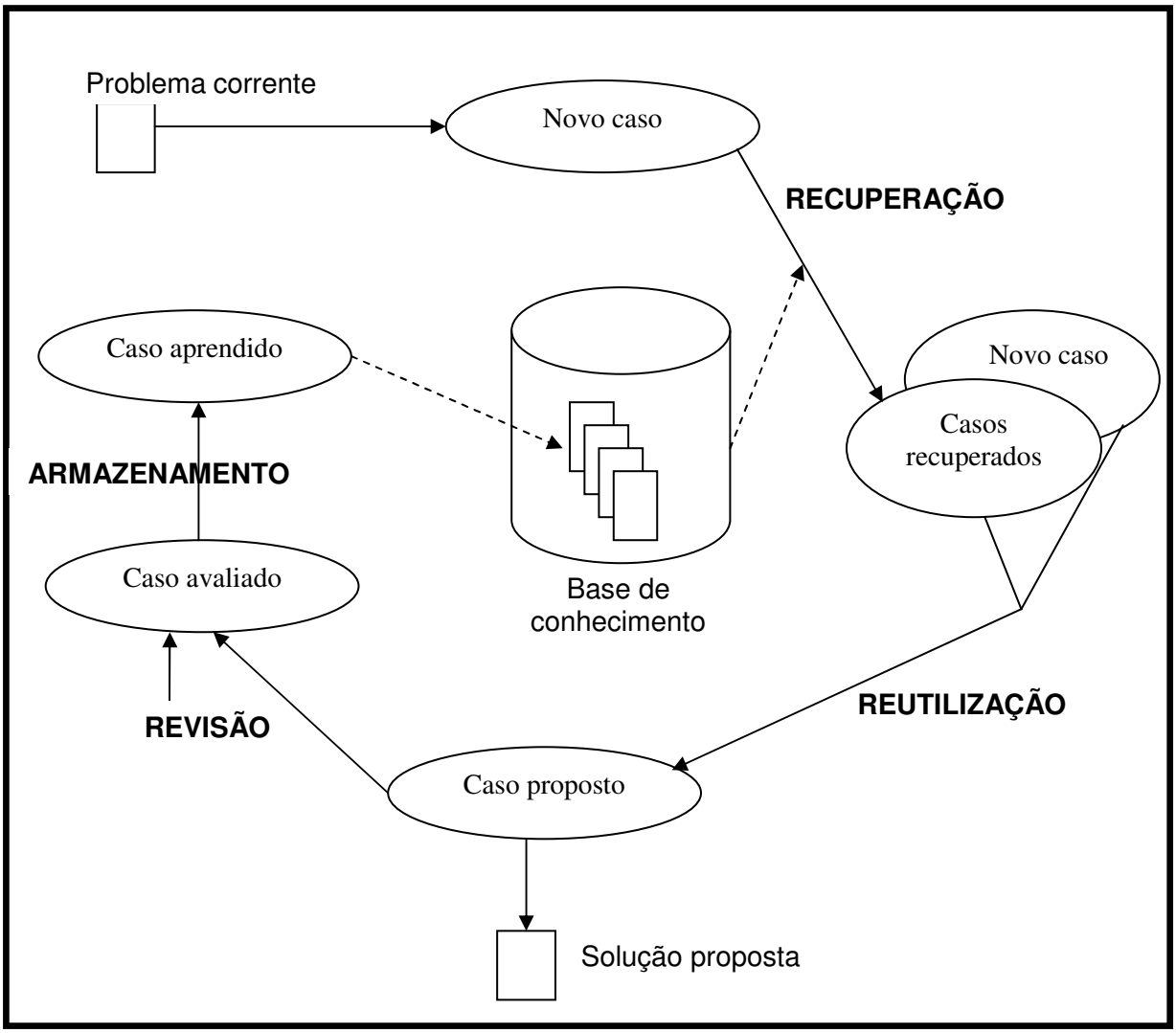


Figura 3.1 - Ciclo do RBC

No centro do ciclo está o conjunto de casos representando as experiências armazenadas na base de casos e um conhecimento geral do domínio. Este conhecimento geral do domínio é aplicado em diferentes passos do processo de RBC e provém, por exemplo, controle do conhecimento para consistência dos casos, a busca por casos similares ou adaptação destes casos para resolução de novos problemas.

Na recuperação, o caso ou um conjunto de casos mais similares da base de casos será determinado, baseando-se na descrição do novo problema. A reutilização da informação e o conhecimento dos casos recuperados serão usados para resolver o novo problema. Durante a revisão será avaliada a aplicabilidade da solução proposta para o problema atual e, se necessário, o caso proposto será adaptado para complementar à resolução da presente situação. Se a solução do caso proposto, durante a fase de revisão, resolveu o problema apresentado, então este caso será retido na base de casos e conseqüentemente será considerado um novo caso.

A idéia fundamental deste ciclo de quatro etapas é, hierarquicamente, distribuir as tarefas e utilizar como um ciclo de raciocínio contínuo (Watson, 2000).

3.2.1 Recuperação

O objetivo desta etapa é recuperar o caso ou casos da base de casos que contenham uma solução mais próxima para um problema atual. A recuperação é feita usando as características do novo caso que são relevantes na solução de um problema. Segundo Aamodt e Plaza (1997), a tarefa de recuperação de casos inicia com a descrição de um problema e termina quando um caso mais similar é encontrado. A Figura 3.2 apresenta o esquema do processo de recuperação em RBC.

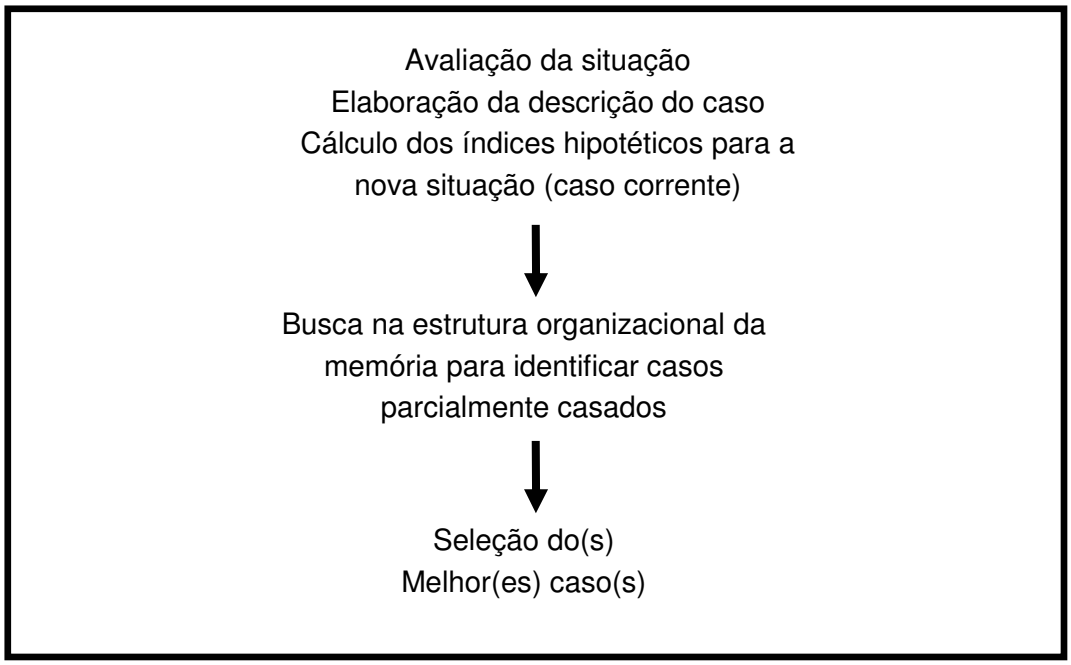


Figura 3.2 - Esquema do processo de recuperação na base de casos

É difícil determinar se um caso é útil para a solução de um problema ou uma situação específica. O que faz um caso ser similar a outro difere de acordo com o domínio e o propósito da aplicação, porém, num contexto geral, pode-se dizer que a semelhança entre casos está na similaridade das características que representam o conteúdo e o contexto das experiências em questão. Um caso pode ser considerado similar ao problema atual, se a solução do caso puder ser reutilizada para resolver o problema atual. Em RBC assume-se que problemas ou situações similares requerem soluções similares.

A questão da similaridade é bastante singular e precisa ser estudada para cada domínio. A similaridade entre casos pode ser determinada por métricas (resultado de uma função), por metas, restrições ou classificações, sendo que nos três últimos casos, uma métrica pode ser utilizada para estabelecer o grau de similaridade dos casos (Peres, 1999). Alguns autores dividem a avaliação da similaridade em dois grupos:

- **similaridade sintática:** comparação sintática dos valores dos atributos, analisando sinônimos, categorias ordinais, intervalos, etc;
- **similaridade semântica:** abrange o significado dos casos relacionando-os aos valores dos atributos, envolvendo o processamento de linguagem natural.

Tanto a similaridade sintática quanto a semântica, quando aplicadas em problemas, retornam valores que medem a similaridade entre os casos. Esta medida de similaridade é

encontrada através de funções e/ou regras que guardam algum conhecimento sobre o domínio do problema, já que para julgar se determinadas situações são parecidas é necessário saber o que exatamente as tornam semelhantes.

Para avaliação de similaridade há necessidade de determinação dos índices que irão orientar quais atributos serão relevantes. Para determinação dos índices há necessidade de conhecimento especializado, isto é, o conhecimento de um especialista sobre o problema abordado. A indexação é a essência do RBC porque orienta a avaliação de similaridade. Os índices determinam o que deve ser comparado entre os casos, no intuito de recuperar os casos mais úteis para resolver o novo caso. Kolodner (1993) apresenta o problema de indexação através de duas sub-tarefas: (a) a definição do vocabulário de indexação; e (b) a atribuição dos índices. A primeira consiste em selecionar dimensões que, quando atribuídas, preencham as funções dos índices e a segunda refere-se à atribuição de valores para estas dimensões.

Há funções numéricas para determinar a similaridade. Uma freqüentemente utilizada é a do "vizinho mais próximo". Segundo Watson (1997), a técnica do vizinho mais próximo é, talvez, a mais utilizada para o estabelecimento da similaridade. Os aspectos de definição e identificação dos índices são fatores fundamentais para uma recuperação de sucesso. A similaridade entre o caso alvo e um caso na base de casos é determinada para cada atributo. Esta medida deve ser multiplicada por um fator peso. A somatória de todos os atributos é calculada e permite estabelecer a medida de similaridade entre os casos da base de casos e o caso alvo. A fórmula para o cálculo da técnica do vizinho mais próximo é a seguinte:

$$Similaridade(T,S) = \sum_{i=1}^{n} f(Ti,Si)wi$$

onde:

T é o caso alvo

S é o caso fonte

n é o número de atributos em cada caso

i é cada atributo individual variando de 1 a n

f é a função de similaridade para o atributo *i* no caso *T* e *S*

w é o peso relativo ao atributo *i*

Este cálculo é utilizado para cada um dos casos da biblioteca de casos, obtendo o *ranking* dos mesmos. As similaridades são usualmente normalizadas para um intervalo entre zero e um (zero quando tiver nenhuma similaridade e um quando a similaridade for exata). A grande dificuldade é determinar o valor dos pesos relativos às características. Em geral o tempo de recuperação aumenta linearmente com o número de casos.

Depois de encontrados os casos similares, é necessário estabelecer uma restrição para recuperação. Um limiar de similaridade (recuperação apenas de casos cuja medida de similaridade ultrapassa esse limiar) ou um número limite de casos pode ser estabelecido.

3.2.2 Reutilização

Para Aamodt e Plaza (1994) a reutilização da solução do caso recuperado em relação ao novo caso foca dois aspectos:

- as diferenças entre o caso passado e o caso atual;
- qual parte do caso recuperado pode ser transferida para o novo caso.

A simples tarefa de classificar as diferenças é abstrata, pois é considerada uma tarefa não relevante, enquanto que a similaridade é considerada relevante e a solução do caso recuperado é transferida para o novo caso como sendo a solução proposta. Este é o tipo trivial de reutilização, contudo há outros sistemas que implementam a reutilização adaptando a solução do caso recuperado para o novo caso.

a) Adaptação

Há situações em RBC em que os casos recuperados podem apresentar uma solução aproximada para o caso atual, exigindo algumas modificações para melhor ajustá-la na resolução deste caso. Estas modificações são chamadas de *adaptação* e podem ser feitas por meio da utilização de conceitos específicos da técnica de RBC, por meio de regras que representam um conhecimento adicional sobre o domínio do problema ou até mesmo por meio de interações com o usuário.

Segundo Watson (1997), a adaptação pode ser feita de várias formas, havendo a possibilidade de ocorrer, em determinadas situações, uso combinado das seguintes tarefas:

- inclusão de um novo comportamento à solução esperada;

- eliminação de um comportamento da solução recuperada;
- substituição de parte de um comportamento.

Kolodner (1993) descreve dois tipos de métodos para proceder a uma adaptação, classificando estes em *adaptação derivativa* e *adaptação transformativa*. A adaptação derivativa é realizada sobre o método de solução apresentado no caso escolhido. Este processo de adaptação aplica o método descrito no caso escolhido, adaptando-o para aplicá-lo no problema de entrada.

A adaptação transformativa é aplicada nos sistemas que transformam o caso recuperado, de modo a solucionar o problema de entrada, por meio de heurísticas ou modelos. Os chamados operadores de transformação são construídos em função das diferenças entre o problema de entrada e o caso escolhido. Na adaptação transformativa há métodos de substituição, descritos a seguir.

- **reinstanciação:** aplica-se este método atribuindo novos valores aos atributos que descrevem o caso. Este método aplica-se quando as funções (no problema escolhido) passam a ser exercidas diretamente no problema de entrada;
- **ajuste de proporções:** é implementado através de interpolação nos valores do caso escolhido, em função de uma variação de intensidade, volume ou número com relação ao problema de entrada;
- **busca local:** é definida pela busca numa estrutura hierárquica, por um substitutivo para algum atributo que esteja impedindo a satisfação de uma restrição. Esta busca é feita somente nas redondezas do caso e pode ser utilizada quando a busca local não fornece a substituição adequada;
- **busca especializada:** é orientada para resolver uma questão que gera discrepância entre os casos (a similaridade não é absoluta se a discrepância não for neutralizada). Ela transcende a busca na memória, seu objetivo é procurar o valor que neutralize a discrepância entre os casos;
- **substituição baseada em casos:** este método executa uma busca na memória que recupera um outro caso similar e, então, provê o valor desejado para a adaptação.

Apesar da adaptação poder ser usada de várias formas e em várias situações, ela não é essencial e, muitos sistemas comerciais que utilizam RBC não a implementam. Eles simplesmente recuperam o caso mais similar e disponibilizam a solução para o usuário, deixando-o livre para proceder à adaptação (Watson, 1997).

3.2.3 Revisão

Quando a solução gerada pela fase de reutilização não está correta, há a oportunidade para aprender com as falhas encontradas. Esta fase é chamada de revisão do caso e consiste em duas tarefas: (a) avaliação da solução gerada para reuso; e (b) reparar a solução usando o conhecimento específico do domínio (Aamodt e Plaza, 1997). A Figura 3.3 mostra o processo de revisão.

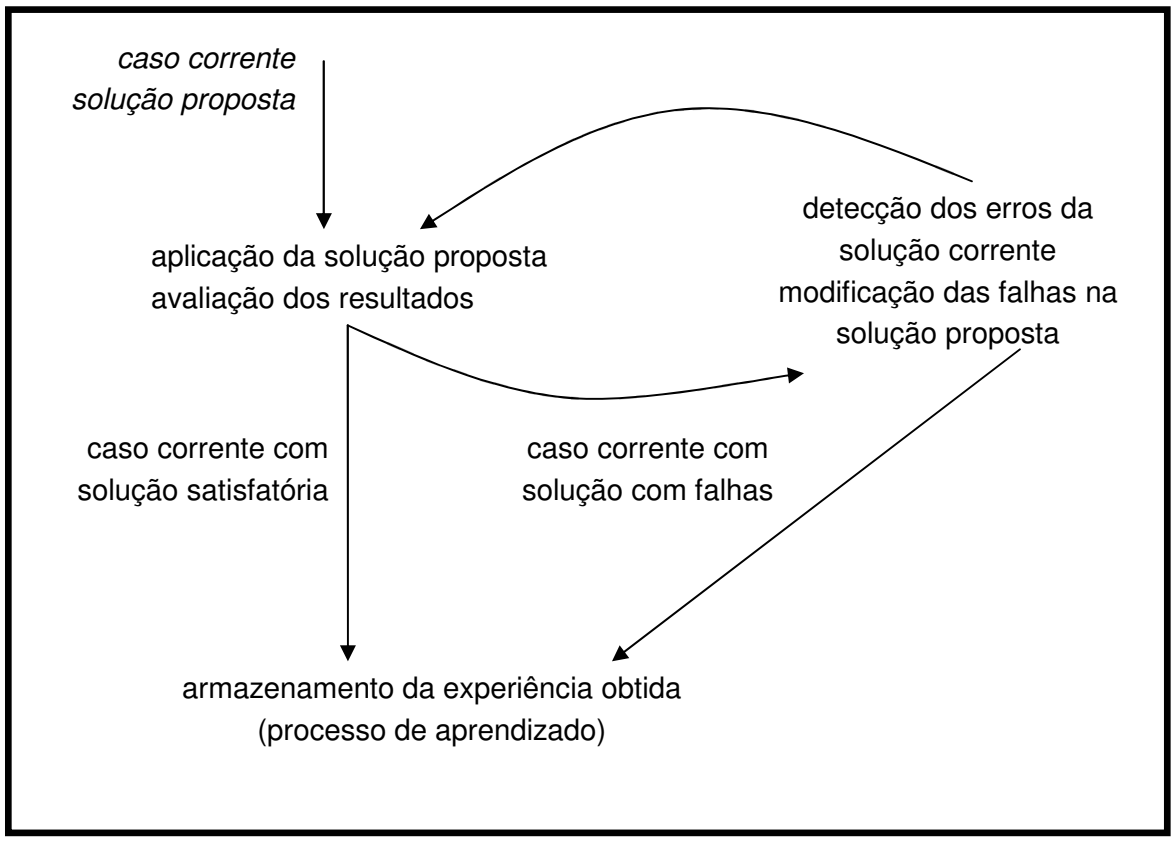


Figura 3.3 - Esquema do processo de revisão

a) Avaliação da solução

A tarefa de avaliação considera o resultado da aplicação da solução em um ambiente real, através do questionamento a um especialista ou através da aplicação de regras que validem a solução. O resultado da aplicação da solução pode levar algum tempo para aparecer, dependendo do tipo de aplicação. Em um sistema de suporte à decisão médica, por exemplo, o sucesso ou falha do tratamento pode levar algumas horas ou até vários meses. O caso pode ainda ser instruído e validado na base de casos por um período intermediário, mas ele deverá ser marcado como um caso validado temporariamente. A solução pode também ser aplicada a um programa de simulação que está habilitado a gerar uma solução correta.

b) Reparação de falha

Reparar um caso envolve encontrar os erros da solução proposta e apresentar explicações para atualização desta solução. Alguns sistemas utilizam um conhecimento casual para gerar uma explicação do porquê certas partes da solução não foram alcançadas. Estes sistemas aprendem as situações gerais que causarão as falhas usando uma técnica de aprendizagem baseada em explicação. Isto é incluído em uma memória de falhas que são usadas para fase de reutilização, fornecendo atalhos para a adaptação. Isto oferece a grande vantagem que é a possibilidade de detectar possíveis erros na fase de adaptação (Aamodt e Plaza, 1997).

Descobrindo as falhas da solução proposta, há necessidade de aplicar a reparação da falha. Esta tarefa usa uma falha explicada para modificar a solução de tal forma que aquela falha não ocorra. Para modificar a falha, um módulo de reparação, que possui conhecimento do domínio, assegura que a causa dos erros não irá ocorrer. Após este processo a solução poderá ser retida, caso haja efetiva certeza da eficiência da solução ou poderá passar novamente pelas etapas de validação e reparação (Aamodt e Plaza, 1997).

