Universidade de Brasília
Departamento de Ciência da Informação e Documentação
Programa de Pós-Graduação em Ciência da Informação

SYMBALL RUFINO DE OLIVEIRA

RECUPERAÇÃO INTELIGENTE DE JURISPRUDÊNCIA: Uma Avaliação do Raciocínio Baseado em Casos Aplicado a Recuperação de Jurisprudências no Tribunal Regional Eleitoral do Distrito Federal

Brasília – DF
2008

SYMBALL RUFINO DE OLIVEIRA

RECUPERAÇÃO INTELIGENTE DE JURISPRUDÊNCIA: Uma Avaliação do Raciocínio Baseado em Casos Aplicado a Recuperação de Jurisprudências no Tribunal Regional Eleitoral do Distrito Federal

Dissertação apresentada à banca examinadora como requisito parcial à obtenção de Título de Mestre em Ciências da Informação pelo Programa de Pós-Graduação em Ciências da Informação do Departamento de Ciência da Informação e Documentação da Universidade de Brasília.
Orientadora: Profª. Drª. Marisa Bräscher Basílio Medeiros

Brasília – DF
2008

Ficha Catalográfica

Oliveira, Symball Rufino de

Recuperação Inteligente de Jurisprudência: Uma Avaliação do Raciocínio Baseado em Casos Aplicado a Recuperação de Jurisprudências no Tribunal Regional Eleitoral do Distrito Federal / Symball Rufino de Oliveira. -- Brasília: UnB / Departamento de Ciência da Informação e Documentação, 2009.

xv, 126 f.: il

Orientadora: Marisa Bräscher Basílio Medeiros

Dissertação (mestrado) – Universidade de Brasília / Programa de Pós-Graduação em Ciências da Informação, 2009.

1. Recuperação da Informação 2. Raciocínio Baseado em Casos 3. Avaliação de Sistemas de Recuperação da Informação 4. Cálculo de Similaridade 5. Jurisprudência Eleitoral I. Medeiros, Marisa Bräscher Basílio. II. Universidade de Brasília, Programa de Pós-Graduação em Ciência da Informação. III. Título.

FOLHA DE APROVAÇÃO

Título: **RECUPERAÇÃO INTELIGENTE DE JURISPRUDÊNCIA:** *Uma Avaliação do Raciocínio Baseado em Casos Aplicado a Recuperação de Jurisprudências no Tribunal Regional Eleitoral do Distrito Federal*

Autor: **Symball Rufino de Oliveira**

Área de Concentração: **Transferência da Informação**

Linha de Pesquisa: **Arquitetura da Informação**

Dissertação submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação do Departamento de Ciência da Informação e Documentação da Universidade de Brasília como requisito parcial para obtenção do titulo de **Mestre em Ciência da Informação**.

Dissertação aprovada em: 18/02/2009

Aprovada por:

Profa Dra Marisa Bräscher Basílio Medeiros
Presidente – (UnB/PPGCInf)

Prof. Dr. Jaime Robredo
Membro Interno – (UnB/PPGCInf)

Prof. Dr. Marcelo Grangeiro Quirino
Membro Externo – (UCB)

Prof. Dr. Rogério Henrique de Araújo Júnior
Suplente – (UnB/PPGCInf)

Esta obra é dedicada

À minha mãe, meu maior exemplo.

À minha irmã Zâmbia, por me ensinar o caminho
de casa, dos estudos, da virtude.

Ao meu irmão Ramsés que me permitiu ficar com um
pouquinho da sua inteligência.

Às minhas irmãs Núbia e Loanda, pelo amor, carinho e amizade.

À Ana Lúcia Fernandes, minha esposa, meu amor, minha amiga e maior
incentivadora.

Aos meus filhos Pedro, Alexandre e Vinícius, por me ensinarem o significado
do amor.

Ao meu pai, Prof. Dr. Décio Rufino de Oliveira, em memória, por ter me ensinado
a conhecer um homem olhando para o que está além dele.

AGRADECIMENTO

Ao Grande Arquiteto do Universo que é Deus, por me permitir chegar até aqui.

À Profª. Drª. Marisa Bräscher Basílio Medeiros, por orientar meus passos no caminho da pesquisa científica durante a realização desse trabalho.

Ao Tribunal Regional Eleitoral do Distrito Federal por ter servidor como base para minha pesquisa.

Aos queridos colegas do Tribunal Regional Eleitoral do Distrito Federal da Secretaria de Tecnologia da Informação e da Secretaria Judiciária pelo apoio incondicional na realização desse trabalho.

“Se o homem não sabe a que porto se dirige, nenhum vento lhe será favorável”.
Lucius Annaeus Sêneca

RESUMO

Trata-se de uma pesquisa cujo objeto é avaliar a medida de precisão de um sistema de recuperação de informação jurídica (jurisprudência) que utiliza técnica de inteligência artificial conhecida como Raciocínio Baseado em Casos (RBC). Nesse modelo as jurisprudências são organizadas sob a forma de casos jurídicos concretos. O raciocínio baseado em casos tem como princípio a idéias de que um caso jurídico passado pode ser útil para resolver um problema atual, desde que exista entre eles algum grau de semelhança. Para estabelecer semelhanças entre casos atuais e passados o modelo estudado propõe o uso de cálculo de similaridade que é realizado com base na comparação de índices temáticos obtidos a partir do processo de indexação realizado por especialistas utilizando-se como apoio um tesauro jurídico. Esta pesquisa utiliza como universo as jurisprudências produzidas pelo Tribunal Regional Eleitoral do Distrito Federal. A amostra foi composta, considerando o recorte dado à pesquisa, por jurisprudências eleitorais produzidas nas eleições gerais de 2006 no Distrito Federal. Para realizar a avaliação do modelo, foi construído um protótipo do sistema de recuperação de informação baseado em casos. Em seguida, avaliou-se o protótipo quanto ao grau de precisão obtido no resultado de um conjunto de buscas. O método adotado para as avaliações foi o mesmo utilizado na Text REtrieval Conference (TREC) de 2007, tarefa Legal Track. Após a coleta dos dados foi elaborado um relatório discutindo a possibilidade do sistema de recuperação de informação baseado em casos ser considerado um paradigma para a recuperação de informação jurídica eleitoral.

Palavras-chave: Recuperação da Informação, Raciocínio Baseado em Casos, Avaliação de Sistemas de Recuperação da Informação, Jurisprudência Eleitoral, Cálculo de Similaridade.

ABSTRACT

This is a research whose object is to evaluate a legal information retrieval system precision. This IR system is based on a model that uses artificial intelligence technique known as Case-Based Reasoning (CBR). In this model the jurisprudences are organized in the form of actual legal cases. The principle of CBR is that a past legal case can be useful to solve a current problem, since there is between them some degree of similarity. To establish similarities between current and past cases the model studied proposes the use of the similarity calculation performed based on comparison of thematic indices. The process of indexing is performed by experts using a thesaurus as a legal support. This research uses jurisprudences produced by the Regional Electoral Tribunal of the Distrito Federal. The sample was composed considering electoral jurisprudence produced in general elections of Distrito Federal, in the year of 2006. To perform the evaluation of the model, a prototype of a case-based information retrieval system was built. Then the prototype precision degree was evaluated from the result of a set of queries submitted to it. The method adopted for the evaluation was the same used in the Text REtrieval Conference (TREC) in 2007 by Legal Track Task. After the data collecting, a report was made to discuss the possibility of the case-based information retrieval system can be considered a paradigm for the legal information retrieval.

Key-words: Information Retrieval, Case-Based Reasoning, Information Retrieval System Evaluation, Electoral Jurisprudence, Similarity Calculation.