3.2.4 Retenção

Depois de realizado o processo de revisão, a solução do caso selecionado pode então ser utilizada para resolver o caso de entrada. Um sistema RBC somente se tornará eficiente quando estiver preparado para aprender a partir das experiências passadas e da correta indexação dos problemas (Kolodner, 1993).

A retenção de casos significa incorporar à base de casos informações úteis relativas à resolução de um novo problema. Este processo corresponde à aprendizagem de um sistema RBC, que é disparado pelas tarefas de avaliação e adaptação de soluções (Aamodt e Plaza, 1994).

O aprendizado de um sistema deve ocorrer de forma ordenada para não tornar a base de casos algo difícil de ser manipulado. A inclusão de novos casos e associação de índices deve ocorrer de forma que o sistema possa raciocinar sobre eles. Melhorias no cálculo do grau de similaridade assim como nas regras de adaptação, também podem ajudar no melhoramento da performance do sistema. A base de casos tende a crescer com o passar do tempo, portanto pode ser necessária a inclusão de um processo de esquecimento para controlar esse crescimento. Este processo pode ser guiado por meio de funções, regras e heurísticas que dependem do domínio de atuação do sistema (Peres, 1999).

A aprendizagem em sistemas de RBC pode ser empregada ao nível dos casos e da base de casos. As bases de casos podem ser estendidas por processos incrementais de aprendizagem se a tarefa e o projeto do sistema permitirem. A partir de um pequeno conjunto de casos, a base de casos pode crescer com novos casos (Leake, 1996).

A aprendizagem no nível dos casos acontece como expressão da aprendizagem com a experiência. A parte do caso destinada ao resultado do emprego de determinada solução ou interpretação serve a este propósito. Mantém-se no caso o registro de seu desempenho ao ser utilizado. Assim, tanto sucessos como fracassos são informados incrementando o conhecimento e as lições embutidas no caso. O registro do resultado de reutilização pode prevenir o usuário em relação às possíveis conseqüências de seu uso. Este procedimento é valioso porque, para compensar a inclusão de informações no caso, o sistema evita a reutilização de sugestões menos favoráveis, resultando no incremento da qualidade da recuperação (Webber-Lee, 1998).

3.3 Indexação

A indexação é a essência do RBC porque orienta a avaliação de similaridade (Kolodner, 1993). A indexação determina o que comparar entre os casos para determinar a similaridade, usando-os com o objetivo de facilidade e rapidez na recuperação (Watson, 1997).

Kolodner (1993) apresenta o problema de indexação através de duas sub-tarefas: (a) a definição do vocabulário de indexação; e (b) a atribuição dos índices. A identificação do vocabulário de indexação consiste em selecionar dimensões que, quando atribuídas, preenchem as funções dos índices. A atribuição dos índices refere-se à atribuição de valores para estas dimensões. Segundo Watson (1997) os índices devem:

- ser preditivos, isto é, devem prever a utilização da informação presente nos casos para diferentes situações de problema;
- endereçar os propósitos onde o caso pode ser usado;
- serem abstratos suficientes para permitir que um caso seja útil em uma variedade de diferentes situações;
- serem concretos suficientes para que possam ser facilmente reconhecidos em situações futuras.

Os índices podem ser escolhidos através de métodos manuais, onde a escolha começa com a análise dos casos para a identificação da utilidade que poderia ter o caso, e sob que circunstâncias. Essas informações devem ser, então, traduzidas para representações que o sistema pode usar, definindo um conjunto de descritores, que são trabalhados de modo a garantir que os índices sejam aplicáveis em âmbito geral e que possam ser reconhecidos no máximo de situações possíveis (Melchiors, 1999).

Uma difícil tarefa na definição dos índices é prever que tipos de situação de consultas irão surgir e que tipos de informação serão necessárias para recuperar casos em situações futuras. Muitos esforços foram feitos para estabelecer regras gerais de vocabulário de índices em classes particulares nas tarefas de RBC, mas esta tarefa acaba ainda sendo desenvolvida para atender os objetivos específicos da recuperação de cada aplicativo que use RBC (Lagemann, 1998).

O processo de indexação é uma oportunidade de superar a deficiência de experiências mal descritas e torná-las úteis e valiosas na realização da tarefa do sistema. Esta meta é conduzida pela correta interpretação da experiência a partir da perspectiva do especialista, permitindo a identificação do significado intrínseco e da correlação entre as entidades ativas participantes na experiência. Uma forma de buscar tais relações é tentar representar as correspondências entre as causas e conseqüências, razões e soluções (Webber-Lee, 1998).

A indexação pode representar um gargalo no desenvolvimento de sistemas de RBC, como é o exemplo da necessidade de indexação automática para viabilizar o sistema. A indexação automática pode ser necessária em sistemas que comportam mecanismos de aprendizagem automática. Além disso, há domínios em que o conhecimento está disponível somente em formato textual o que exige grandes bases de casos. Os domínios do Direito, Economia e Medicina são exemplos nos quais um mecanismo de indexação automática é plenamente justificado (Webber-Lee, 1998).

3.4 RBC em recuperação de informação

Segundo Korfhage (1997), qualquer sistema que faça algo mais que recuperar informação não está classificado como um sistema de recuperação de informação. Para formular um sistema aplicando a técnica de RBC há necessidade de conhecimento prévio sobre o domínio do problema, isto é, o sistema de RBC é aplicado sobre um domínio específico.

Como discutido no capítulo 2 a recuperação de informação é a tarefa de encontrar documentos relevantes sobre um conjunto de documentos, em resposta a uma necessidade de informação de um usuário. As soluções proporcionadas pela pesquisa em um conjunto de documentos, trazidas pela comunidade de recuperação de informação, guardam afinidade com RBC. O ponto fundamental desta afinidade trata do interesse da comunidade de RBC na tarefa de recuperar informação, isto é, RBC efetiva essencialmente a recuperação da informação a partir de uma consulta (Smeaton, 2001).

No contexto de RBC a definição de um vocabulário de índices e a construção da medida de similaridade, em particular, são cruciais, considerando que há previamente uma coleção de documentos em uma base de dados.

Os sistemas que utilizam RBC apresentam a necessidade de conhecimento específico sobre o domínio do problema, onde características importantes do domínio devem ser observadas, como palavras-chaves, termos e expressões. Sendo assim, pode haver a necessidade de construir um *thesaurus*, ampliando a informação para a busca.

Outra característica definida em RBC é a construção da função que indicará a similaridade entre a informação de busca e a informação na base de casos. Este conhecimento sobre o domínio do problema oferece a possibilidade de construir uma

medida de similaridade mais eficiente, como, por exemplo, definir pesos para palavras-chaves, expressões e termos, influenciando diretamente na recuperação de documentos (Lenz, 2001).

Outra característica importante em RBC é a possibilidade de adaptação. Esta característica, particularmente, evidencia mais uma vez a abrangência maior do que apenas recuperar informação. A utilização do conceito de adaptação em sistemas possibilita que as soluções de situações passadas de recuperação de informação possam ser adaptadas para auxiliar a recuperação de informações atual.

Segundo Ramirez (2001), devido às características de RBC, as áreas que possuem maior tendência em utilizá-las tem por característica:

- dificuldade em formalização;
- necessidade de análise e planejamento;
- importância de convencimento;
- necessidade de cenários hipotéticos.

3.5 RBC em aplicações na Internet

A Internet tornou-se rapidamente o principal meio de troca de informação, tanto pelo volume de informações disponível como pelo constante crescimento do mesmo. Esta vasta quantidade de informação gerou a necessidade de mecanismos eficientes para encontrar o que se procura. É neste contexto que se insere a aplicação de RBC, técnica capaz de incluir conhecimento de domínio nas consultas realizadas.

A Internet como agente facilitador funciona como instrumento para diversas aplicações, por exemplo, busca por informações, comunicação e comércio eletrônico. Em muitas dessas áreas RBC é perfeitamente aplicável, sendo que e em algumas, além de melhorar o resultado final, diminuirá a quantidade de etapas necessárias para se obter um determinado serviço ou a desejada informação.

Há várias arquiteturas propostas utilizando RBC em aplicações na Internet. Marinilli *et al*. (2001), por exemplo, propõem uma arquitetura de sistema utilizando RBC para filtragem de informação na Internet. O sistema propõe a capacidade de selecionar documentos coletados na Internet, de acordo com o interesse e características de cada usuário. De modo geral o usuário deve formar um perfil, indicando as áreas de interesse e

submeter uma *query* em forma de texto. O sistema faz uma busca utilizando algum serviço de busca já disponível, (*Altavista* - www.altavista.com por exemplo) e sobre os documentos recuperados e o perfil do usuário é aplicado um cálculo que indica o grau de similaridade do documento com os parâmetros informados pelo usuário. O resultado apresenta a relação de documentos, que possuem similaridade igual ou superior à indicada pelo usuário, ordenados pelo maior grau de similaridade. Em testes comparativos entre a utilização do sistema e somente a utilização do site de busca *Altavista*, houve uma melhora superior a 30%, segundo os autores.

Outra área que está utilizando a técnica de RBC é o comércio eletrônico. Vollrath *et al*. (2001) mostram a vantagem de utilizar RBC nesta área por meio de uma aplicação em uma empresa que vende mais de 130 tipos de diferentes produtos. Os vendedores preenchem parâmetros que estão divididos em categorias indicando o grau de importância. Estes parâmetros geram uma *query* que é pré-processada em um servidor por meio de uma busca simples pelos parâmetros, retornando a lista dos produtos que tenham alguma relação com estes parâmetros. O resultado desta *query* é enviado ao servidor que contém o módulo de RBC, e então é feita a aplicação dos mecanismos da técnica de RBC sobre este pré-resultado. Como resultado são apresentados os dez produtos com maior grau de similaridade. Caso os produtos recuperados não satisfaçam o vendedor, há possibilidade de alterar o grau de importância de cada produto e submeter uma nova consulta.

Watson (2001) descreve um sistema utilizando RBC para auxiliar vendedores de sistemas de aquecimento, ventilação e ar condicionado para engenheiros da construção civil. Antes da utilização do sistema com RBC, quando os engenheiros tinham dúvidas sobre algum dos produtos, tinham que entrar em contato com um vendedor que teria que esclarecer qual tipo de produto seria mais indicado, isto é, o vendedor deveria saber, por meio das características da construção, informadas pelo engenheiro, qual produto se adaptaria melhor ao ambiente da construção. Para facilitar a decisão do vendedor, a empresa, que já possuía uma base de dados contendo mais de 10.000 registros, e considerando que cada registro contém sessenta atributos e representa uma construção, desenvolveu um sistema utilizando RBC para informar que tipo de produto seria mais indicado para um determinado ambiente em uma construção. O sistema utiliza um *Applet*[1]

[1] Applet: programa feito na linguagem Java, para os usuários fazerem download de bytecodes pela Internet, através de um navegador Web (Netscape Navigator, Internet Explorer) e executar em suas próprias máquinas (Horstmann e Cornell, 2001).

que é utilizado pelo vendedor que serve para descrição das características da construção a ser instalado o produto. Inseridas as características, o *Applet* gera um arquivo XML contendo as informações da *query* de busca. No servidor há um módulo de RBC que irá recuperar as construções cujas características sejam mais similares. Feito este processo, é gerado um novo arquivo XML contendo as informações das características das construções mais similares e, então, é enviado ao *Applet* do vendedor, servindo de auxílio para este indicar qual produto é mais indicado para o ambiente da construção.

A HP – Hewlett-Packard (2001) possui um site para auxílio aos usuários, um *help desk* sobre problemas com impressoras. Ao entrar no site o usuário se depara com uma questão e várias opções de resposta para o que possa estar acontecendo com sua impressora. A partir dessa resposta outras questões serão requeridas, sendo que as novas perguntas são formalizadas em função da resposta da pergunta anterior, uma vez que cada pergunta possui um conjunto de alternativas como resposta. As perguntas podem ser sobre problemas que estão ocorrendo ou pedindo a confirmação de procedimentos, instruções, que poderiam ser feitos pelo usuário, por exemplo, o cabo da impressora foi conectado ao computador. Cada resposta é armazenada no servidor, formando um perfil, para então aplicar a técnica de RBC. Depois de algumas perguntas respondidas, o servidor aplica a técnica de RBC comparando este perfil de problema sobre um banco de dados que contém problemas semelhantes juntamente com sua solução. Ao final do processo o usuário receberá um indicativo ou indicativos dos problemas que poderão estar acontecendo com sua impressora, juntamente com as possíveis soluções, ordenadas por grau similaridade.

A utilização da técnica de RBC para este tipo de sistema da HP simplifica muito a aquisição de soluções para o cliente, fornecendo de uma maneira rápida e dinâmica, condições para utilização por qualquer usuário. Além do usuário, que utiliza este site, os funcionários que dão algum tipo de suporte podem ser auxiliados e quando houver novos funcionários não há necessidade de treinamento intensivo, pois à medida que o funcionário for utilizando o sistema, conhecerá os produtos da empresa e pode fornecer suporte adequado.

Há outros sistemas em outras áreas que poderão ser utilizados com a técnica de RBC na Internet. Uma característica que pode ser notada, mais uma vez, é que cada sistema deverá possuir um domínio específico, onde o procedimento do cálculo de

similaridade é singular, pois deverá ser abordado utilizando características particulares de cada domínio, para maximizar a sua utilização.

3.6 Considerações finais

Este capítulo abordou a técnica de Raciocínio Baseado em Casos. O primeiro tópico apresentou um breve histórico da criação de RBC, descrevendo seus principais fundamentos conceituais. A composição da arquitetura de RBC constitui várias fases, recuperação, reutilização, revisão e retenção. Outro tópico abordado refere-se à utilização de índices para o processo de recuperação, sendo esta, uma etapa de especial relevância em tarefas de recuperação de informação.

Um dos tópicos abordados neste capítulo é a utilização da técnica de RBC em recuperação de informação. Uma das etapas de RBC é a recuperação e esta, juntamente com as outras etapas, pertencentes a sua arquitetura, fornece subsídios suficientes para formalizar um sistema de recuperação de informação inteligente. Abordou-se também o emprego da técnica de RBC em sistemas na Internet, fornecendo mais subsídios para obtenção de informação. Descreveram-se exemplos em diferentes áreas de sistemas já em funcionamento e discutiram-se as melhoras significativas nas respostas.

O método de construção de perfis de busca em sites na Internet, proposto nessa dissertação, tem na técnica RBC o elemento tecnológico central. Cada busca ou interação de um usuário na Internet é vista como um caso da técnica RBC. As novas buscas são consideradas casos de entrada e o acionamento do RBC é automático e transparente para o usuário. Assim, como se verá no próximo capítulo, RBC no modelo proposto do presente trabalho ocupa o papel de tecnologia que viabiliza a memorização, indexação e recuperação de consultas a um site, de forma a facilitar o processo de buscas de futuros usuários do site. A aplicação de RBC e, sobretudo, o modelo conceitual proposto e transformado em ferramenta aplicada a sites conhecidos, estão descritos no próximo capítulo.

4 A Ferramenta RBNet

4.1 Introdução

A Internet nos proporciona o acesso a grande quantidade e diversidade de informação, despertando interesse de vários usuários na tentativa de suprir a sua necessidade de informação. Na utilização de sites de busca (em especial aqueles que disponibilizam grandes volumes de informação) é comum que usuários na procura de uma determinada informação, levem muito tempo para conseguí-la ou mesmo não a consigam obtê-la (embora ela se encontre no site).

As informações de cada consulta podem ser utilizadas para auxiliar outros usuários em consultas similares. Quando um usuário faz uma consulta e recebe a listagem contendo os sites relacionados à expressão de busca (palavra ou texto), aqueles que julgar interessantes são visitados. Estas informações, expressão da busca e a lista dos sites visitados, por exemplo, podem ser consideradas como informações do perfil de consulta do usuário. Este perfil é composto por todas as informações de interação do usuário com o site de busca, portanto, é formado dinamicamente, pois, por exemplo, um site visitado faz parte do perfil da consulta.

As informações pertencentes ao perfil de uma consulta podem ser armazenadas e servir como instrumento facilitador de recuperação de informação. Assim, o usuário além de receber a listagem dos sites relativos à sua consulta, pode receber a listagem de consultas cujas informações do perfil possuem alguma semelhança com as informações do perfil de consulta do usuário atual. A relação de consultas similares é ordenada pelo grau de similaridade em ordem decrescente.

Com as informações dos perfis de consulta dos usuários armazenadas, caso o usuário queira verificar as informações relativas a uma determinada consulta similar recuperada, há diminuição da carga de trabalho do computador servidor responsável por varrer a base de dados recuperando as informações da consulta, o que torna a busca ainda mais rápida, uma vez que cada perfil de consulta possui, entre outras informações, a relação dos sites que recupera e a relação somente daqueles sites visitados.

O modelo concebido, prototipado e testado tem este objetivo. Por meio da utilização de RBC, armazenar as informações dos perfis de consultas na forma de casos e recuperar essas informações por meio de uma função que calcula o grau de similaridade, para auxiliar os demais usuários em suas buscas. À medida que o usuário interage com os sites recuperados, o perfil de sua consulta incorpora mais informações, como já foi mencionado. A cada alteração das informações do perfil de consulta atual, uma nova recuperação de consultas é disparada, apresentando um refinamento automático na relação de consultas similares recuperadas, conseqüentemente, melhor qualificação dos resultados apresentados.

4.2 Arquitetura da ferramenta RBNet

A ferramenta RBNet tem como principal objetivo o aproveitamento das interações de usuários em sites de disponibilização de informações, por meio de uma busca, como subsídios a novos usuários que utilizam os mesmos recursos em novas buscas. Para a ferramenta atender a essa definição, foi necessário conceber uma arquitetura que é composta pelos seguintes módulos: perfil das consultas, módulo de entrada, módulo de saída, cálculo do grau de similaridade e a base de casos. A Figura 4.1 apresenta uma visão geral do esquema de recuperação de informação com RBNet, detalhado mais a frente.