LISTA DE FIGURAS

Figura 1 – Arquitetura da gestão eletrônica de jurisprudência do Tribunal Regional Eleitoral do Distrito Federal 26
Figura 2 – Tela de cadastramento de processo no Sistema de Acompanhamento de Documentos e Processos 27
Figura 3 – Tela de cadastro no Inteiro Teor de Acórdãos e Resoluções 28
Figura 4 – Tela de consulta do Sistema de Jurisprudências 29
Figura 5 – Tela de cadastro do Sistema de Jurisprudências 30
Figura 6 – Tela de consulta de jurisprudência no sitio do Tribunal Regional Eleitoral do Distrito Federal 32
Figura 7 – Tela de consulta pelo campo busca livre 33
Figura 8 – Exemplo solução de um problema atual com base em um caso passado 36
Figura 9 – Modelo básico do enfoque Raciocínio Baseado em Casos 37
Figura 10 – Ciclo do raciocínio baseado em casos 38
Figura 11 – Processo de recuperação da informação 43
Figura 12 – Taxonomia dos modelos de recuperação da informação 44
Figura 13 – Processo geral de recuperação da informação 58
Figura 14 – Representação gráfica da revocação 62
Figura 15 – Gráfico comparativo entre precisão e revocação 63
Figura 16 – Esquema de medida de efetividade dos resultados 67
Figura 17 – Representação esquemática do processo de recuperação de informação 68
Figura 18 – Exemplo de representação de um tópico 71
Figura 19 – Organograma da Secretaria Judiciária 82
Figura 20 – Motor de Conhecimento baseado em casos para Recuperação de Acórdãos 84
Figura 21 – Processamento da pesquisa textual 90
Figura 22 – Exemplo de distância de vizinho-mais-próximo 94
Figura 23 – Gráfico do resultado da precisão para o especialista 1 103
Figura 24 – Gráfico do resultado da precisão para o especialista 2 104
Figura 25 – Gráfico do resultado da precisão para o especialista 3 105
Figura 26 – Gráfico do resultado da precisão para o especialista 4 106
Figura 27 – Gráfico do resultado da precisão para o especialista 5 107
Figura 28 – Gráfico do resultado consolidado da precisão 108

LISTA DE QUADROS

Quadro 1 – Combinação das expressões de relevância de Saracevic 53
Quadro 2 – Relação de servidores participantes 82
Quadro 3 – Pesos dos pontos de acesso 86
Quadro 4 – Atributos da tabela de acórdão 88
Quadro 5 – Exemplo de caso de entrada 89
Quadro 6 – Gabarito de metas de recuperação 91
Quadro 7 – Exemplo de cálculo de similaridade local 93
Quadro 8 – Exemplo de cálculo da média principal de precisão 100
Quadro 9 – Exemplo de cálculo da média principal de precisão 101

LISTA DE SIGLAS

CBR – Case-Based Reasoning - Raciocínio Baseado em Casos

CI – Ciência da Informação

CIA – Central Intelligence Agency – Agência Central de Inteligência

CORPJ – Coordenadoria de Registros Partidários e Jurisprudência Eleitoral

DARPA – Defense Advanced Research Projects Agency – Agência de Defesa para Projeto de Pesquisa Avançado

EI – Entidade da Informação

EUA – Estados Unidos da América

FRBR – Functional Requirements for Bibliographic Records - Requisitos Funcionais para Registros Bibliográficos

GIF – Graphics Interchange Format

HC – Habeas Corpus

IA – Inteligência Artificial

IBM – International Business Machines

ICMC - Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo

ITAR – Inteiro Teor de Acórdãos e Resoluções

MARC – MAchine Readable Cataloging - Catalogação Legível por Computador

MOP – Memory Organization Packages - Pacotes de Organização de Memória

MPP – Média Principal da Precisão

NI – Necessidade de Informação

NILC – Núcleo Interinstitucional de Lingüística Computacional

NIST – National Institute of Standards and Technology – Instituto Nacional de Padrões e Tecnologia

OPAC – Online Public Access Catalog – Catálogo em Linha de Acesso Público

PHP – Personal Home Pages

PREC-R – Precisão-R

RBC – Raciocínio Baseado em Casos

RDF – Resource Description Framework – Linguagem de Descrição de Conteúdos Web

RI – Recuperação de Informação

SADP – Sistema de Acompanhamento de Documentos e Processos

SGML – Standard Generalized Markup Language – Meta Linguagem Padrão de Marcação

SJU – Secretaria Judiciária

SJUR – Sistema de Jurisprudência

SRI – Sistema de Recuperação de Informação

STAIRS – STorage And Information Retrieval System – Sistema de Armazenamento e Recuperação de Informação

TDDE – Tipo De DEcisão

TDEC – Tabela de dados estatísticos consolidados

TIPSTER – Termo utilizado para descrever pessoa que provê informação estratégica

TREC – Text REtrieval Conference - Conferência de recuperação de texto

TRE-DF – Tribunal Regional Eleitoral do Distrito Federal

TRF1 – Tribunal Regional Federal da 1ª Região

TSE – Tribunal Superior Eleitoral

UFRGS – Universidade Federal do Rio Grande do Sul

UFSC – Universidade Federal de Santa Catarina

UNIVALI – Universidade do Vale do Itajaí

US – United States – Estados Unidos

SUMÁRIO

INTRODUÇÃO 16
1. Formulação do problema 18
2. Objetivo da Pesquisa 19
2.1 Objetivo Geral 20
2.2 Objetivos Específicos 20
3. Justificativa 21
4. Estrutura da Dissertação 21
REFERENCIAL TEÓRICO 22
1. Jurisprudência Eleitoral 22
1.1. Metodologia utilizada para análise e indexação de jurisprudências 23
1.2 Comentários ao Acórdão 24
1.3 Sistemas de recuperação de jurisprudência eleitoral 25
1.3.1 Arquitetura de software de recuperação de jurisprudência 25
1.3.2 Etapas do processamento de Acórdãos 26
1.3.3 Recuperação de jurisprudência eleitoral 30
2. O Raciocínio Baseado em Casos 34
2.1 Histórico 34
2.2 Definição 35
2.3 Elementos do raciocínio baseado em casos 36
2.4 O Ciclo do raciocínio baseado em casos 37
2.5 O raciocínio baseado em casos aplicado à recuperação de informação Jurídica 39
3. Sistemas de Recuperação da Informação 40
3.1 Evolução histórica 41
3.2 Os modelos de sistemas de recuperação da informação 42
3.2.1 Modelo Booleano 45
3.2.2 Modelo Dinâmico 47
4. Avaliação de Sistemas de recuperação de informação 50
4.1. Histórico e estudos de laboratório 51
4.2 Avaliação 52
4.3 Relevância 52
4.3.1 Relevância lógica 54
4.3.2 Relevância Situacional 55

4.4 Paradigmas da avaliação de sistemas de recuperação de informação 55
4.5 Critérios de avaliação 57
4.6 Medidas de avaliação 60
4.6.1 Revocação (Recall) e precisão (precision) 61
4.6.2 Técnicas de recuperação de informação 63
4.6.3 Problemas da revocação e precisão 64
4.6.4 Outras medidas de desempenho do sistema 65
4.7. Coleções de referência 65
4.7.1 Coleção Storage and Information Retrieval System 66
4.7.2 Coleção Text Retrieval Conference 68
4.7.2.1 Conjunto de documentos 71
4.7.2.2 Conjunto de necessidade de informação 71
4.7.2.3 Julgamento de relevância 72
4.7.2.4 Avaliação do resultado da busca 72
4.7.2.5 Legal Track 73
4.7.2.6 Críticas aos métodos de avaliação utilizados na Text Retrieval Conference 74
4.7.3 Iniciativas de criação de coleções de teste no Brasil 75
4.7.3.1 Coleção .gov.br 75
4.7.3.2 Corpus Yes, user! 76
MÉTODOS 79
1. Fundamentos 79
2. Método Aplicado à Pesquisa 79
2.1 Fases da pesquisa 81
2.2 Universo 81
2.3 Especialistas participantes 81
2.4 Necessidades de informação 82
2.5 Amostra 83
2.6 Instrumento 83
2.6.1 Processo de construção do motor recuperação baseado casos 84
2.6.2 Tesauro jurídico eleitoral 96
2.7 Coleta de dados 98
2.8 Análise de dados 98
RESULTADOS 103
1. Especialista 1 103

2. Especialista 2 104
3. Especialista 3 105
4. Especialista 4 105
5. Especialista 5 106
6. Resultado consolidado 107
CONCLUSÃO 109
REFERÊNCIAS 112
APÊNDICE A – Resultado do Especialista 1 120
APÊNDICE B – Resultado do Especialista 1 Tabelado 122
APÊNDICE C – Resultado do Especialista 2 123
APÊNDICE D – Resultado do Especialista 2 Tabelado 125
APÊNDICE E – Resultado do Especialista 3 126
APÊNDICE F – Resultado do Especialista 3 Tabulado 128
APÊNDICE G – Resultado do Especialista 4 129
APÊNDICE H – Resultado do Especialista 4 Tabelado 131
APÊNDICE I – Resultado do Especialista 5 132
APÊNDICE J – Resultado do Especialista 5 Tabulado 134
APÊNDICE K – Resultado Consolidado Tabulado 135
ANEXO A – Formulário de indexação: Modelo TSE 136
ANEXO B – Exemplo de Acórdão 137
ANEXO C – Formato padrão de um tópico 138
ANEXO D – Instrumento de coleta de dados 139
ANEXO E – Formulário de indexação: Modelo de RBC 140
ANEXO F – Tabela de dados estatísticos consolidados (TDEC) 141