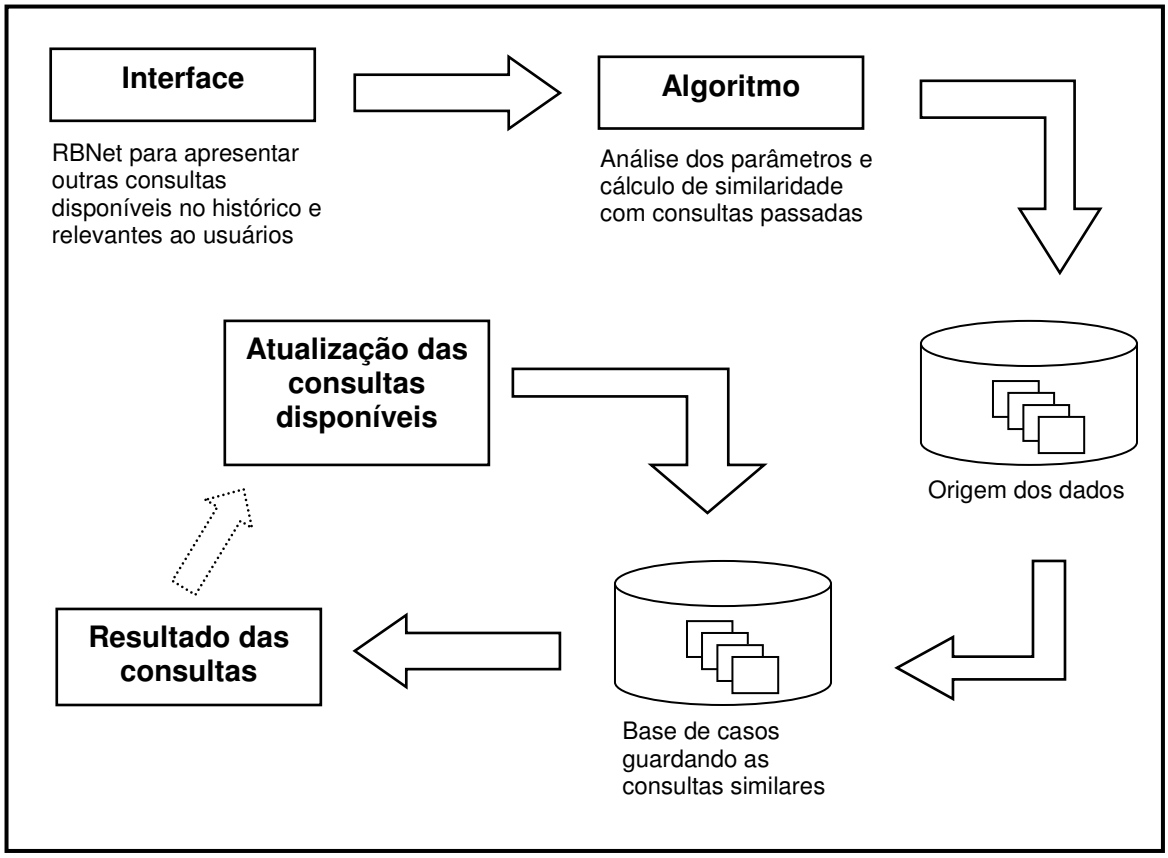


Figura 4.1 - Esquema de recuperação de informação com RBNet

4.2.1 Perfil de consulta

O perfil de consulta é composto por todas as informações que caracterizam uma consulta e mais as informações de administração desta no servidor. As informações destes perfis estão armazenadas na base de casos. Cada perfil de consulta representa um caso na base de casos.

A composição do perfil de uma consulta inclui os seguintes elementos:

- **identificador**: o identificador é único e representa uma consulta na base de casos;
- **expressão da consulta**: representa as palavras informadas para aplicar na busca sobre a base de dados e base de casos;
- **parâmetros**: a quantidade de parâmetros é variável para cada sistema, podendo ou não existir. Para definir estes parâmetros há necessidade do conhecimento de um especialista no domínio da aplicação;

- ***query* da consulta**: a *query* da consulta é armazenada para prever futuras atualizações dos registros recuperados pelas consultas da base de dados, fazendo parte da administração da base de casos. Em resumo, objetiva manter atualizada a relação de registros recuperados pelas consultas;
- **registros recuperados**: é a relação dos registros recuperados por cada consulta;
- **registros visitados**: é a relação dos registros visitados em cada uma das consultas.

4.2.2 Módulo de entrada

Quando o usuário acessa um site de busca há necessidade de informar o que se deseja buscar. Este procedimento, como já foi mencionado, exige a digitação de uma palavra ou frase (expressão de busca) e, opcionalmente, para sites que oferecem parâmetros de busca, a escolha ou não de opções desses parâmetros. A ferramenta proposta pode ser aplicada em qualquer site de busca, desde que haja o conhecimento de um especialista no domínio da aplicação, pois este conhecimento será necessário para a definição da modelagem do RBNet. A disponibilização de parâmetros para busca, além da expressão de busca, é importante para melhorar a definição da busca e ampliar a capacidade de ação da técnica de RBC sobre a base de casos, oferecendo condições para recuperação mais efetiva de perfis de consultas similares.

O módulo de entrada que compõe o perfil de consulta, é composto por:

- **expressão de busca**: pertence ao site de busca e representa a informação explícita a ser buscada, isto é, são as palavras ou frase digitada no campo texto (por exemplo, "*inteligência artificial*").
- **parâmetros**: pertence ao site de busca. São os parâmetros utilizados para melhor representação da consulta (por exemplo, recuperar sites somente em português).
- **grau mínimo de similaridade**: pertence à interface RBNet. É o valor percentual que indica o grau mínimo de similaridade para recuperar uma consulta, que pode variar de 0% até 100%. Por exemplo, se o grau de

similaridade é igual a 70%, são recuperadas somente as consultas com grau de similaridade igual ou superior a este percentual.

- **registros visitados**: cada registro visitado possui um identificador único, que é armazenado na base de casos, fazendo parte do perfil de consulta.

4.2.3 Módulo de saída

O módulo de saída representa as informações que são recuperadas e apresentadas ao usuário. Estas informações resultam da recuperação de registros armazenados na base de dados (ou em páginas de site) e a recuperação das consultas similares armazenadas na base de casos. O módulo de saída está diretamente relacionado com o módulo de entrada, pois as informações resultantes dependem das informações constantes no módulo de entrada.

As informações contidas no módulo de saída são:

- **consultas similares**: pertence à interface RBNet. É o local onde são apresentadas as informações das consultas que possuem o grau mínimo de similaridade indicado no módulo de entrada.
- **registros recuperados**: pertence ao site de resultados. São os registros recuperados da base de dados, diretamente associados às informações do módulo de entrada.
- **detalhamento da consulta similar**: são as informações que identificam uma consulta similar recuperada. As informações apresentadas, por exemplo, expressão de busca, total de registros visitados e as opções dos parâmetros utilizados, podem ter *links* para informações mais detalhadas da consulta, como, por exemplo, a relação dos registros que foram visitados.
- **detalhamento do documento**: os registros recuperados possuem *link* indicando o caminho para visualização completa do documento ou página equivalente.

4.2.4 Cálculo do grau de similaridade

No modelo proposto há, em dois momentos distintos, a utilização do cálculo do grau de similaridade, portanto, tem-se a necessidade de definir duas fórmulas diferentes.

Estas fórmulas, que possuem semelhança, foram baseadas na técnica do vizinho mais próximo, descrita no capítulo sobre RBC (Watson, 1997).

A primeira fórmula é utilizada no momento em que os dados de uma nova consulta (entende-se por nova consulta quando uma nova expressão é inserida) são submetidos a uma nova busca no servidor. Neste processo são recuperadas as consultas com grau de similaridade igual ou superior ao indicado pelo usuário. Esta fórmula é aplicada para cada caso na base de casos, isto é, para cada perfil de consulta armazenado, uma vez que há comparação entre as informações do perfil da consulta atual e as informações de cada perfil de consulta armazenado na base de casos. Esta primeira fórmula é expressa da seguinte forma:

$$Sim(X) = \left[\left(\frac{\frac{QPI}{QPP} + \frac{QPI}{QPC}}{2} \right) * wp \right] + \sum_{i=0} Opi * wi$$

X = (QPI, QPP, QPC, Op)

Onde: ***QPI*** total de palavras iguais encontradas da relação entre as palavras digitadas do perfil de consulta atual e as palavras contidas no perfil da consulta armazenada na base de casos

QPP quantidade de palavras digitadas para busca

QPC quantidade de palavras contidas no perfil de consulta armazenada na base de casos

Op parâmetro para busca

wp peso aplicado às palavras encontradas (definido pelo especialista da aplicação)

w peso aplicado ao(s) parâmetro(s)

Os parâmetros *Op* e *w* estão diretamente relacionados ao domínio da aplicação, pois cada aplicação pode ter parâmetros específicos com seus respectivos pesos, não havendo necessidade destes pesos serem iguais para todas as opções dos parâmetros utilizados na aplicação. Os parâmetros *wp* e *w* são determinados pelo especialista da aplicação, portanto,

para cada aplicação um valor de peso deve ser definido. Dessa forma, por exemplo, o parâmetro *categoria* pode ter mais influência em uma busca do que o parâmetro *data de atualização* em um site de busca disponível na Internet, ficando a critério do especialista do domínio da aplicação definir.

Para exemplificar a utilização da fórmula de similaridade segue a seguinte situação:

- o site de busca contém um campo texto para digitar a expressão da busca, um parâmetro para escolha da língua (*português*, *inglês*, *etc*) e outro parâmetro para escolher a categoria (*universidade*, *comercial*, *etc*);
- no campo texto é digitada a seguinte expressão para busca "*inteligência artificial distribuída*";
- para o parâmetro língua foi escolhida a opção "*português*";
- para o parâmetro categoria foi escolhida a opção "*educacional"*;

A base de casos, que contém as informações dos perfis das consultas armazenados, possui o perfil de uma consulta com as seguintes informações:

- expressão de busca: "*inteligência artificial*"
- opção do parâmetro língua: "*português*";
- opção do parâmetro categoria: "*comercial*";

Extraindo os valores das informações utilizadas no exemplo, obtêm-se as seguintes informações para os parâmetros utilizados na fórmula do cálculo do grau de similaridade:

- *QPI* é igual a 2, representando às palavras iguais (*inteligência artificial*), resultantes da comparação entre à expressão de busca do perfil de consulta atual (*inteligência artificial distribuída*) e a expressão de busca do perfil de consulta armazenado na base de casos (*inteligência artificial*).
- *QPP* é igual a 3, indicando o total de palavras contidas (expressão de busca) no perfil de consulta atual (*inteligência artificial distribuída*).
- *QPC* é igual a 2, representando o total de palavras contidas no perfil de consulta armazenado (*inteligência artificial*).
- O valor de *op* é 1, indicando que somente um dos parâmetros (*língua igual a português*) é comum entre o perfil de consulta atual e o perfil de consulta armazenado na base de casos.

- O valor do peso aplicado às palavras encontradas, *wp*, é igual a 0,6 (valor definido pelo especialista da aplicação);
- O peso aplicado aos dois parâmetros (*língua e categoria*) é o mesmo, 0,2 (valor definido pelo especialista da aplicação).

Uma observação importante a ser feita, refere-se ao somatório dos valores dos pesos que sempre devem ser igual a 1 (neste exemplo os valores são: 0,6 – valor do peso aplicado à expressão de busca; 0,2 – peso aplicado ao parâmetro língua; e 0,2 – peso aplicado ao parâmetro categoria), pois representa a similaridade de 100%. Os dados estão postados na fórmula a seguir, juntamente com o resultado, sendo o valor do grau de similaridade.

$$Sim = \left[\left(\frac{\frac{2}{3} + \frac{2}{2}}{2} \right) * 0{,}6 \right] + (1 * 0{,}2) = 0{,}698 * 100 \Rightarrow 69{,}8\%$$

Aplicando a fórmula sobre as informações especificadas obtêm-se o valor de 0,698, indicando que a consulta armazenada contém 69,8% de similaridade com a consulta atual. Observando as informações dos perfis de consulta, tanto o perfil de consulta atual como o perfil de consulta armazenado, nota-se que a similaridade não pode ser 100%, afinal a expressão da busca do perfil de consulta atual não é igual à expressão de busca contida no perfil de consulta armazenado na base de casos e, também, apenas um dos valores dos parâmetros é igual em ambas às consultas.

Como o perfil de consulta é formado dinamicamente, cada visita em um registro recuperado é acrescentada ao perfil de consulta, necessitando de nova atualização na relação de consultas similares recuperadas. Para isto, uma nova fórmula para o cálculo do grau de similaridade é necessária, pois as informações relativas aos registros visitados devem ser incorporadas, afinal, os registros visitados podem indicar o (novo) rumo de pesquisa do usuário.

$$Sim(X) = \left[\left(\frac{\frac{QPI}{QPP} + \frac{QPI}{QPC}}{2} \right) * wp \right] + \sum_{i=0}^{n} Opi * wi + \left[\left(\frac{\frac{QRI}{QRV} + \frac{QRI}{QRVC}}{2} \right) * wr \right]$$

***X** = (QPI, PQQ, QPC, Op, QRI, QRV, QRVC)*

Onde: ***QPI*** total de palavras iguais encontradas da relação entre as palavras digitadas do perfil de consulta atual e as palavras contidas no perfil da consulta armazenada na base de casos

QPP quantidade de palavras digitadas para busca

QPC quantidade de palavras contidas no perfil de consulta armazenada na base de casos

Op parâmetro para busca

wp peso aplicado às palavras encontradas (definido pelo especialista da aplicação)

w peso aplicado ao(s) parâmetro(s)

QRI quantidade de registros visitados iguais (resultado da comparação entre os registros visitados contidos no perfil de consulta atual e os registros visitados contidos no perfil de consulta armazenado na base de casos)

QRV quantidade total de registros visitados, presentes no perfil de consulta atual

QRVC quantidade total de registros visitados, presentes no perfil de consulta armazenada na base de casos

wr peso aplicado aos registros encontrados (definido pelo especialista da aplicação)

Para exemplificar a utilização desta nova fórmula são usadas as mesmas informações do exemplo anterior, exceto pela alteração dos valores dos pesos, além de considerar a informação referente à relação dos registros visitados. O perfil de consulta atual é agregado de mais uma informação, que é a visita a um site recuperado. Já o perfil de consulta armazenado na base de casos é composto pela relação de três sites visitados. O site visitado, presente no perfil de consulta atual, é igual a um dos sites visitados contidos

no perfil da consulta armazenada na base de casos. Desta forma, os valores ficam expressos como se segue:

- *QRI* tem valor igual a 1, representando a quantidade de registros iguais visitados em ambos os perfis, perfil de consulta atual e perfil de consulta armazenado na base de casos.
- *QRV*, que é a quantidade de registros visitados, presente no perfil de consulta atual, é igual a 1.
- *QRVC* é igual a 3, indicando a quantidade de registros visitados contidos no perfil de consulta armazenado na base de casos.
- O novo valor de *wp* é 0,4, ao invés de 0,6, que é aplicado à expressão de busca (peso definido pelo especialista da aplicação).
- O valor de *w1* e *w2*, aplicado aos parâmetros (*língua e categoria*), é igual a 0,15, ao invés de 0,2 (peso definido pelo especialista da aplicação).
- *wr* é igual a 0,3, que é o peso aplicado a relação dos registros encontrados (peso definido pelo especialista da aplicação).

$$Sim = \left[\left(\frac{\frac{2}{3} + \frac{2}{2}}{2} \right) * 0{,}4 \right] + (1 * 0{,}15) + \left[\left(\frac{\frac{1}{1} + \frac{1}{3}}{2} \right) \right] * 0{,}3 = 0{,}682 * 100 \Rightarrow 68{,}2\%$$

Como se pode notar, a similaridade diminuiu, se comparado com a similaridade resultante do exemplo da primeira fórmula. A razão desta diminuição do valor do grau de similaridade se refere à inclusão de uma nova informação para este cálculo do grau de similaridade, que são os registros visitados, além da mudança dos pesos aplicados à expressão de busca e aos parâmetros, presentes nos perfis de consulta.

Com relação aos pesos utilizados nesta segunda fórmula, há possibilidade de definir valores dinâmicos, sendo que esta dinamicidade deverá ser definida previamente por um especialista no domínio da aplicação. Por exemplo, na medida em que registros recuperados são visitados o peso aplicado ao parâmetro que representa os registros recuperados visitados pode aumentar e, conseqüentemente, o valor dos pesos aplicados aos

demais parâmetros diminuem, influenciando diretamente a relação de consultas similares recuperadas. Neste caso, pode não haver tanta importância às informações previamente inseridas para a consulta (*expressão de busca* e *opções de parâmetros*), pois a visita aos registros recuperados pode determinar um novo rumo de busca, na verdade, há uma especificação mais detalhada da informação que se está realmente buscando.

4.2.5 Base de casos

A base de casos contém as informações sobre os perfis de consultas, armazenados em forma de casos, portanto, cada caso representa um perfil de consulta. Quando uma consulta é efetuada, as informações pertencentes ao perfil desta consulta são armazenadas na base de casos, formando um novo perfil, conseqüentemente, um novo caso. Cada perfil de consulta armazenado é identificado por um código único.

A estrutura de armazenamento é composta por três tabelas, sendo que o código do perfil de consulta é o elemento de relação entre as tabelas. Há uma tabela principal chamada *EN_DADOS_CONSULTA* que contém as informações genéricas do perfil de consulta. Nesta tabela cada registro armazenado representa um perfil de consulta. Os atributos que compõem esta tabela são responsáveis pelas informações sobre a expressão de busca, as opções dos parâmetros escolhidos e a descrição da *query* utilizada para recuperar os registros na base de dados.

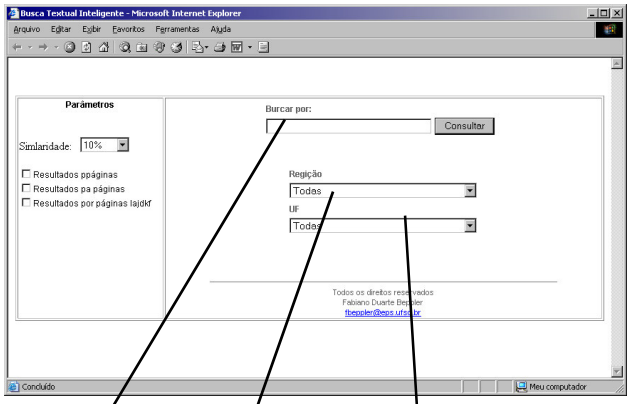
Na segunda tabela, chamada *EN_REGISTROS_RECUPERADOS*, ficam as informações sobre os registros recuperados pela consulta, isto é, cada registro desta tabela é composto por uma chave única, composta pelo identificador do registro recuperado da base de dados e o identificador do perfil da consulta, armazenado na tabela *EN_DADOS*_CONSULTA. O identificador do registro recuperado da base de dados é o elemento responsável para localizar de maneira objetiva as demais informações sobre o mesmo. Como uma consulta pode recuperar mais de um registro, obrigatoriamente todos os registros recuperados são relacionados com a consulta, portanto, um perfil de consulta pode ser composto por vários registros recuperados da base de dados. Por exemplo, se uma determinada consulta recupera 35 sites (isto é, registros contidos na base de dados), conseqüentemente, há 35 registros armazenados na tabela *EN_REGISTROS_RECUPERADOS*, vinculados diretamente ao perfil de consulta através do identificador que está armazenado na tabela *EN_DADOS_CONSULTA*.

Na terceira tabela, chamada *EN_REGISTROS_VISITADOS*, ficam as informações sobre os registros visitados, também como parte do perfil de consulta. Cada registro desta tabela é formado por uma chave composta pelo identificador do perfil de consulta e o identificador do registro visitado, recuperado da base de dados, muito similar a tabela *EN_REGISTROS_RECUPERADOS*. Como a tabela *EN_REGISTROS_RECUPERADOS*, a tabela *EN_REGISTROS_VISITADOS* também pode conter vários registros para um determinado perfil de consulta, pois quando se recuperam vários registros é normal à visita em mais de um destes. Por exemplo, após fazer uma consulta tem-se como resultado 42 sites, mas somente 5 destes sites são visitados. O identificador de cada um dos cinco sites na base de dados é armazenado na tabela *EN_REGISTROS_VISITADOS* juntamente com o identificador do perfil de consulta à que pertence.

Quando uma consulta é efetuada e esta não recupera nenhum registro da base de dados, seu perfil não será armazenado na base de casos. Outra observação importante a ser destacada refere-se ao procedimento quanto à tabela *EN_REGISTROS_VISITADOS*, pois o perfil de uma consulta pode ser formado sem nenhum registro visitado, isto é, nenhum dos registros recuperados da base de dados foi visitado. Isto não implica nas informações armazenadas nas demais tabelas. Portanto, dois perfis de consultas podem ter os mesmos parâmetros de busca e índice de similaridade diferente, dependendo das ações de cada usuário quanto aos resultados.

A Figura 4.2 mostra um exemplo de como a base de casos pode ficar. A tabela *EN_DADOS_CONSULTA*, que armazena as informações gerais sobre os perfis de consultas, é formada pelos seguintes atributos: identificador da consulta; texto representando a expressão de busca; parâmetro que representa a região; parâmetro que representa a unidade da federação – UF; e *query* da consulta. Esta primeira tabela está relacionada com as outras duas por meio do identificador da consulta, como descrito no modelo.