INTRODUÇÃO

A busca por novos modelos de organização e recuperação de informação tem sido um dos maiores esforços dos pesquisadores da ciência da informação (CI). Nesse sentido essa dissertação alinha-se com estudos atuais porque desenvolve uma pesquisa de aplicação e avaliação de um modelo de recuperação de informação jurídica que utiliza recursos da inteligência artificial com o objetivo de melhorar o índice de precisão obtido no resultado de busca por informação jurídica.

A preocupação científica de desenvolver soluções para o problema evidenciado pela rápida expansão do volume de informações ocorrida no pós-guerra dos anos 50 deu origem ao termo recuperação da informação (RI) cunhado por Mooers (1951). A partir daí, bibliotecas e centros de informação passaram a se preocupar em elaborar técnicas de organização da informação a fim de atender com maior rapidez e precisão às necessidades de informação (NI) expressas sob a forma de consulta por seus usuários e utilizada na busca por informação. Dentre as soluções utilizadas, Svennonius (2000) descreve os diferentes padrões de catalogação utilizados para organizar e descrever as unidades informacionais, tais como Dubblin Core, Marc, FRBR, SGML, RDF Schema, entre outros.

Um padrão de catalogação permite acesso a um item individual dentro de uma coleção de unidades informacionais (TAYLOR, 2004). Para tanto, utiliza-se campos de registros de informação, tais como: autor, título, assunto, que funcionam como ponto de acesso às unidades de informação.

Utilizando-se das tecnologias da informação, mecanismos automatizados de recuperação da informação foram desenvolvidos visando facilitar o processo de busca por informações em grandes coleções organizadas em bases de dados eletrônicas. Esses mecanismos são baseados nos princípios da organização da informação a fim de permitir descrição física, a catalogação e descrição de conteúdo dos documentos, a indexação, a classificação e a elaboração de resumos. O modelo clássico de recuperação da informação utiliza estratégia de comparação entre palavras de entrada (representam a necessidade de informação do usuário) e os campos de registros de informação definidos como ponto de acesso a um determinado documento.

Os pontos de acesso são elementos de representação da informação, resultado do processo de sua organização. Existem pontos de acesso relativos aos aspectos físicos do documento (autor, título, editor, etc.) e aqueles relativos aos aspectos de conteúdo (termos de indexação, número de classificação, resumo, etc.). O procedimento de definição dos campos

de registros de informação que representariam os pontos de acesso ao documento realizado por especialista no assunto abordado recebeu o nome de indexação (TAYLOR, 2004).

Com intuito de avaliar a qualidade do processo de indexação utilizado em uma coleção, diversos métodos foram desenvolvidos. Entre os principais destacamos estudo de avaliação de quatro sistemas de indexação utilizados para recuperação de Informação que foi realizado na faculdade de aeronáutica de Cranfield, na Inglaterra. Desenvolvido no ano de 1960, esse trabalho tinha como objetivo avaliar o grau de precisão, definido como a relação entre os documentos relevantes e os documentos recuperados em um resultado de busca (LANCASTER e FAYEN, 1973), e o grau de revocação, entendido como a taxa referente à relação entre os documentos relevantes obtidos no resultado e o total de documentos relevante da coleção (CLEVERDON, 1964), obtidos a partir da busca textual. A principal contribuição do trabalho de Cleverdon (1964) foi a definição de um método de avaliação baseado em medidas de desempenho de um sistema de indexação na recuperação da informação. Swanson (1960, appud MARON e BLAIR, 1985) e Salton (1970, appud MARON e BLAIR, 1985), utilizando os métodos de avaliação de Cleverdon(1964), realizam estudos de avaliação do resultado da busca em texto completo por meio do uso de computador, obtendo resultados muito animadores, mas utilizam um conjunto muito pequeno de documentos em texto completo. As pesquisas de Maron e Blair (1985) utilizando base de dados de texto completo sucederam alguns estudos como os realizados Swanson (1960) e Salton (1970). Esse estudo representou o primeiro experimento de recuperação de informação em base de dados textuais de larga escala. As pesquisas de Maron e Blair (1995) concluíram que o método de recuperação de informação baseado na indexação e busca por palavras-chave obteve o índice máximo de precisão de 25%, ou seja, para cada busca somente 25% dos documentos relacionados no resultado foram considerados relevantes.

Essa descoberta torna-se a chave para uma preocupação da comunidade científica que se vê diante de um crescente volume de informações textuais sendo armazenadas em base de dados e, em contra partida, a percepção cada vez maior da existência de uma lacuna científica que exige uma solução para os sistemas de recuperação de informação para que apresentem métodos alternativos que busquem melhorar os índices de precisão em seus resultados de busca.

Para Bräscher (2002), esquemas de representação da informação desenvolvidos em outras disciplinas, como a inteligência artificial, podem ser usados para aprimorar a construção de sistemas de recuperação da informação que alcancem melhores índices de precisão no resultado da busca. Na linha de pensamento de Bräscher (2002), Weber-Lee

(1998), Bueno (1999) e Braga Júnior (2001) propõem modelos de sistema de recuperação de informação relacionado a inteligente artificial que utiliza teoria do raciocínio baseado em casos na recuperação de informações jurídicas.

O Raciocínio Baseado em Casos (RBC), no entendimento de Harmon e King (1988), é uma das áreas de estudo da inteligência artificial que se ocupa de sistemas que usem o conhecimento simbólico para simular o comportamento de especialistas. A idéia básica do enfoque raciocínio baseado em casos é resolver um novo problema relembrando uma situação anterior similar, dessa forma, reutilizando informação e conhecimento daquela situação (REISBECK e SCHANK, 1989).

Com base em estudos de Weber-Lee (1998), Bueno (1999) e Braga Júnior (2001), essa pesquisa apresenta uma proposta de avaliação do modelo de recuperação de informação baseado em casos a partir da sua aplicação no ambiente do Tribunal Regional Eleitoral do Distrito Federal e posterior verificação do seu índice de precisão no resultado de busca por informações jurídicas. É sua intenção verificar se o modelo aplicado melhora a relevância dos resultados das buscas na base de dados textual, aproximando o resultado de uma consulta à necessidade de informação do usuário.

1. Formulação do problema

Jurisprudência é a forma pela qual os tribunais respondem ao caso concreto, firmando entendimento pacífico para futuros casos similares. É, portanto hábil ferramenta de orientação não só aos advogados, bem como aos magistrados (MARINONI, 2001).

A fim de prover serviço de acesso às bases de jurisprudência eleitoral, o Tribunal Regional Eleitoral do DF desenvolveu um sistema de Recuperação de Informação (RI) de jurisprudência eleitoral, disponibilizado no seu sitio na Internet, que visa recuperar decisões judiciais (sentenças, acórdãos, súmulas) que sejam úteis ou relevantes para apoiar as atividades de magistrados, juristas, advogados, estudantes e cidadãos (usuários), pertencentes ou não ao quadro de pessoal do Tribunal.

De maneira geral, os usuários de sistemas de recuperação da informação formulam suas sentenças em linguagem natural, por meio de uma consulta que pode ser convertida em um conjunto de palavras-chaves formadas a partir da retirada dos termos de ligação (stopwords[1]) utilizados na sentença original e, ainda, pela redução das palavras à sua raiz

[1] Palavras comuns como preposições e artigos que são removidas para facilitar o processamento da linguagem natural

gramatical (stemming[2]) (RIJSBERGEN, 1979). No caso do Tribunal Regional Eleitoral do Distrito Federal, para busca de casos jurídicos relevantes, tais palavras-chaves são então comparadas com termos índices criados a partir do processo de indexação das jurisprudências eleitorais.