A "*consulta1*" recupera três sites e a "*consulta2*" recupera quatro sites, todos armazenados na tabela *EN_REGISTROS_RECUPERADOS* que contém a relação dos sites recuperados. Já a tabela *EN_REGISTROS_VISITADOS* com a relação dos sites visitados contém somente três registros, sendo que um dos registros visitados está relacionado com a "*consulta1*" e os outros dois sites visitados estão relacionados com a "*consulta2*".



EN_DADOS_CONSULTA				
Identificador	**Expressão da busca**	**Opção 1 = Região**	**Opção 2 = UF**	***Query***
Consulta1	Inteligência aplicada	Sul	SC	Selecione os registros que contenham "inteligência aplicada" que estejam na região "sul" e na UF de "SC"
Consulta2	Saúde coletiva	Nordeste	RN	Seleciona os registros que contenham "saúde coletiva" e que estejam na região "nordeste" e na UF de "RN";

EN_REGISTROS_RECUPERADOS	
Ident. da consulta	**Ident. do registro**
Consulta1	Registro23
Consulta1	Registro38
Consulta1	Registro51
Consulta2	Registro12
Consulta2	Registro25
Consulta2	Registro27
Consulta2	Registro48

EN_REGISTROS_VISITADOS	
Ident. da consulta	**Ident. do registro**
Consulta1	Registro38
Consulta2	Registro27
Consulta2	Registro48

Figura 4.2 - Exemplo da base de casos

4.3 Descrição da arquitetura

A Figura 4.3 apresenta a arquitetura conceitual do Sistema RBNet. Os elementos genéricos para sua aplicação são o *site de buscas*, *repositório de informações* e *páginas de resultados*. Estes três elementos estão presentes em qualquer site Internet que disponibiliza informações. O *site de busca* caracteriza a interface pela qual os usuários declaram a informação que desejam obter. Deste, todos os parâmetros podem servir de atributos para os casos tratados no RBNet. O *repositório de informações* consiste no local onde residem

os dados disponíveis no site. Estes podem estar em base de dados relacional, diretório de páginas HTML ou qualquer local com informações digitais. O terceiro elemento presente em todo site de buscas é a *página de resultados*. Trata-se do conjunto de páginas montadas pelo site com os resultados das buscas realizadas.

O segundo conjunto de elementos traz os componentes do RBNet. O primeiro é o processo de tornar os parâmetros de busca do site como atributos dos casos. O mesmo ocorre para os resultados das consultas. O RBNet tem o elemento de *Interface RBNet*, que apresenta casos semelhantes e permite que o usuário recupere o histórico destas consultas sempre que desejar. O último elemento do RBNet é o *Banco de Casos Anteriores*, onde são armazenadas consultas passadas.

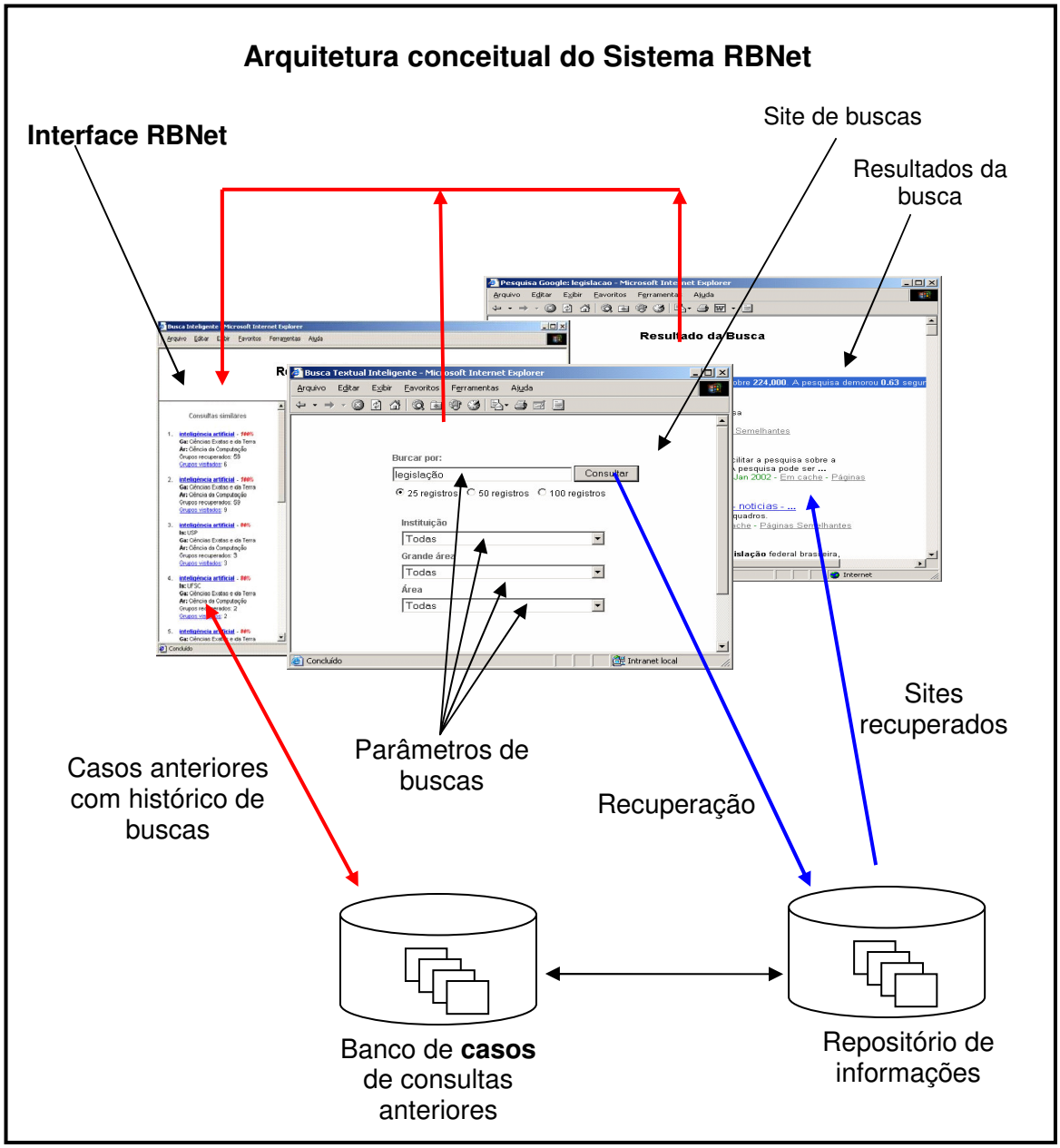


Figura 4.3 - Arquitetura conceitual do Sistema RBNet

O esquema anterior é genérico o suficiente para atender a qualquer site de busca na Internet. De forma geral, o usuário irá interagir com o site de busca, onde a ferramenta RBNet foi inserida, de maneira similar aos sites de busca disponíveis na Internet, como, por exemplo, o *Cadê* (www.cade.com.br) e *Yahoo* (www.yahoo.com). É necessário inserir a expressão para a busca e, se desejar, selecionar algumas opções de parâmetros disponíveis com objetivo de minimizar ou maximizar a abrangência da busca, como, por exemplo, língua (sites em língua portuguesa, inglesa, etc) e quantidade de registros por página.

Nestes sites, após o preenchimento das informações necessárias, a *query* é enviada ao servidor para ser executada. Ao chegar no servidor é disparado um processo de busca sobre os dados armazenados no banco de dados, e os registros ou sites que atendem aos parâmetros requeridos são enviados para o usuário; é o resultado da consulta. A Figura 4.4 mostra o processo utilizado em sites de busca na Internet.

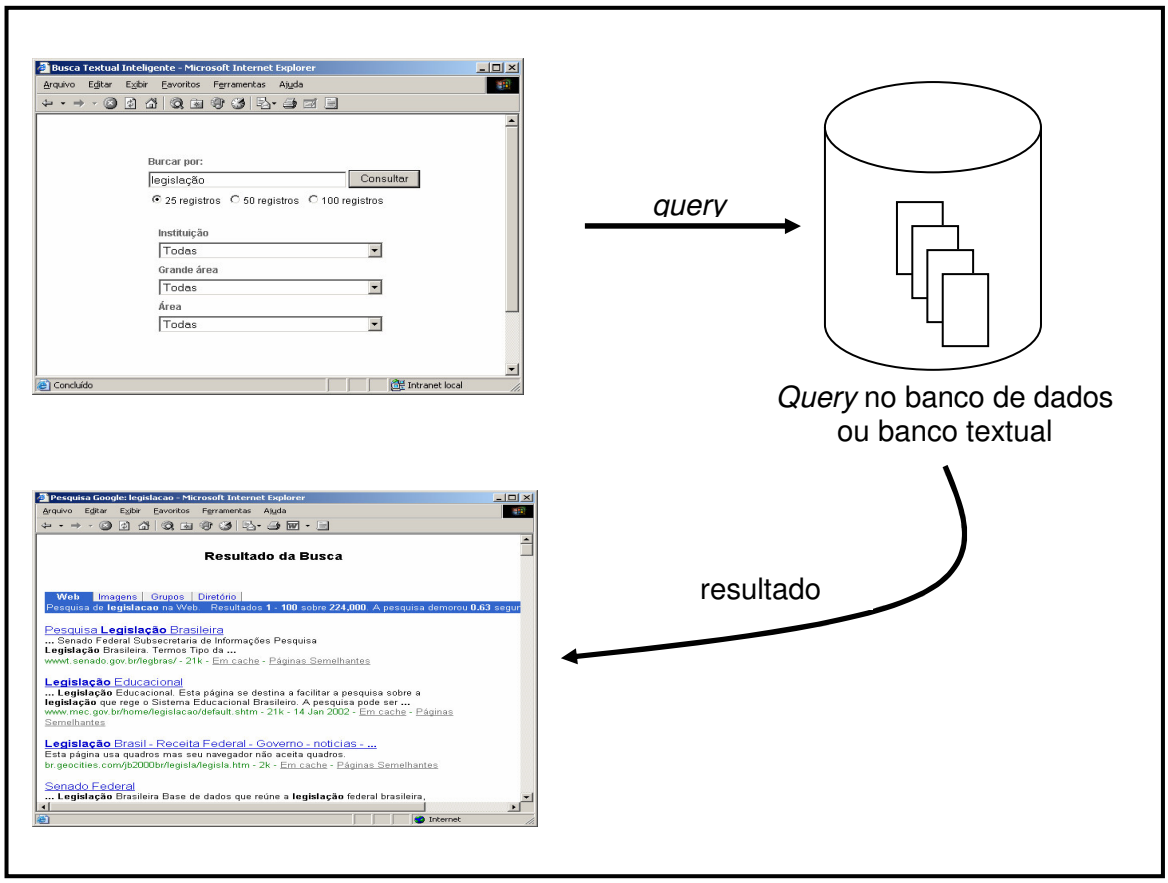


Figura 4.4 - Esquema de consulta de sites Internet

Na arquitetura RBNet pode haver componentes adicionais, como, por exemplo, o grau de similaridade, que tem como função restringir a recuperação de consultas a somente aquelas que possuem grau de similaridade igual ou superior ao indicado. Depois de configurados os parâmetros da consulta e os parâmetros da ferramenta RBNet, uma *query* é formada e submetida ao servidor.

Quando a *query* chega no servidor, é disparado um processo de busca disponível no site (por exemplo, linguagem de consulta em banco de dados relacional SQL[1], sobre a base

[1] SQL (*Structured Query Language*) é a linguagem padrão que permite todo tipo de operações em banco de dados relacionais.

de dados). Terminado este procedimento, é disparado um novo processo para busca de consultas similares sobre a base de casos, que contém os perfis de consulta armazenados. Neste processo, utilizar-se-á a técnica de RBC, possibilitando a recuperação de consultas com grau de similaridade igual ou superior ao grau de similaridade indicado pelo usuário, como mostra a Figura 4.5.

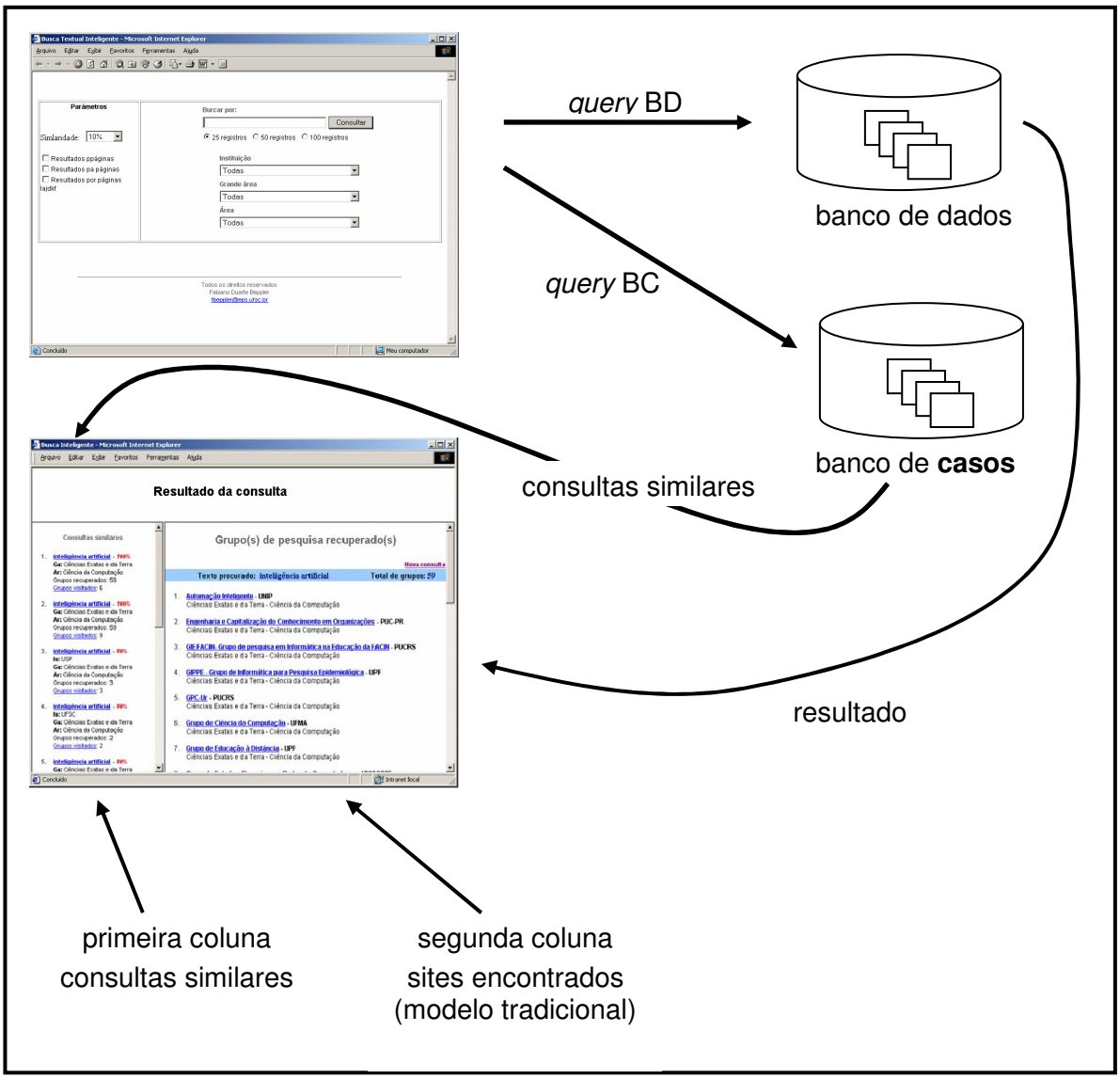


Figura 4.5 - Ciclo de interação do usuário com o site com o RBNet

Na área de *Interface do RBNet*, após completados os dois processos de busca, são apresentadas duas colunas contendo as informações recuperadas. A primeira coluna (área RBNet) apresenta a relação das consultas similares com suas respectivas informações, constantes no perfil de cada consulta, como, por exemplo, expressão da busca, grau de

similaridade e total de registros que recupera. Na segunda coluna (página de resultados do site) é apresentada a lista dos registros recuperados pelo método tradicional de busca.

Há possibilidade de visualizar, por meio da *Interface RBNet*, os registros que cada consulta recupera, por meio de um *link* sobre a descrição da expressão de busca da consulta. Na área de *Interface RBNet* ao se clicar sobre a expressão de uma determinada consulta recuperada (1), uma nova requisição é enviada ao servidor, contendo a informação dos parâmetros desta consulta recuperada. O servidor não necessita fazer uma nova busca sobre toda a base de dados, pois a lista de registros que a consulta recupera faz parte do perfil de consulta, que está armazenado na base de casos. Desta forma, a base de casos informa a base de dados (2) quais os registros que devem ser mostrados na interface do site. Como resultado destes procedimentos, a *Interface RBNet* é preenchida novamente, com as novas consultas similares (3) e a interface do site com a lista dos sites que a consulta similar selecionada recupera (4). A Figura 4.6 mostra a ação do usuário sobre as consultas recuperadas.

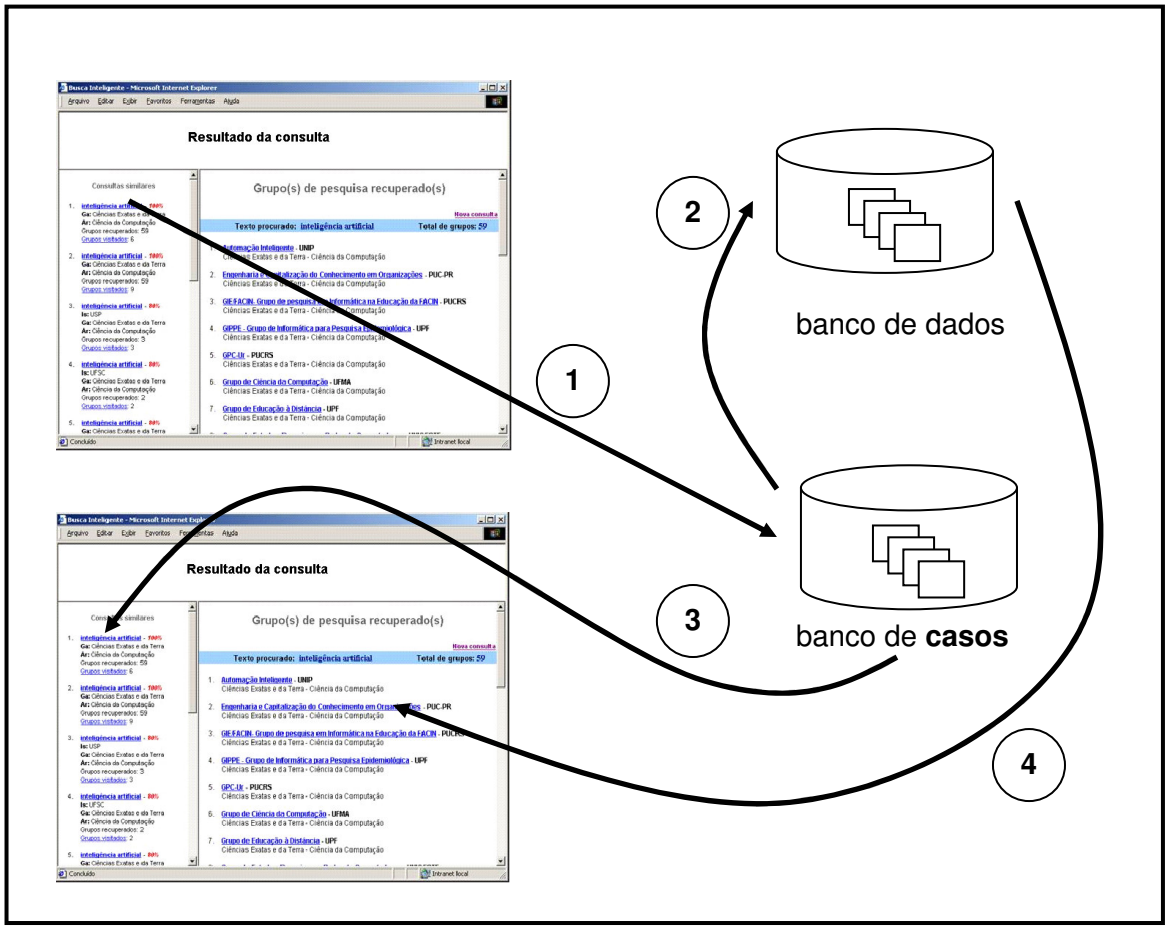


Figura 4.6 - Ação do usuário sobre as consultas recuperadas

Na interface do site, onde são apresentados os resultados recuperados pelo procedimento de busca disponível, geralmente tem-se *link* sobre o título de cada registro recuperado, apontando para o endereço de localização deste na Internet. Quando o usuário visita um dos registros recuperados, é disparado um novo processo no servidor para recuperar novas consultas similares, pois o perfil da consulta atual foi acrescido de novas informações.

O perfil da consulta é formado, também, pelas interações do usuário com as informações recuperadas, conforme já definido no respectivo tópico. Ao visitar um registro recuperado, ocorre uma atualização do perfil da consulta e automaticamente na relação de consultas similares, pois, neste caso, o cálculo de similaridade considera, também, a relação dos registros visitados. Este processo de atualização varre a base de casos através do RBC e por meio deste novo cálculo de similaridade, compara o perfil da consulta atual com os perfis de consulta armazenados, recuperando as consultas com grau de similaridade igual ou superior ao indicado. Ao ocorrer este procedimento, o grau de similaridade das consultas recuperadas é alterado, podendo diminuir ou aumentar.

Com a variação do grau de similaridade, as consultas antes não recuperadas podem ser consideradas válidas e consultas anteriormente recuperadas podem ser dispensadas por não atenderem o grau de similaridade desejado. Esta variabilidade é determinada pelo perfil da consulta, pois um registro visitado faz parte do seu perfil e esta informação é considerada no cálculo do grau de similaridade. Com o novo cálculo, as consultas com a maior quantidade de registros visitados iguais aos registros visitados pela consulta atual, possuem, conseqüentemente, maior grau de similaridade.

Quando instalado como recurso de um site, o RBNet supõe a divisão das telas de consultas, com a inclusão do mecanismo de recuperação inteligente. A Figura 4.7 a seguir apresenta o esquema de funcionamento do RBNet quando incluso em um site de consultas.

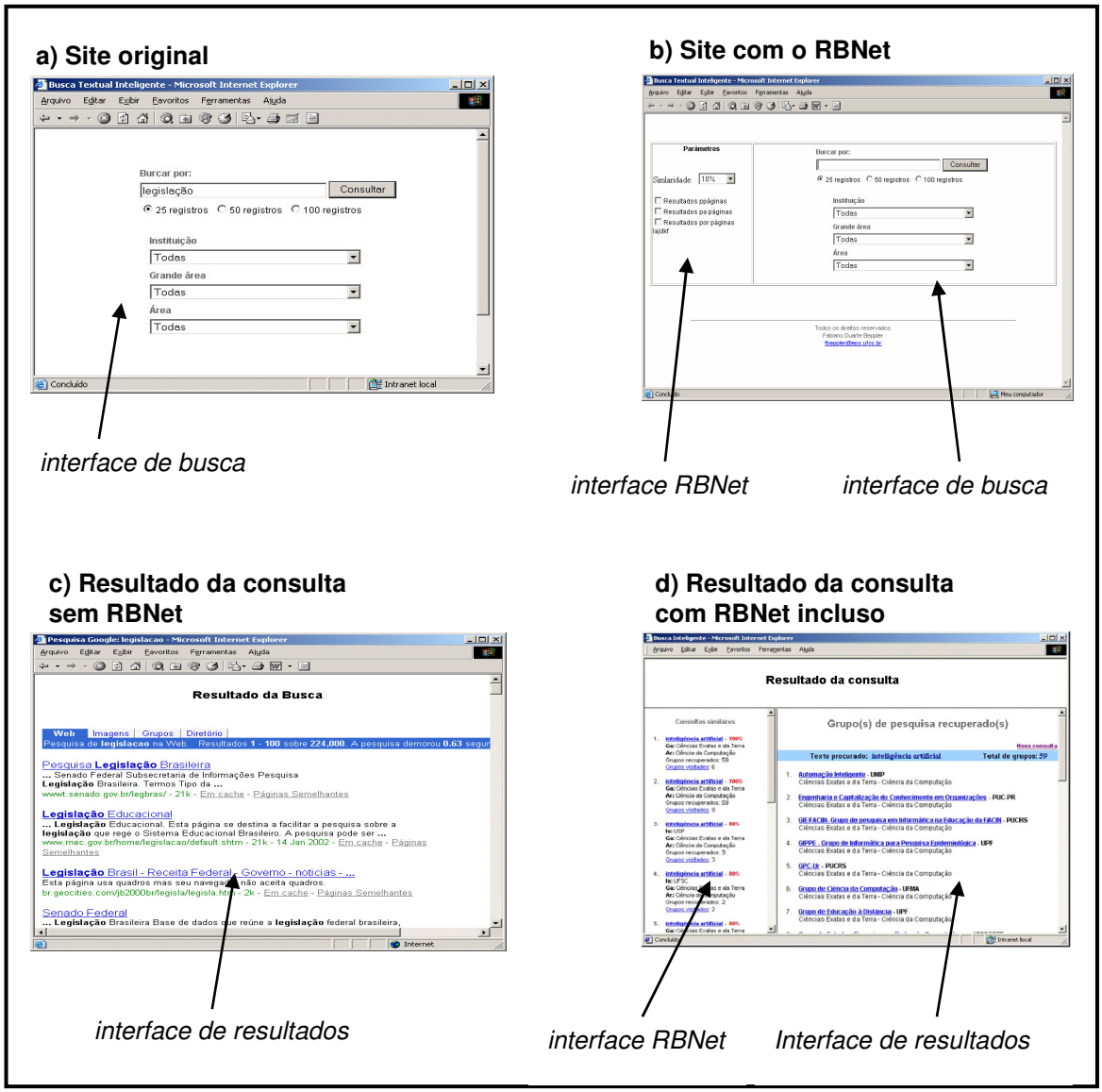


Figura 4.7 - Alteração de layout de sites que incluem o RBNet em seus mecanismos de busca

A Figura 4.7 apresenta as situações de consulta e de apresentação de resultados de busca antes e após a inclusão do RBNet. No caso da página de consulta, a *Interface RBNet* inclui informações gerais sobre RBNet e permite ao usuário retirá-lo da consulta. Na tela de resultados, a *Interface RBNet* mostra os casos de consulta anteriores e os respectivos graus de similaridade entre a consulta atual e os casos passados. Ao clicar sobre um caso na área do RBNet, o usuário pode receber de retorno todos os registros recuperados pela consulta ou somente a relação dos registros que foram visitados, pois estas informações estão armazenadas na base de casos.

O perfil da consulta começa a ser formado quando o usuário faz uma nova busca, sendo que as primeiras informações armazenadas são os parâmetros utilizados para a

consulta, expressão de busca e as opções de parâmetros disponíveis. Este perfil é formado enquanto o usuário age sobre as informações recuperadas, interage com as consultas recuperadas ou visita algum dos registros recuperados.

Outra informação presente no perfil da consulta é a *query* utilizada para recuperar os registros relacionados aos parâmetros da busca. Esta informação é útil para os sites de busca que são alimentados dinamicamente, isto é, o armazenamento de novas informações ou atualizações em sua base de dados é constante. Com esta *query* armazenada no perfil da consulta, há possibilidade de atualização da base de casos referente aos registros que a consulta recupera, isto é, atualização do perfil de consulta.

A conclusão do perfil da consulta ocorre quando o usuário faz uma nova consulta, informando uma nova expressão de busca, novas opções para os parâmetros ou, ainda, quando não houver interação do usuário com o site após um determinado tempo pré-estabelecido pelo servidor da Internet (*timeout*)[1].

4.4 Utilização de outros modelos de recuperação

O modelo proposto para a ferramenta RBNet utiliza a tecnologia de RBC, como já foi apresentado, para efetuar o processo de recuperação de informação na base de casos. Isto não significa que, para utilizar a ferramenta, necessariamente a tecnologia definida e implementada para recuperar informações da base de casos seja o RBC. Como foi descrito no capítulo sobre recuperação de informação, há vários modelos de recuperação de informação que podem ser utilizados na ferramenta. A análise que se deve fazer é verificar as vantagens e desvantagens na utilização do modelo de recuperação escolhido, adaptando suas particularidades para uma estrutura personalizada, objetivando usufruir suas vantagens.

Se, por exemplo, for utilizado o modelo vetorial, apresentado no capítulo sobre recuperação de informação, ao invés de RBC para o módulo de recuperação de consultas similares, é necessário fazer adaptações na base de casos. As informações dos casos devem ser armazenadas em vetores contendo as palavras de cada caso, havendo necessidade de criar uma forma de atribuir um peso para cada palavra. Uma das vantagens de utilizar o

[1] Timeout significa que não houve resposta durante um determinado tempo previsto e a sessão foi fechada.

modelo vetorial é a existência implícita do módulo que calcula o grau de similaridade. Uma das desvantagens em utilizar este modelo é a velocidade na recuperação dos casos similares, pois o processo de recuperação compara cada item do vetor *query* com cada item de cada vetor da base de casos, ou base de vetores. Este processo pode ser rápido em uma base de casos pequena, mas como a base de casos é acrescida de novos casos dinamicamente, a demora na recuperação de casos similares pode ser um grande problema, ou melhor, uma grande desvantagem.

4.5 Considerações finais

O presente capítulo abordou o modelo proposto para recuperação inteligente de informações da ferramenta RBNet. Primeiramente esboçou as vantagens da utilização de sites de busca para encontrar informações na Internet. Mostrou também o conhecimento desperdiçado pelos modelos de sites utilizados para buscar informação, que é o reaproveitamento da informação das buscas dos usuários no auxílio à busca de informações de outros usuários.

Os módulos de entrada e saída, componentes da ferramenta RBNet, descrevem as informações que são utilizadas através da interação do usuário com o site, presentes no módulo de entrada, e as informações que são apresentadas ao usuário no final da execução dos processos, presente no módulo de saída.

Detalhou-se o cálculo do grau de similaridade, que representa o valor de semelhança entre a consulta de entrada e as consultas armazenadas na base de casos. Este item descreve os cálculos do grau de similaridade necessários para serem aplicados no processo de recuperação, expondo a necessidade de cálculos diferentes em momentos diferentes, com intuito de tornar ainda mais eficaz o processo de recuperação.

A base de casos é composta pelas informações das consultas e outras informações referentes à interação de processos sobre a base de dados e a recuperação de informações. Estas informações, descritas de forma detalhada, formam o perfil de uma consulta e são armazenadas na base de casos.

A ferramenta RBNet tem como principal objetivo o aproveitamento das interações de usuários em sites de disponibilização de informações como subsídios a novos usuários que utilizam os mesmos recursos em novas buscas. Para verificar o grau com que a

arquitetura proposta para o RBNet cumpre com seu objetivo, discute-se no próximo capítulo sua aplicação a um site de disponibilização de informações sobre Ciência & Tecnologia. Trata-se do Diretório dos Grupos de Pesquisa, projeto do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) que tem no grupo de pesquisa sua unidade central de análise.

5 Aplicação do RBNet utilizando os dados do Diretório dos Grupos de Pesquisa no Brasil

5.1 Introdução

Para demonstrar a viabilidade e usabilidade do RBNet, foi escolhida a base de dados do projeto do Diretório dos Grupos de Pesquisa no Brasil, projeto do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Uma das ferramentas do projeto do Diretório dos Grupos de Pesquisa no Brasil é a Busca Textual, que permite recuperar informações sobre qualquer dado presente na base do Diretório. Para Busca Textual, a base de dados do Diretório disponibiliza, além das informações cadastradas nos grupos de pesquisa, informações complementares que foram migradas da base de Currículo Lattes, como, por exemplo, as referentes à produção científica, tecnológica e artística dos integrantes dos grupos. Esta Busca Textual contém funcionalidades semelhantes a outro site de busca tradicional na Internet, como por exemplo o *Google* (www.google.com).

Em aproximadamente dois anos de uso, cerca de metade dos usuários que utilizaram esta Busca Textual sobre as informações do Diretório realizou somente uma consulta. O RBNet pode ser utilizado para auxiliar os usuários a encontrar as informações requeridas e, por meio das consultas recuperadas, oferecer uma visão mais abrangente das informações que estão contidas no Diretório. Além disso, acredita-se que o RBNet possa classificar de forma indireta os usuários do site.

A comunidade científica, principal usuário do site do Diretório dos Grupos de Pesquisa, tem como uma de suas características a multiplicidade de interesse, segundo os diversos domínios do conhecimento. Um dos recursos que o RBNet pode oferecer é justamente a classificação indireta e transparente destes interesses. Assim, um usuário interessado em determinado tema (ex.: procurando pesquisas em "*violência urbana*") conhecerá o que outros usuários consultaram em temas correlatos. O resultado pode ser, inclusive, o aumento no número de consultas realizadas.

5.2 Diretório dos Grupos de Pesquisa no Brasil

O Diretório dos Grupos de Pesquisa no Brasil é um projeto desenvolvido pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) desde 1992. O projeto tem por objetivo formar bases censitárias sobre todos os grupos de pesquisa em atividade no país ligadas a ensino superior e a empresas estatais. As informações disponíveis indicam que na quarta versão com dados relacionados ao triênio 1997 – 2000 o Diretório conseguiu alcançar uma cobertura acima de 80% dos grupos de pesquisa em atividade no país nestas instituições (CNPq 2001).

As informações constantes na base de dados do Diretório de Grupos de Pesquisa dizem respeito: (a) aos recursos humanos participantes nos grupos; (b) às linhas de pesquisa em andamento; (c) às especialidades do conhecimento; (d) setores de atividades envolvidos; (e) os cursos de mestrado e doutorado com os quais o grupo interage; e (f) à produção científica e tecnológica nos dois anos e meio imediatamente anteriores à época da captura dos dados. Além disso, cada grupo é localizado no espaço e no tempo (Guimarães, 1999).

A primeira versão (1.0) apresentou informações referentes ao segundo semestre de 1993, onde a produção científica e tecnológica corresponde ao triênio 1990-1992. A versão 2.0 apresentou informações referentes ao segundo semestre de 1995 e a produção científica e tecnológica correspondeu ao biênio 1993-1994. A versão 3.0 apresentou informações referentes ao segundo semestre de 1997 e à produção científica e tecnológica do período de 1º de janeiro 1995 a 30 de junho de 1997. A versão (4.0) apresenta informações sobre os grupos coletados no primeiro semestre de 2000 e à produção científica e tecnológica do período de janeiro de 1997 a dezembro de 2000. Esta produção, relativa aos pesquisadores e estudantes cadastrados no Diretório no primeiro semestre de 2000, é oriunda da base de Currículo Lattes do CNPq com informações disponíveis em 1º de junho de 2001 (CNPq, 2001).

A base de dados do Diretório possui três finalidades principais (CNPq, 2000 e 2001):

- no que se refere à sua utilização pela comunidade científica e tecnológica no dia-a-dia do exercício profissional, é um eficiente instrumento para o intercâmbio e para troca de informações. Com “precisão e rapidez”, é capaz de

responder quem é quem, onde se encontra, o que está fazendo e o que produziu recentemente.

- no âmbito do planejamento e gestão em ciência e tecnologia (C&T), o Diretório é, talvez, a mais poderosa ferramenta já desenvolvida no Brasil. Seja no nível das instituições, seja no das sociedades científicas ou, ainda, no das várias instâncias de organização político-administrativa do país, a base de dados é uma fonte inesgotável de informação.
- a base de dados, na medida em que se pretende recorrente, virá a ter cada vez mais um importante papel na preservação da memória da atividade científico-tecnológica no Brasil.

Segundo CNPq (2001), desde a versão 3.0, o Diretório é capaz de descrever com precisão os limites e o perfil geral da atividade científico-tecnológica no Brasil. Igualmente, é capaz de fornecer aos interessados uma grande massa de informação, bastante diversificada, sobre detalhes de quem realiza as atividades, como e onde se realizam e sobre o quê.

A versão 4.0 do Diretório integrou-se à Plataforma Lattes[1] por meio do sistema de Currículo Lattes, utilizando as informações relativas aos dados dos pesquisadores e estudantes e suas respectivas produções. O Sistema de Currículo Lattes é o instrumento eletrônico de atualização curricular do MCT, CNPq, Finep e da CAPES/MEC cujo principal objetivo é aprimorar a qualidade das informações curriculares (CNPq, 2001a).

Após completar o censo das informações, o Diretório disponibiliza um site com Busca Textual, Plano Tabular, Súmula e Estratificação dos grupos de pesquisa brasileiros. O site de Busca Textual constitui o alvo da pesquisa e a aplicação do RBNet. Na versão 4.1, em um ano havia cerca de 70.000 acessos ao site, com 93,3% de utilização da Busca Textual e com 50% destes realizando apenas uma busca. Dada a riqueza de informação do Diretório, que permite mapear e encontrar pesquisadores, estudantes e grupos por temas de pesquisa, o nível de consultas está aquém dos recursos disponibilizados.

Ao propor a inserção do RBNet no site da Busca Textual, está-se considerando a oportunidade para que os novos usuários possam se valer das consultas anteriores e, com

[1] A Plataforma Lattes é um conjunto de sistemas computacionais do CNPq que visa compatibilizar e integrar as informações em toda interação da agência com seus usuários.

isso, cheguem mais rapidamente às informações que desejam obter. Nas seções a seguir, apresenta-se o processo de inserção do RBNet no site da Busca Textual do Diretório.

5.3 Descrição da base de dados

A definição metodológica mais importante na constituição da base de dados foi a de sua unidade de análise. O grupo de pesquisa foi definido como um conjunto de indivíduos organizados hierarquicamente (CNPq, 2000):

- cujo fundamento organizador são a experiência, o destaque e a liderança no terreno científico ou tecnológico;
- em que há envolvimento profissional e permanente com atividades de pesquisa;
- no qual o trabalho organiza-se em torno de linhas comuns de pesquisa;
- que, em algum grau, compartilha instalações e equipamentos.

Cada grupo de pesquisa organiza-se em torno de uma liderança (eventualmente duas), que é a fonte das informações constantes na base de dados. O conceito de grupo admite aquele composto por apenas um pesquisador. Na quase totalidade dos casos, esses grupos se compõem de pesquisador e de seus estudantes (CNPq, 2000).

Uma das principais características do Diretório de Grupos de Pesquisa é o caráter institucional de suas informações. Os dados constantes na base têm responsabilidade compartilhada entre Dirigentes Institucionais de Pesquisa (pró-reitores ou diretores de P&D), líderes de grupos de pesquisa, pesquisadores e estudantes destes grupos.

A identificação dos líderes dos grupos e a certificação dos grupos na base de dados, são responsabilidades dos dirigentes de pesquisa das instituições participantes. As informações referentes ao grupo como um todo, aos pesquisadores, aos estudantes, ao pessoal de apoio técnico e às linhas de pesquisa são de responsabilidade dos líderes dos grupos de pesquisa. Alguns dados pessoais sobre os pesquisadores, estudantes e aqueles relativos à produção científica, tecnológica e artística são de responsabilidade de cada pesquisador, que as informam com a atualização dos seus Currículos Lattes.

O site de Busca Textual ao qual se aplicou o RBNet tem como repositório de informações a base de dados da versão 4.0 do Diretório. As informações contidas referem-se aos grupos de pesquisa de 224 instituições que participaram da atual versão. O total de

grupos de pesquisa é de 11.760, contendo 48.781 pesquisadores e 41.539 linhas de pesquisa. O total de estudantes presentes na quarta versão do Diretório é de 60.225 (CNPq, 2000). A Tabela 5.1 mostra a evolução histórica do projeto do Diretório dos Grupos de Pesquisa no Brasil.