Uma inconveniência desta abordagem é que na recuperação baseada em palavras-chaves geralmente há uma distância semântica entre a necessidade do usuário e o conjunto de jurisprudências que são recuperados, ou seja, nem sempre as jurisprudências recuperadas são de interesse dos usuários. Baeza-Yates e Ribeiro-Neto (1999) definem esse sistema de recuperação de informação como modelo clássico. São exemplos, o modelo booleano, o modelo vetorial, o modelo *fuzzy*. O Tribunal Regional Eleitoral do Distrito Federal utiliza-se do modelo booleano para recuperar jurisprudências eleitorais. Em uma pesquisa com a coleção STAIRS, acrônimo de STorage And Information Retrieval System, Maron e Blair (1985) concluíram que modelo de sistema de recuperação de informação booleano apresentam índice de precisão máximo de 25%, ou seja, será necessário ao profissional do direito que utilize o serviço de consulta a jurisprudência eleitoral do Tribunal Regional Eleitoral do Distrito Federal ler cem casos jurídicos para poder encontrar vinte e cinco entre eles que lhe seja relevante.

A introdução de uma ferramenta de recuperação de informação que utiliza o raciocínio baseado em casos pode ser uma solução para melhorar a qualidade do resultado da pesquisa em bases de jurisprudência eleitoral, aproximando o resultado da consulta da necessidade de informação (NI) dos usuários.

Contudo, a seguinte questão ainda precisa ser respondida: o uso de um sistema de recuperação de informação que utiliza o raciocínio baseado em casos, pautado no modelo de recuperação inteligente de jurisprudência (WEBER-LEE, 1998; BUENO, 1999; BRAGA JÚNIOR, 2001), produzirá melhores índices de precisão, comparados aos do modelo clássico (BLAIR e MARON, 1985), aprimorando, dessa forma, o sistema de recuperação de jurisprudência eleitoral do Tribunal Regional Eleitoral do Distrito Federal?

2. Objetivo da Pesquisa

Contribuir para a consolidação de um modelo inteligente de recuperação de jurisprudência que se estabeleça como um novo paradigma para a construção de sistemas de recuperação de informação que apresentem resultados de busca mais relevantes.

[2] Processo de redução de uma palavra à sua parte essencial do item lexical, semelhante ao radical

2.1 Objetivo Geral

Aplicar o modelo de Braga Júnior (2001) no Tribunal Regional Eleitoral do Distrito Federal para avaliar se esse modelo melhora a medida de precisão no resultado da busca por informações jurídicas eleitorais.

Melhorar o nível de satisfação dos usuários em relação à recuperação de jurisprudência eleitoral, elevando o nível das taxas atuais de precisão, diminuindo o esforço do usuário nas consultas realizadas na base de jurisprudência eleitoral do Distrito Federal;

Buscar um sistema de recuperação de informação para área jurídica que permita aproximar os resultados de uma busca por jurisprudência da necessidade de informação do profissional do direito.

Possibilitar que magistrados, advogados, estudantes do direito e outros interessados tenham acesso mais preciso aos acórdãos que mais se assemelhem às suas necessidades de argumentação jurídica;

Comprovar cientificamente, por meio do levantamento e confronto do resultado obtido nas métricas de precisão, a possibilidade de aplicação e adequação do modelo de raciocínio baseado em casos para recuperação inteligente de jurisprudência na Justiça Eleitoral.

2.2 Objetivos Específicos

Desenvolver um protótipo baseado no modelo em estudo que utilize a técnica de inteligência artificial conhecida como raciocínio baseado em casos;

Construir uma base de casos jurídicos a partir de um corpus de jurisprudências definido;

Realizar comparações entre um caso jurídico novo informado pelo usuário e os existentes na base de casos jurídicos utilizando-se o cálculo de similaridade para definir o grau de proximidade entre os casos jurídicos, podendo ser essa proximidade máxima (100% de similaridade entre os casos jurídicos comparados) ou inexistente (0% de similaridade entre os casos jurídicos comparados);

Permitir a comparação entre os casos jurídicos novos que representam a necessidade de informação dos usuários e os casos jurídicos existentes na base de casos jurídicos eleitorais (jurisprudências);

Relacionar os casos jurídicos da base de casos que possuam no mínimo 50% de similaridade com o caso jurídico novo;

Avaliar se esse protótipo melhora o índice de precisão no resultado da busca quando comparado ao sistema de recuperação de informação atualmente utilizado pelo Tribunal Regional Eleitoral do Distrito Federal.