Tabela 5.1 Evolução histórica do projeto Diretório dos Grupos de Pesquisa no Brasil

Unidades / Ano	1993	1995	1997	2000
Instituições	99	158	181	224
Grupos de pesquisa	4.404	7.271	8.632	11.760
Linhas de pesquisa	15.854	21.523	25.483	41.539
Pesquisadores	21.541	26.779	34.040	48.781
Estudantes	33.565	61.345	115.696	60.225

Fonte: CNPq

Ao longo de suas quatro primeiras versões, o Diretório foi aumentando o mapeamento da pesquisa nacional, aproximando-se cada vez mais de seus objetivos referentes ao caráter censitário da sua base de dados. Com isto, é importante ressaltar que não é adequado concluir fatos sobre a pesquisa nacional tendo por base a comparação longitudinal no tempo (ex.: deduzir aumento de 100% de pesquisadores envolvidos em pesquisa, entre 93 e 2000). Porém, a comparação de números relativos tem em outras bases nacionais correspondência nas deduções (ex.: a participação de doutores na pesquisa nacional cresce a média de 2% a cada 2 anos).

5.4 Aplicação de RBNet à busca textual do Diretório de Grupos de Pesquisa (v. 4.0)

5.4.1 Perfil de consulta

O perfil de consulta representa as informações que caracterizam uma busca realizada por um usuário. Além das informações obtidas da interação do usuário, há, também, as informações referentes à manutenção da base de casos, pois a base de dados pode ser alimentada dinamicamente, o que torna necessário atualizar a informação referente à busca de cada perfil de consulta, armazenado na base de casos.

O perfil da consulta é composto por:

- **identificador**: o identificador é único e representa uma consulta na base de casos;
- **expressão da consulta:** palavra ou frase digitada pelo usuário para busca na base de dados;
- **parâmetros:** são as opções dos parâmetros disponíveis ao usuário. No caso da busca textual sobre as informações dos grupos de pesquisa, os parâmetros disponíveis são: instituição, grande área do conhecimento e área do conhecimento;
- **registros recuperados:** é a relação dos grupos de pesquisa recuperados pela consulta;
- **registros visitados:** é a relação de grupos de pesquisa visitados em cada consulta**;**
- ***query*:** é o SQL utilizado para fazer a recuperação dos grupos de pesquisa na busca sobre a base de dados. Esta informação é importante para o caso de atualização de informações na base de dados, desta forma, a relação de grupos de pesquisa recuperados em cada consulta mantém-se atualizada.

5.4.2 Módulo de entrada

O módulo de entrada é composto pelas informações necessárias para se efetuar inicialmente uma busca. Essas informações fazem parte do perfil de consulta, e são descritas a seguir:

- **expressão de busca:** representa a informação a ser buscada, isto é, palavra ou texto inserido no campo texto, por exemplo *"energia elétrica"*;
- **parâmetros:** na aplicação os parâmetros disponíveis para buscas são informações ligadas diretamente ao grupo de pesquisa. Um dos parâmetros é a instituição a qual o grupo pertence. Cada grupo de pesquisa está vinculado somente a uma instituição. Este parâmetro possui 224 opções, representando a quantidade total de instituições participantes do Diretório, isto é, instituições que possuem algum grupo de pesquisa. Os outros parâmetros de consulta são a grande área do conhecimento e a área do conhecimento predominantes no grupo (no Diretório cada grupo de pesquisa possui uma única grande área do

conhecimento e área do conhecimento predominante). O parâmetro que representa a grande área do conhecimento é composto por 8 opções, representando todas as opções de macro classificação disponíveis aos grupos de pesquisa. O parâmetro referente à área do conhecimento é composto por 76 opções, todas como subclassificações de uma grande área do conhecimento (quando o usuário escolhe uma grande área do conhecimento, as opções da área do conhecimento não são mais 76 e sim as áreas do conhecimento relativas à grande área do conhecimento escolhida).

- **grau mínimo de similaridade:** no RBNet o usuário pode definir o percentual do grau de similaridade mínimo para recuperação de consultas. A faixa de valores para esta opção varia de 10 até 100, onde 100 refere-se aos perfis de consultas cem por cento similares ao perfil da consulta atual. Por exemplo, caso o valor do grau de similaridade escolhido seja de 85%, somente as consultas com grau de similaridade igual ou superior a 85% são recuperadas.
- **registros visitados:** na aplicação, tratam-se das páginas específicas a cada grupo de pesquisa que o usuário consultar seu detalhamento. Todos os grupos de pesquisa visitados fazem parte do perfil de consulta, que servem de auxílio para recuperação de consultas similares.

5.4.3 Módulo de saída

O módulo de saída é formado pelas informações apresentadas ao usuário após a submissão de alguma requisição (busca de informações). Na aplicação o módulo de saída é composto pelas seguintes informações:

- **consultas similares:** são as informações referentes às consultas similares recuperadas da base de casos através da técnica de RBC. As seguintes informações são apresentadas: expressão de busca, opção de cada parâmetro escolhido, total de grupos de pesquisa que recupera, total de grupos de pesquisa visitados e o percentual do grau de similaridade;
- **registros recuperados:** são os grupos de pesquisa recuperados, referentes as informações inseridas para busca. As informações apresentadas de cada grupo

de pesquisa recuperado são: nome do grupo, nome do líder do grupo, grande área e área de conhecimento predominante do grupo de pesquisa;

- **detalhamento do grupo de pesquisa:** são as informações que compõem o grupo de pesquisa. O usuário pode visualizar o detalhamento de todas as informações do grupo de pesquisa, como, por exemplo, dados de identificação, linhas de pesquisa em que atua, informações individuais de cada pesquisador, estudante e produção C&T.

5.4.4 Cálculo do grau de similaridade

No RBNet há a aplicação do cálculo do grau de similaridade em dois momentos. Para cada uma das situações há uma fórmula específica para o cálculo do grau de similaridade, como foi definido na arquitetura do RBNet.

Quando o usuário submete a consulta ao servidor, ocorre a execução de dois processos. O primeiro processo é responsável pela busca de grupos de pesquisa sobre a base de dados e o segundo processo (utilizando a técnica de RBC e conseqüentemente o cálculo do grau de similaridade) é aplicado sobre a base de casos para recuperar as consultas com similaridade igual ou superior à similaridade definida pelo usuário. Neste cálculo do grau de similaridade, são usadas a expressão da consulta e as opções dos parâmetros escolhidos (instituição, grande área e área do conhecimento). Na aplicação realizada, simulações com a ferramenta e o meu próprio conhecimento da base de dados indicaram como sugestão para o peso aplicado à expressão da consulta o valor 0,4 e para cada uma das opções dos parâmetros o valor do peso de 0,2. Desta forma, a fórmula utilizada na Busca Textual do Diretório é a seguinte:

$$Sim(X) = \left[\left(\frac{\frac{QPI}{QPP} + \frac{QPI}{QPC}}{2} \right) * 0{,}4 \right] + (Is * 0{,}2) + (Gr * 0{,}2) + (Ar * 0{,}2)$$

$\boldsymbol{X}$ = *(QPI, QPP, QPC, Is, Gr, Ar)*

Onde: ***QPI*** total de palavras encontradas da relação entre as palavras digitadas e as palavras da consulta armazenada

QPP quantidade de palavras digitadas para busca

QPC quantidade de palavras contidas na consulta armazenada

Is pode valer um se a instituição escolhida é igual a instituição da consulta armazenada, do contrário vale zero

Gr pode valer um se a grande área escolhida é igual a grande área da consulta armazenada, do contrário vale zero

Ar pode valer um se a área escolhida é igual a área da consulta armazenada, do contrário vale zero

Como já foi descrito na arquitetura do RBNet há a possibilidade de definição para os parâmetros a serem utilizados. Para a aplicação do RBNet sobre a base de dados do Diretório foram definidos três parâmetros *Is*, *Gr* e *Ar* com seus respectivos pesos. Nesta primeira fórmula os pesos aplicados a esses parâmetros são iguais, diferentemente da segunda, onde o peso dos parâmetros é diferente para cada um.

Para exemplificar este cálculo de similaridade supõe-se que se está buscando pela expressão "*data mining*" e escolhendo-se a grande área "*Ciências Exatas e da Terra*". Na base de casos há uma consulta similar, contendo também a expressão de busca "*data mining*", mas como opção do parâmetro grande área "*Engenharias*". Para o cálculo é necessário definir os valores correspondentes para cada elemento da fórmula. O valor de *QPI* é igual 2, pois a expressão é composta por duas palavras, sendo que as duas palavras são iguais às palavras da consulta armazenada. Para *QPP* e *QPC* os valores são 2, correspondendo ao total de palavras informadas. Como a instituição e a área não foram selecionadas em ambas as consultas, o valor de *Is* e de A*r* é 1. Já o valor do parâmetro *Gr* é 0, pois esta opção selecionada para busca é diferente do valor contido na consulta armazenada. Aplicando os dados na fórmula verifica-se que a similaridade da consulta armazenada na base de casos é igual a 80%.

$$Sim = \left[\left(\frac{\frac{2}{2} + \frac{2}{2}}{2} \right) * 0{,}4 \right] + (1 * 0{,}2) + (0 * 0{,}2) + (1 * 0{,}2) = (0{,}80 * 100) \Rightarrow 80\%$$

Depois de concluídos esses processos, os dados resultantes são apresentados ao usuário, oferecendo duas possibilidades de interação. A primeira consiste em recuperar os grupos de pesquisa a partir da consulta similar recuperada. A segunda possibilidade é visitar um dos grupos de pesquisa recuperados referente aos dados informados pelo usuário. Caso o usuário escolha a segunda possibilidade, novamente são disparados dois novos processos no servidor. O primeiro processo é o da seleção dos dados do grupo de pesquisa escolhido para a visualização detalhada. O segundo processo consiste em uma nova busca de consultas similares, sobre a base de casos, utilizando uma nova fórmula para o cálculo do grau de similaridade.

Para este novo cálculo são usados, além da expressão da consulta e das opções dos parâmetros da consulta, a relação dos grupos de pesquisa visitados pelas consultas. Os valores dos pesos são diferentes para este novo cálculo, afinal, há mais uma variável a ser considerada. Desta forma, o novo valor do peso para a expressão da consulta é de 0,3, para cada uma das opções dos parâmetros é 0,1 e para os grupos visitados o valor do peso é de 0,4. A seguir, há demonstração da fórmula para o novo cálculo de similaridade.

$$Sim(X) = \left[\left(\frac{\frac{QPI}{QPP} + \frac{QPI}{QPC}}{2} \right) * 0{,}3 \right] + (Is * 0{,}1) + (Gr * 0{,}1) + (Ar * 0{,}1) + \frac{\frac{QRI}{QRV} + \frac{QRI}{QRVC}}{2} * 0{,}4$$

***X** = (QPI, QPP, QPC, Is, Gr, Ar, gvi, gva, gvca)*

Onde: ***QPI*** total de palavras encontradas da relação entre as palavras digitadas e as palavras da consulta armazenada

QPP quantidade de palavras digitadas para busca

QPC quantidade de palavras contidas na consulta armazenada

Is pode valer um se a instituição escolhida é igual à instituição da consulta armazenada, do contrário vale zero

Gr pode valer um se a grande área escolhida é igual a grande área da consulta armazenada, do contrário vale zero

Ar pode valer um se a área escolhida é igual à área da consulta armazenada, do contrário vale zero

QRI quantidade de grupos iguais visitados

QRV quantidade de grupos visitados pela consulta atual

QRVC quantidade de grupos visitados pela consulta armazenada

Para exemplificar este novo cálculo de similaridade supõe-se que um grupo de pesquisa resultante da busca, exemplificada anteriormente, é visitado. A consulta recuperada como similar possui dois grupos visitados, dentre os quais um deles é igual ao grupo visitado na consulta atual. Para o cálculo do grau de similaridade os demais valores permanecem, alterando somente o valor dos pesos. Para *QRI* o valor é igual a 1, representando que somente um grupo foi visitado por ambas consultas. O valor de *QRV* também é igual a 1, pois na consulta atual é visitado somente um grupo e o valor de *QRVC* é 3, representando os dois grupos visitados pela consulta armazenada. A seguir, seguem os números postados na fórmula, que tem como resultado o novo valor de similaridade da consulta, 77%.

$$Sim = \left[\left(\frac{\frac{2}{2} + \frac{2}{2}}{2} \right) * 0{,}3 \right] + (1 * 0{,}1) + (0 * 0{,}1) + (1 * 0{,}1) + \left[\left(\frac{\frac{1}{1} + \frac{1}{3}}{2} \right) \right] * 0{,}4 = 0{,}77 * 100 \Rightarrow 77\%$$

Este novo valor indica que a consulta anterior possuía um grau de similaridade igual a 80% e após visitar um grupo recuperado sua similaridade diminuiu para 77%. Se o usuário visitar outro grupo, e este grupo ter, também, sido visitado pela consulta armazenada, o grau de similaridade irá aumentar para 83%, indicando que outro usuário já procurou algo muito similar ao que se está procurando agora, oferecendo a possibilidade de visualizar, além de todos os grupos de pesquisa recuperados pela consulta armazenada, somente os grupos de pesquisa visitados pelo usuário.

5.4.5 Base de casos para a aplicação

A inclusão da ferramenta RBNet inclui a base de casos no conjunto de bases que o site acessa. Esta base de casos irá conter as informações requeridas para aplicação do modelo. A base de casos é composta por três tabelas relacionadas, como mostra a Figura

5.1. Nestas se dá o armazenamento das informações dos perfis das consultas feitas pelos usuários sobre as informações dos grupos de pesquisa.

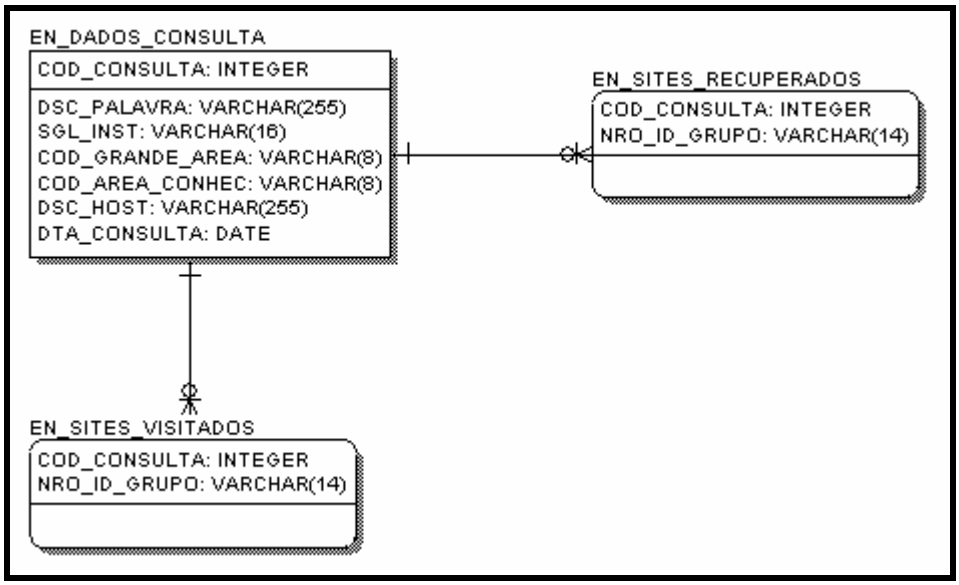


Figura 5.1 - Modelo das tabelas utilizadas na base de casos

A primeira tabela, chamada EN_DADOS_CONSULTA, contém as informações gerais do perfil das consultas, descritos pelos seguintes atributos:

- **código da consulta:** é um código único para cada consulta que serve como identificador da consulta e como objeto de ligação entre as outras tabelas que armazenam as outras informações relacionadas à consulta;
- **descrição do texto da consulta:** este atributo contém a expressão utilizada para a busca. No site da Busca Textual é obtido da janela de edição dos termos da consulta a grupos;
- **sigla da instituição:** contém a opção do parâmetro sigla da instituição, caso tenha sido escolhido, servindo para refinamento da consulta, como parâmetro na busca;
- **código da grande área do conhecimento:** representa a identificação da grande área do conhecimento, caso escolhida, para refinamento da consulta, representando um parâmetro para busca;

- **código da área do conhecimento:** contém a informação sobre a área do conhecimento do grupo de pesquisa, caso escolhida, também para refinamento da consulta, como um de parâmetro da busca.

A segunda tabela, chamada EN_SITES_RECUPERADOS, contém as informações sobre os grupos recuperados por cada consulta e é composta por:

- **código da consulta:** este código é obtido da tabela EN_CONSULTA, significando que os registros que contém o mesmo código fazem parte da mesma consulta, e representa os grupos de pesquisa que são recuperados por uma determinada consulta;
- **número de identificação do grupo de pesquisa:** esta informação se refere ao identificador único da tabela de grupos de pesquisa, assim é possível encontrar os grupos recuperados pelas consultas sem necessitar que uma nova busca seja efetuada.

A terceira e última tabela, chamada EN_SITES_VISITADOS, contém as informações dos grupos de pesquisa visitados em cada consulta, e é formada por:

- **código da consulta:** este código representa que uma determinada consulta teve um ou vários grupos de pesquisa visitados pelo usuário;
- **número de identificação do grupo visitado:** representa o número de identificação do grupo de pesquisa que foi visitado pelo usuário.

5.5 Interações do usuário

No que se refere à seção de buscas e consultas, o usuário não nota diferenças no site da busca textual após a inclusão do RBNet. A configuração do site de buscas, com localização de termos, filtros, etc, permanece a mesma. A diferença para o usuário está no resultado de suas consultas, que já aparece comparado aos casos de consultas anteriores. A primeira tarefa é inserir a palavra ou frase que representa a informação que se deseja buscar. Além desta interação há a possibilidade de refinar a busca escolhendo uma das opções dos parâmetros fornecidos (*instituição*, *grande área* ou *área do conhecimento*). A escolha dos parâmetros é opcional, isto é, o usuário pode simplesmente digitar a palavra ou frase a ser buscada e clicar sobre o botão “consultar”.

O resultado da consulta é a recuperação de uma lista de consultas similares e de uma lista dos grupos de pesquisa recuperados. A partir deste momento, o usuário pode utilizar o site somente com cliques do mouse sobre os grupos de pesquisa recuperados ou sobre as consultas similares recuperadas.

As consultas similares ajudam na busca das informações desejadas, utilizando as informações de consultas feitas por outros usuários. A informação desejada, muitas vezes não é encontrada por meio dos grupos de pesquisa recuperados pela palavra ou frase digitada na busca, mas por meio dos grupos de pesquisa das consultas similares.

Para cada clique do usuário sobre as informações recuperadas, novas consultas similares são apresentadas e conseqüentemente novos grupos de pesquisa também. Como o cálculo do grau de similaridade é aplicado a cada clique sobre os grupos de pesquisa recuperados, a relação de consultas similares altera-se dinamicamente. Afinal, a idéia é possibilitar a visualização de resultados de consultas com perfil similar ao perfil da consulta que está sendo executada pelo usuário.

Verificando estes procedimentos, nota-se que há muitos casos em que a informação buscada e os grupos de pesquisa resultantes são totalmente irrelevantes, se considerar os grupos de pesquisa que realmente forneceram as informações desejadas. O usuário consegue obter o auxílio de outros usuários de interesse similares e, sem necessitar trocar nenhuma informação com estes usuários obtém conhecimento armazenado por meio das informações dos perfis de consultas, facilitando a busca das informações desejadas.

5.6 Protótipo

Para demonstrar a aplicabilidade de RBNet ao site de buscas do Diretório de Grupos de Pesquisa no Brasil, desenvolveu-se um site protótipo, tomando-se a busca textual de grupos de pesquisa. Para diferenciar este protótipo do site de buscas do Diretório, aplicação protótipa é chamada de “Recuperação Inteligente de Informações”. A primeira tela disponibiliza as opções de interação para a busca de informações ao usuário. Este site é semelhante a um site de busca genérico disponível na Internet, como já foi mencionado. A busca é feita sobre as informações do nome do grupo de pesquisa, título da linha de pesquisa e palavras-chave da linha de pesquisa. A Figura 5.2 mostra a visualização do site pelo usuário.

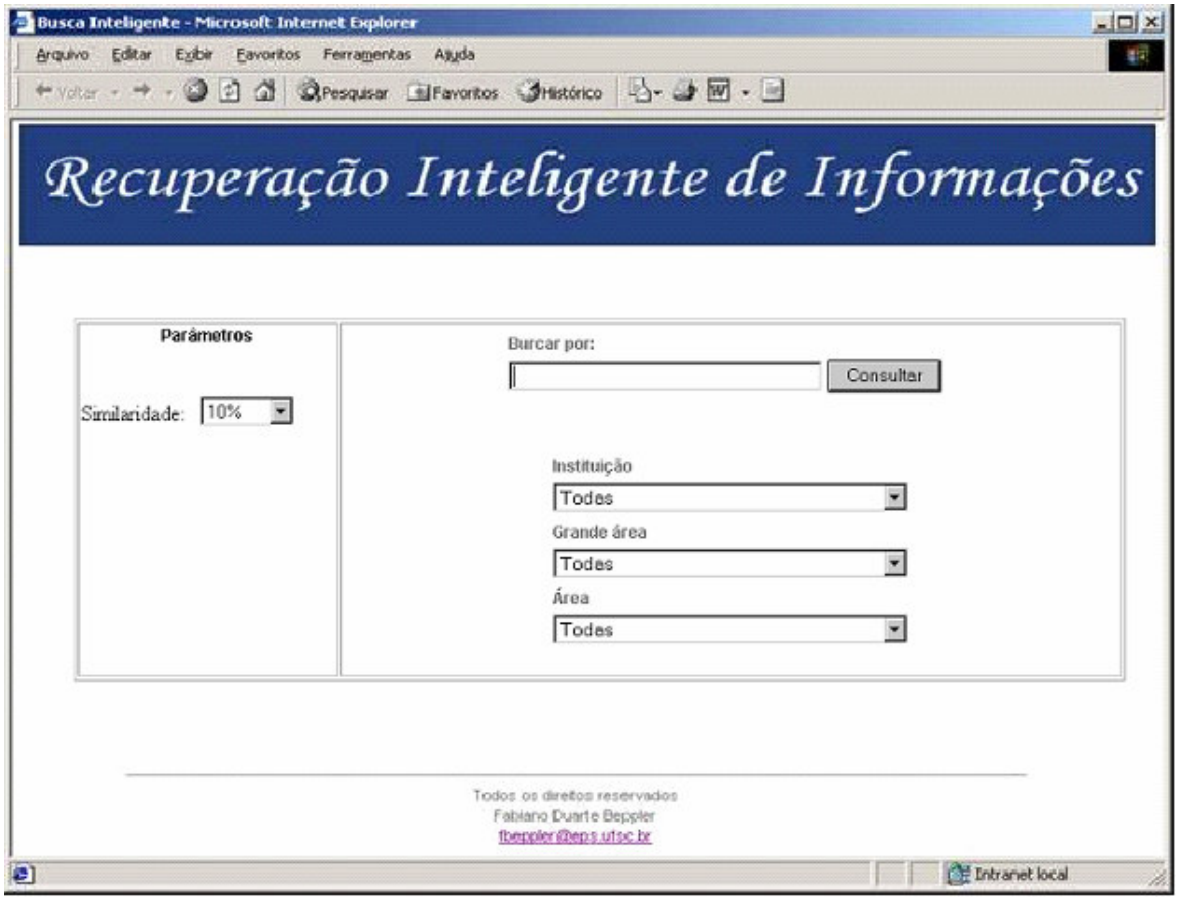


Figura 5.2 - Visualização do site para consultas

Após configurar a consulta e submeter à busca, o usuário receberá como resposta uma nova tela contendo duas colunas. A primeira coluna contém as informações das consultas similares e a segunda contém os grupos de pesquisa recuperados por meio das informações configuradas para a busca. A Figura 5.3 mostra um exemplo de tela para interação do usuário com as informações recuperadas.

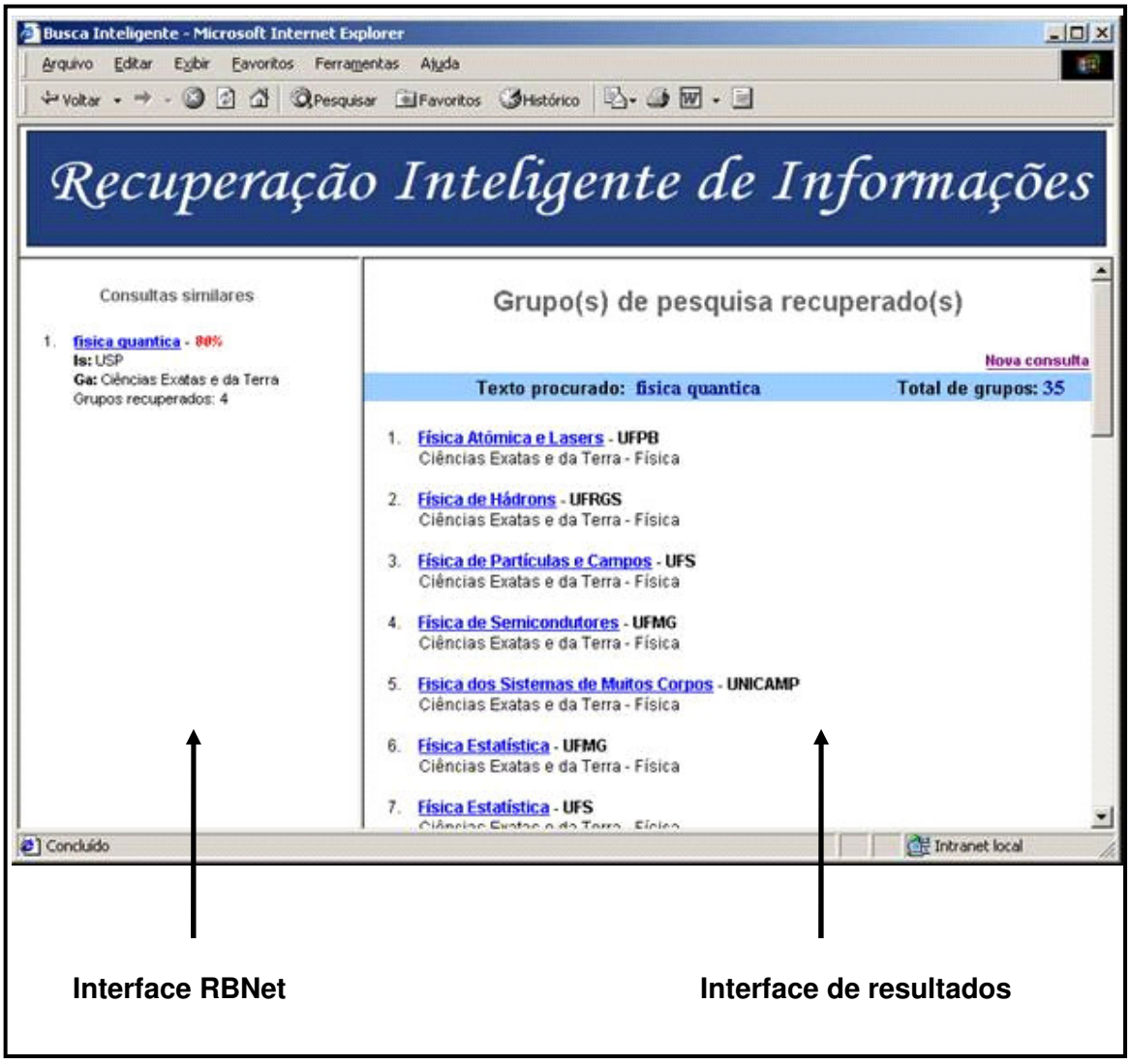


Figura 5.3 - Exemplo do site para visualização e interação com as informações recuperadas

A facilidade de interação do usuário com o site está na simplicidade. A medida em que o usuário interage, pode verificar a utilidade de poder recuperar a informação desejada, sendo auxiliado por vários outros usuários e, da mesma forma, auxiliar outros usuários que consultarem o Diretório de Grupos de Pesquisa com expressão e/ou parâmetros semelhantes.

5.6.1 Exemplo

Para exemplificar o funcionamento do protótipo, serão utilizados os seguintes parâmetros:

Expressão da busca:	Inteligência artificial
Instituição:	Todas
Grande área:	Ciências Exatas e da Terra
Área do conhecimento:	Ciência da Computação
Similaridade:	60%

A Figura 5.4 mostra como ficou o site com os parâmetros configurados.

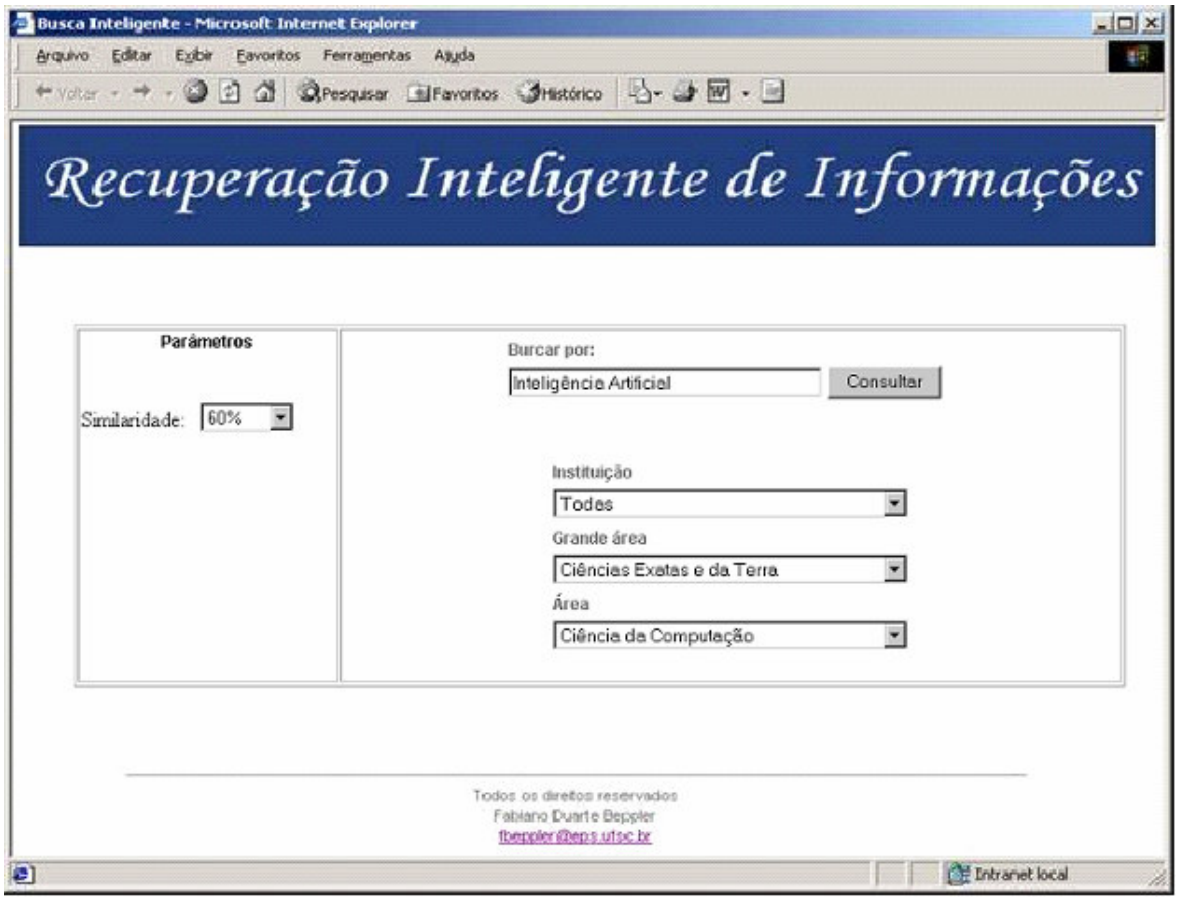


Figura 5.4 - Tela contendo os dados do exemplo postados

Após o preenchimento das informações necessárias é preciso clicar sobre o botão "*Consultar*", submetendo os dados. Os dados são então enviados ao servidor para que se execute os processos de recuperação de grupos de pesquisa e de recuperação de consultas similares. Terminada a execução dos dois processos, o resultado é apresentado ao usuário, como mostra a Figura 5.5.

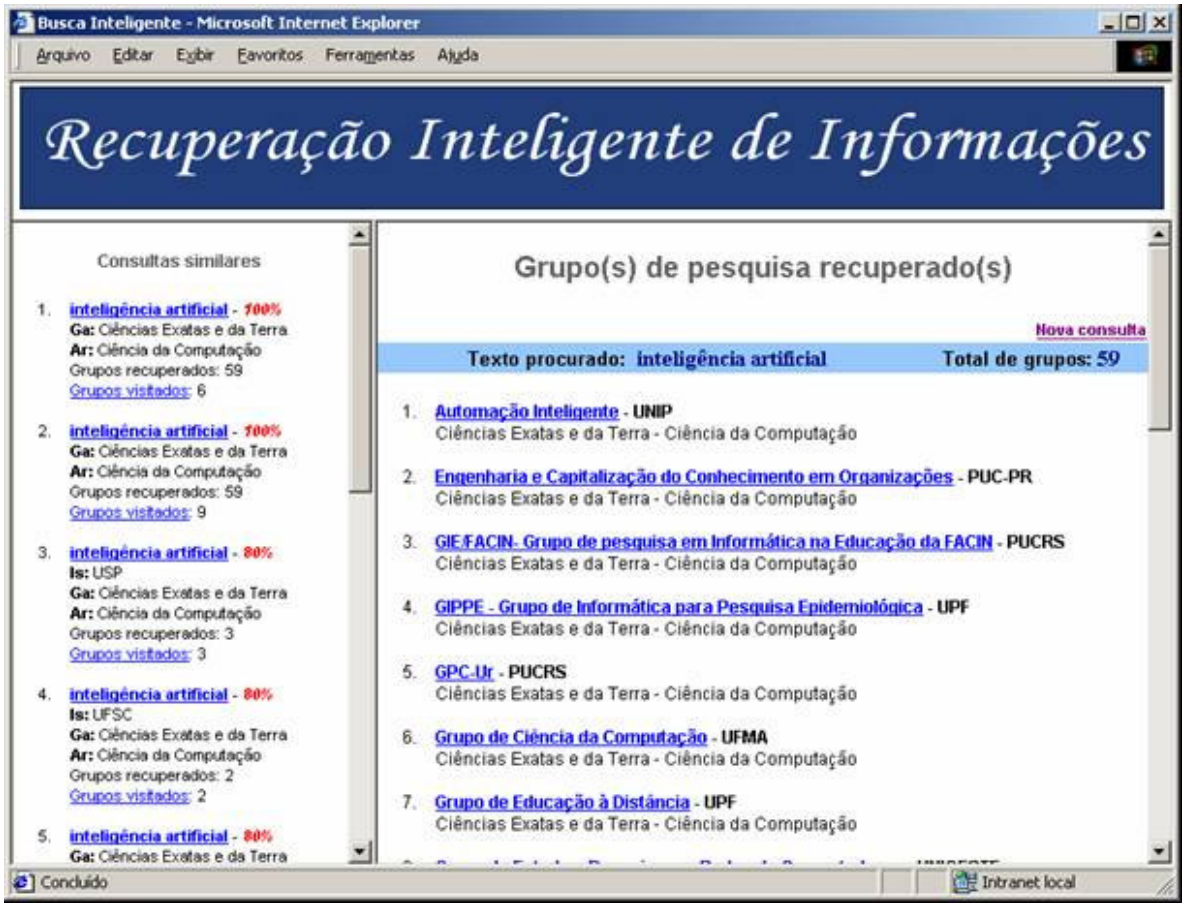


Figura 5.5 - Resultado da consulta relativo aos dados do exemplo

Como resultado da busca sobre a base de dados houve a recuperação de 59 grupos de pesquisa, mostrados na coluna da direita. Estes 59 grupos de pesquisa possuem alguma relação com os dados informados para busca. Já na coluna da esquerda, é apresentada a relação de consultas similares, com um total de 12, isto é, o mecanismo RBNet verificou que apenas 12 perfis de consultas armazenados na base de casos possuem similaridade igual ou superior a 60% com o perfil da consulta atual. As duas primeiras consultas possuem similaridade igual a 100%, significando que os parâmetros, requeridos para este cálculo de similaridade, dos perfis de consulta são os mesmos do perfil da consulta atual (busca *"inteligência artificial"* em qualquer instituição, mas na grande área *"Ciências Exatas e da Terra"* e na área *"Ciências da computação"*). A terceira e a quarta consulta recuperadas possuem grau de similaridade de apenas 80%, verificando-se que apenas a opção do parâmetro instituição é diferente em ambos os perfis de consulta (na terceira consulta se refere à instituição da USP e a quarta à instituição da UFSC).

Para exemplificar a aplicação de RBNet com o segundo cálculo do grau de similaridade, que considera a visita sobre os grupos recuperados, para o exemplo, os grupos visitados são apenas os grupos cujo nome contém a expressão "*inteligência artificial*". A Figura 5.6 mostra como ficou a tela de apresentação após visitar um grupo de pesquisa.

Figura 5.6 - Tela mostrando exemplo após visita de um grupo de pesquisa recuperado

Visitando o primeiro grupo de pesquisa recuperado (que possui a expressão "*inteligência artificial*" no nome grupo), verifica-se que a relação de consultas similares foi alterada. Antes havia 12 consultas similares, agora a quantidade de consultas foi reduzida para apenas 4, isto é, dos perfis de consulta presentes na base de casos, apenas 4 cumprem o mínimo grau de similaridade requerido, 60%. Nenhuma das consultas recuperadas possui grau de similaridade igual a 100%, como mostra a Figura 5.6, onde a consulta mais similar tem 82,2%. Quando foi clicado sobre o nome do grupo para visualização completa dos seus dados, além da chamada ao servidor requerendo os dados do grupo de pesquisa, automaticamente o perfil da consulta atual foi alterado, requerendo

uma nova recuperação de consultas de similares. Este processo, já descrito no capítulo da ferramenta RBNet, varre a base de casos com o cálculo do grau de similaridade considerando a relação dos grupos de pesquisa visitados, presente no perfil de consulta.

A primeira consulta recuperada, por exemplo, contém grau de similaridade igual a 82,2%. Observando-se as informações dos perfis de consulta nota-se que o grau de similaridade não pode ser de 100%, pois a quantidade de grupos visitados é diferente nos perfis, enquanto o perfil atual visitou apenas 1 grupo de pesquisa o perfil de consulta recuperado possui uma relação de 9 grupos de pesquisa visitados. Agora, se os próximos grupos de pesquisa visitados forem os mesmos desta consulta, o grau de similaridade irá aumentar, podendo chegar até 100%, indicando, conseqüentemente, que outro usuário fez a mesma busca.

Visitando o próximo grupo (que possui a expressão "*inteligência artificial*" no nome do grupo), verifica-se que a ordem das consultas similares foi alterada. A segunda consulta agora tem mais similaridade que a primeira, mesmo não contendo os mesmos parâmetros gerais de configuração do perfil de consulta (expressão de busca e opções dos parâmetros instituição, grade área e área), como mostra a Figura 5.7.

Figura 5.7 - Tela mostrando exemplo após visita de outro grupo de pesquisa recuperado

Essa mudança no grau de similaridade das consultas recuperadas indica o caminho que o usuário da consulta atual está tomando. Neste exemplo, ficou claro que os parâmetros preenchidos inicialmente para busca não estão tendo tanta influência se considerar o caminho que o usuário está percorrendo e a relação de consultas similares recuperadas. Este perfil dinâmico acrescido de novas informações a cada visita em grupos recuperados tende a direcionar o usuário, já que automaticamente há atualização das consultas similares recuperadas, ao caminho da verdadeira informação requerida.

Por exemplo, se um usuário estiver interessado em informações sobre *conjuntos difusos*, mas insere a expressão de busca *inteligência artificial* e só visita grupos de pesquisa que tenham alguma relação com *conjuntos difusos*, pode ser encaminhado diretamente a consultas com perfis voltados para *conjuntos difusos*. Este é mais um exemplo que reforça a capacidade de ação da ferramenta RBNet, por que o usuário não precisa saber qual é a expressão de busca exata, mas apenas navegar pelas informações recuperadas que tenham alguma relação com a informação desejada.

5.7 Resultados da aplicação

A utilização de RBNet sobre a base de dados do Diretório dos Grupos de Pesquisa demonstrou eficiência e fácil aplicabilidade. A idéia de utilizar a informação do quê os usuários estão buscando e onde estão encontrando esta informação serve como instrumento de efetivo auxílio para qualquer usuário com interesse semelhante.

Outra característica do RBNet é ampliar a visualização dos dados da base de dados, mostrando, não somente os grupos de pesquisa recuperados por sua consulta, mas também as informações de consultas similares feitas anteriormente por outros usuários. Esta visualização oferece a oportunidade de observar o tipo de informação pesquisada e como esta informação foi buscada, servindo como subsídio para novas consultas ou ainda para análises do perfil de interesse dos usuários.

A técnica de RBC utilizada para encontrar as informações de consultas similares proporciona, entre outras coisas, o grau de similaridade das consultas contidas na base de casos. Como a interação do usuário com o resultado das informações recuperadas é dinâmica, o modelo atingiu eficiência ao atualizar as informações de quais consultas são similares (pois a formação do perfil da consulta representa qual informação o usuário realmente está interessado e o quê realmente necessita).

À medida que o usuário visualiza as informações dos grupos de pesquisa recuperados, o perfil de sua consulta está sendo definido, qualificando melhor a informação desejada, e como esta informação é utilizada para recuperar consultas similares, a informação também será mais qualificada, o que pode fornecer atalhos para encontrar mais rápido a informação procurada, pois uma das opções fornecidas das consultas similares recuperadas é visualizar somente a relação de grupos que foram visitados pela consulta similar recuperada.

5.8 Considerações finais

No presente capítulo foi descrito um breve histórico sobre o Diretório dos Grupos de Pesquisa no Brasil, cujas informações foram utilizadas no protótipo. Este histórico expressa a evolução da sua construção, desde a versão 1.0 referente ao censo de 1992 até a versão 4.0 referente ao censo de 2000, por meio de números referentes a cada um dos censos efetuados.

A verificação da aplicabilidade de RBNet teve por base o protótipo, com detalhamento de cada um dos módulos componentes. A base de casos é descrita por meio do detalhamento de cada um dos atributos componentes das tabelas necessárias. Os atributos de entrada e saída definem quais informações foram consideradas em cada uma das etapas. Os cálculos do grau de similaridade, utilizados em dois momentos diferentes, foram demonstrados e exemplificados. O perfil da consulta que é armazenado na base de casos, também foi detalhado. A interação do usuário com o site é mostrada por meio da especificação das facilidades e vantagens da utilização do modelo.

Como resultado foram apresentadas as vantagens e uma aplicação do RBNet, principalmente a utilização da técnica de RBC para auxiliar a recuperação de informações. Entre os benefícios verificados destacam-se a facilidade de consulta agregada aos novos usuários (que se valem do conhecimento dos demais), o potencial aumento de visibilidade da fonte de informações do site (base nacional de grupos de pesquisa) e a formação de uma memória das interações dos usuários, com rico potencial de análises posteriores quanto ao perfil das consultas realizadas no site.

6 Conclusões e trabalhos futuros

O aumento do volume de informação na Internet, crescentes em bases diárias, exige mecanismos capazes de prover, de maneira rápida e fácil, a recuperação somente da informação desejada. Foi nesse aspecto que esta dissertação focou seu estudo, apresentando uma ferramenta inteligente para auxílio ao procedimento de recuperação de informação na Internet, o RBNet. Para estabelecer um modelo apropriado para esta tarefa, foi necessário o estudo de arquiteturas de modelos de recuperação de informação, mostrando, por meio de exemplos reais, as vantagens de cada um dos modelos apresentados.

Outro tópico abordado foi a técnica de Raciocínio Baseado em Casos (RBC), que em seus conceitos e em sua aplicabilidade apresenta vantagens para recuperar informações, segundo o preceito de uso de memória de consultas anteriores, como pressuposto em RBNet. Assim discute-se a arquitetura do RBC conceituando cada uma das fases componentes. RBC baseia-se na característica de se valer de experiências passadas para resolução de problemas futuros, conceito este, diretamente associado à proposição da ferramenta RBNet.

A ferramenta RBNet tem o intuito de auxiliar usuários a recuperarem informações a partir de conhecimento acumulado com as buscas já efetuadas por outros usuários. RBNet define o conceito de perfis de busca (parâmetros e opções de cada consulta, por exemplo, palavra ou frase digitada para a busca) e utiliza estes perfis para fornecer subsídios às buscas similares já efetuadas por outros usuários. Estes perfis são armazenados em forma de casos em uma base de casos previamente modelada, propiciando a utilização da técnica de RBC para a recuperação dos perfis de buscas semelhantes à atual (valendo-se do módulo de similaridade, previsto na arquitetura da técnica de RBC). Assim, pode-se apresentar somente consultas acima de uma determinada similaridade (previamente definida pelo usuário). Além disso, como a interação do usuário é dinâmica, RBNet propõe a construção dinâmica do seu perfil de consulta. Isto significa que, à medida que o usuário visita sites recuperados, seu perfil e a relação de consultas similares são automaticamente atualizados. Deste modo, somente aquelas que possuem maior semelhança com o perfil da consulta atual são apresentadas ao usuário. Outra vantagem da ferramenta é a velocidade de recuperar os sites das consultas similares, pelo fato da relação dos sites recuperados

estar contida no perfil da consulta. Esta vantagem em especial, atinge diretamente a utilização do banco de dados, pois não é necessário efetuar uma nova consulta, diminuindo a utilização dos mecanismos de busca sobre as informações disponíveis, e conseqüentemente aumentando a velocidade da recuperação.

Para verificar a usabilidade da ferramenta RBNet, foi desenvolvido um protótipo usando como base de dados as informações dos grupos de pesquisa disponibilizados pelo projeto desenvolvido pelo CNPq, denominado Diretório dos Grupos de Pesquisa no Brasil. Esta aplicação demonstrou a eficiência da utilização do RBNet, que além de fornecer mais subsídios para efetivação de novas buscas, forneceu uma nova forma de visualização das informações contidas na base de dados dos grupos de pesquisa, incluindo apresentação das buscas similares recuperadas. Como a atualização das consultas recuperadas é dinâmica, à medida que o usuário visita os sites recuperados, há refinamento automático das consultas recuperadas, apresentando consultas com perfis cada vez mais similares ao perfil da consulta atual. Como é possível recuperar as informações de uma consulta similar, o usuário, após digitar o termo de busca, poderá encontrar a informação desejada somente por meio de cliques (que podem ser sobre os sites recuperados pela sua consulta ou utilizar os sites recuperados pelas consultas semelhantes).

6.1 Trabalhos futuros

1. Generalização da camada de entrada (interface) do RBNet para permitir configuração automática das informações utilizadas para a similaridade.
2. Generalização das informações para a similaridade, para que parâmetros contextuais ao domínio do problema possam ser inclusos de forma configurável.
3. Implementar um módulo de esquecimento que irá interagir sobre a base de casos, atuando sobre os perfis de consultas que não foram utilizadas durante um período previamente determinado. Este procedimento irá influenciar diretamente a velocidade da recuperação de consultas similares sobre a base de casos, pois o procedimento de recuperação analisa cada caso contido na base de casos.

4. Pode-se acrescentar ao RBNet métodos de CRM (*Customer Relationship Management*), tendo como base os casos de consulta e suas semelhanças. Análises como nível de profundidade da consulta comparada com seus parâmetros pode indicar dificuldade de uso de determinados usuários à totalidade da base.
5. Para ampliar o escopo de recuperação de casos de consulta, é recomendável considerar a utilização de *thesaurus* temáticos ao domínio do site. Assim, por exemplo, a consulta pelo termo "redes neuronais" pode incluir nos casos recuperados consultas a "redes conexionistas". Para tal, RBNet deve incluir *thesaurus*, adicionar ao cálculo da similaridade parâmetros utilizados pelo usuário e seus respectivos sinônimos (com pesos para reduzir similaridade).
6. Transformar a ferramenta RBNet em um serviço web via metodologia *Web Service*.

7 Bibliografia

AAMODT, A.; PLAZA, E. **Case-based reasoning: foundational issues, methodological variations and system approaches.** Artificial Intelligence Communications, Vol. 7, 1994;

ATKINSON, John A. **Text mining: principles and applications.** Revista Facultad de Engenieria, volume 7, Chile, 2000. Disponível em: http://www.electa.uta.cl/~revista/2000/mining-ja.pdf. Acesso em: abril de 2002;

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Mordern information retrieval.** New York, Addison-Wesley, 1999;

BR-BUSINESS. **Os cookies.** Disponível em: http://www.br-business.com.br/brb/cookies.htm. Acesso em: dezembro de 2001;

CNPq. **Help do sistema instituição.** Disponível em: http://www.cnpq.br/diretorio4. Acesso em: maio de 2000;

CNPq. **Diretório dos grupos de pesquisa no Brasil versão 4.0.** Disponível em: http://www.cnpq.br/plataformalattes/dgp/versao4/informacoes_gerais/index.html Acesso em: dezembro de 2001;

CNPq. **Plataforma de currículo Lattes.** Disponível em: http://www.cnpq.br/plataformalattes/curriculo/index.html. Acesso em: junho de 2001a;

DAVEMPORT, Thomas H.; PRUSAC, Laurence. **Ecologia da informação : por que só a tecnologia não basta para o sucesso na era da informação.** São Paulo: Futura, 1998;

DIXON, Mark. **An overview of document mining technology.** Disponível em: http://citeseer.nj.nec.com/rd/26840091%2C91%2C1%2C0.25%2CDownload/http%253A%252F%252Fciteseer.nj.nec.com/cache/papers/cs/6563/http%253AzSzzSzwww.geocities.comzSz%257EmjdixonzSzmarkzSzwritingszSzdixm97_dm.pdf/dixon97overview.pdf. Acesso em: abril de 2002;

DORRE, Jochen; GERSTL, Peter; SEIFFERT, Roland. **Text mining: finding**

nuggets in mountains of textual data. Disponível em: http://maya.cs.depaul.edu/~classes/ect584/papers/dorre.pdf. Acesso em: abril de 2002;

FILHO, Jagurta Lisboa. **Estruturação e modelagem de bancos de dados.** Disponível em: http://www.dpi.ufv.br/~jugurta/bd/estmodbd-gisbr2001.pdf. Acesso em: abril de 2002;

FILHOa, Jayme Teixeira. **Gerenciando conhecimento.** Livro disponível na Internet. Disponível em: http://www.gerenciandoconhecimento.com.br/frames.htm. Acesso em: abril de 2002;

FUHR, Norbert. **Models for integrated information retrieval and database systems.** Disponível em: http://citeseer.nj.nec.com/35392.html. Acesso em: junho de 2001;

GLORIA 10. **Application Service Providers.** Disponível em: http://www.enfase.com/gloria10/asp.htm. Acesso em: dezembro de 2001;

GUIMARÃES, Reinaldo; GALVÃO, Gerson; COSAC, Silvana M.; LOURENÇO, Ricardo S.; *et al*. **A pesquisa no Brasil e hierarquização dos grupos de pesquisa a partir dos dados do Diretório dos Grupos de Pesquisa no Brasil.** CNPq, 1999;

GUDIVADA, Venkat N.; RAGHAVAN, Vijay V.; GROSKY,Willian I.; KASANAGOTTU, Rajesh. **Information retrieval on the world wide web.** IEEE, 1997. Disponível em: http://www.computer.org/intelligente/. Acesso em: maio de 2001;

HEWLETT-PACKARD. **American company, founded in 1939.** Palo Alto, United States of America. Disponível em: http://pawnt139.external.hp.com/servlet/Setec?product=LaserJet5si. Acesso em: junho de 2001;

HORSTMANN, Cay S., CORNELL, Gary. **Core Java 2. Volume I – Fundamentos.** São Paulo, Makron Books, 2001;

HUANG, Lan. **A Survey on Web Information Retrieval Technologies.** New York, University of New York, 1999;

IDG, Revista eletrônica IDG NOW. Disponível em: http://idgnow.terra.com.br/idgnow/carreira/2002/01/0004. Acesso em: abril de 2002;

KATZ, Boris. **From sentence processing to information access on the world wide web.** Massachusetts Institute of Technology. Disponível em: http://www.ai.mit.edu/people/boris/webaccess/. Acesso em: abril de 2002;

KOLODNER, Janet. **Case-based reasoning.** Los Altos, Morgan Kaufmann, 1993;

KORFHAGE, Robert R.. **Information Storage and Retrieval.** New York, Wiley Computer Publishing, 1997;

KURAMOTO, Hélio. **Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais.** Disponível em: http://www.ibict.br/cionline/250296/25029605.pdf. Acesso em: outubro de 2001;

LAGEMANN, Gerson Volney. **RBC para o problema de suporte ao cliente nas empresas de prestação de serviço de software: o caso Datasul.** Dissertação de mestrado, UFSC, 1998;

LEAKE, David. **Case-Based Reasoning: Experiences, Lessons e Future Directions.** California, AAAI Press/The MIT Press, 1996;

LENZ, Mario. **Textual CBR and information retrieval – A comparison.** University Berlin. Disponível em: http://citeseer.nj.nec.com/lenz98textual.html. Acesso em: junho de 2001;

LETSCHE, Todd A.; BERRY, Michel W. **Large-Scala Information Retrieval with Latent Semantic Indexing.** University of Tennessee, 1996. Disponível em: http://www.cs.utk.edu/~berry/lsi++/. Acesso em: abril de 2001;

MARINILLI, Mauro; MICARELLI, Alessandro; ESCIARRONE, Filippo. **A case-based approach to adaptative information filtering for the WWW.** Universtità di Roma, Itália. Disponível em: http://www.contrib.andrew.cmu.edu/~plb/WWWUM99_workshop/marinilli/marinilli.html. Acesso em: junho de 2001;

MELCHIORS, Cristina. **Raciocínio baseado em casos aplicado ao**

gerenciamento de falhas em redes de computadores. Dissertação de mestrado, Porto Alegre – UFRGS, 1999;

MILLER, Ethan; et al. **Performance and scalability of a large-scale N-gram basead information retrieval system.** Disponível em: http://jodi.ecs.soton.ac.uk/Articles/v01/i05/Miller/miller2.pdf. Acesso em: abril de 2002;

OLIVEIRA, Fabio Abreu Dias de; NAVAUX, Philippe Olivier Alexandre. **Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa.** UFRGS. Disponível em: http://www.inf.ufrgs.br/procpar/disc/cmp135/trabs/992/Parser/parser.html#_Toc470452819. Acesso em: junho de 2001.

PERES, Sarajane M. **Raciocínio Baseado em Casos para Avaliação de Planos de Rotas.** Dissertação, UFSC, 1999;

RAMIREZ, Carlos; COOLEY, Roger. **Case-based reasoning model applied to information retrieval.** University of Kent at Canterbury. Disponível em: http://www.cs.ukc.ac.uk/people/staff/rec/pubs.html. Acesso em: junho de 2001;

RIBEIRO, Rita A.; MOREIRA, Ana M.. **Intelligent query model for business characteristics.** Disponível em: http://falcon.cc.ukans.edu/~teska/FUZZY.html. Acesso em: abril de 2002;

RIJSBERGEN, C. J. van. **Information retrieval.** Disponível em: http://www.dcs.gla.ac.uk/Keith/Preface.html. Acesso em: março de 2001;

RIJSBERGEN, C. J. van. **Information retrieval.** Disponível em: http://www.dcs.gla.ac.uk/~iain/keith/. Acesso em: junho de 2001a;

SALTON, G., MCGILL, M. J.. **Introduction do modern information retrieval.** New York, McGraw Hill, 1983;

SCHANK, R. **Dynamic memory: a theory of reminding and learning in computers and people.** Cambridge University Press, 1982;

SMEATON, A. **Information retrieval: still butting heads with natural language**

processing? Disponível em: http://lorca.compapp.dcu.ie/~asmeaton/pubs-list.html. Acesso em: junho de 2001;

TAKEMURA, Roberto Y.. **Controle inteligente: lógica difusa.** Disponível em: http://www.din.uem.br/ia/controle/fuz_prin.htm. Acesso em: maio 2001;

TKCH, Daniel. **Turning information in knowledge.** IBM software and solutions. Disponível em: http://www.us.ibm.com. Acesso em: outubro de 2001;

VENERUCHI, Edilene A., LIMA, José V.. **Linguagens de consulta e recuperação de documentos hipermídia no ambiente web.** Disponível em: http://www.inf.ufrgs.br/pos/SemanaAcademica/Semana99/edilene/edilene.html. Acesso em: janeiro de 2001;

VOLLRATH, Ivo; WILK, Wolfgang; BERGMANN, Ralph. **Case-based reasoning support for online catalog sales.** IEEE, 1998. Disponível em: http://computer.org/internet. Acesso em: maio de 2001;

WANGENHEIM, Chistiane G. von. **Case-based reasoning – a short introduction.** I-Konow, UFSC, 2000;

WATSON, Ian. **Appying case-based reasoning: tecniques for enterprise systems.** San Francisco: Morgan Kaufmann, 1997;

WATSON, Ian. **CBR is a methodology not a technology.** University of Salford – UK, 2000. Disponível em www.ai-cbr.org. Acesso em: agosto de 2000;

WATSON, Ian. **A case-based reasoning application for engineering Sales support using introspective reasoning.** University of Auckland, New Zealand. Disponível em: http://www.aaai.org. Acesso em: junho de 2001;

WEBBER - LEE, Rosina. **Raciocínio baseado em casos.** 1996. Disponível em: http://www.eps.ufsc.br/teses98/rosina/index.html. Acesso em: abril de 2000;

WEBBER – LEE, Rosina. **Intelligent jurisprudence research.** Florianópolis: UFSC, 1998 (Tese de doutorado em Engenharia de Produção);

WIVES, Lendro Krug; LOH, Stanley. Hiperdictionary: a knowledge discovery tool to help information retrieval. Porto Alegre, UFRGS, 2000;

WOHL, Amy D.. **Intelligent text mining. Create business intelligence.** Disponível em: http://www-

4.ibm.com/software/data/iminer/fortext/download/amipap.html. Acesso em: outubro de 2001;

YUWONO, Budy; LAM, Sávio L. Y.; YING, Jerry H.; LEE, Dik L.. **A world wide web resource discovery system.** Disponível em: http://www.cs.ust.hk/cgi-bin/IndexServer/. Acesso em: novembro de 2001;

ZAKON, Robert Hobbes. **Hobbes Internet Timeline.** Disponível em: http://www.zakon.org/robert/internet/timeline/. Acesso em: março de 2002;

ZADEH, Loft A. **Fuzzy Sets**. Information and Control. 1965;

ZANASI, Alessandro. **Web mining through the online analyst.** Disponível em: http://kmtools.crie.ufrj.br/km/ferramentas/artigos/web_mining.pdf. Acesso em: outubro de 2001;