3. Justificativa

Considerando a grande concentração de órgãos do poder judiciário no âmbito do Distrito Federal, a recuperação de informação jurídica passa a ocupar um importante papel no cenário das pesquisas por jurisprudências. Deve-se considerar, ainda, que a evolução científica proposta para mecanismo de recuperação da informação jurídica, na medida em que busca a elevação qualitativa do atendimento dos seus usuários, pode ser vista como instrumento de modernização do processo de atendimento ao público. Além de permitir aos órgãos do poder judiciário promover o aprimoramento dos seus sistemas de recuperação de informação provendo maior satisfação ao seu público alvo por permiti-los recuperar jurisprudências que satisfaçam às suas necessidades de argumentações jurídicas com maior grau de precisão.

4. Estrutura da Dissertação

Essa dissertação foi organizada em capítulos assim descritos: No capítulo 1 são apresentadas as referências teóricas que sustentam conceitos que foram adotados na pesquisa, tais como raciocínio baseado em casos, jurisprudência eleitoral, sistemas de recuperação de informação, avaliação de sistemas de recuperação de informação. Além de uma revisão da literatura no que diz respeito às teorias e métodos de avaliação de sistemas de recuperação de informação, discutindo sua principal medida de avaliação que será utilizada nessa pesquisa. No capítulo 2, a dissertação apresenta a estrutura metodológica utilizada na pesquisa. É apresentado o tipo de pesquisa adotada, a definição da amostra utilizada, o pessoal envolvido, as métricas utilizadas, as variáveis definidas. No capítulo 4, encontra-se a análise dos dados. Os dados coletados e tabulados são apresentados para apoiar a formulação das conclusões da pesquisa. Por fim, no capítulo 4 são apresentadas as conclusões obtidas, confrontando-se o pressuposto da pesquisa ao resultado do caso em análise para verificar sua validade para o domínio analisado, servindo de subsídio para construção de uma explanação sobre o caso estudado. Uma lista de referências utilizadas na pesquisa e a apresentação dos anexos necessários ao melhor entendimento do trabalho estão ao final da dissertação.

REFERENCIAL TEÓRICO

1. Jurisprudência Eleitoral

O Tribunal Regional Eleitoral do Distrito Federal, como órgão julgador, está organizado, de acordo com seu regimento interno, em duas diferentes instâncias decisórias. Na primeira estão os Cartórios Eleitorais, cuja jurisdição pode incluir um município ou comarca, ou então abranger mais de um município ou comarca. Sua administração fica a cargo de um Juiz Eleitoral que deve ser obrigatoriamente um Juiz de Direito em exercício na Justiça Comum. Caberá ao juiz eleitoral o julgamento de matéria eleitoral submetida à instância decisória da sua competência. Suas decisões, pautadas na lei e na sua consciência, são proferidas sob a forma de sentenças. Neste caso, as decisões são de caráter monocrático, ou seja, representam o entendimento de um único juiz. Na segunda instância está a Corte do Tribunal, composta por sete membros: Presidente, Vice-Presidente e cinco juízes que compõem o Tribunal Pleno (TRE-DF, 2008). Compete ao Tribunal processar e julgar matérias eleitorais e administrativas. As decisões do Pleno serão lavradas em forma de acórdão ou resolução e estas, portanto, são decisões de órgão colegiado, ou seja, tomadas por colégio de juízes, proferidas com base em argumentos fundamentados acerca da existência, ou não, de um direito aplicável à determinada situação concreta. As decisões monocráticas ou colegiadas, que são formadas por sentenças, acórdãos e resoluções, formam a base da jurisprudência do Tribunal Regional Eleitoral do Distrito Federal.

A jurisprudência é o complexo de decisões pronunciadas pelos órgãos judiciários no efetivo desenvolvimento da função jurisdicional (Guimarães, 1994). Marinoni (2001, p. 156) define a jurisprudência "como a forma pela quais os tribunais respondem ao caso concreto, firmando entendimento pacífico para futuros casos similares". Sendo, portanto, hábil ferramenta de orientação não só aos advogados, bem como aos magistrados. Para Reale (1994), Jurisprudência é fonte formal do Direito Positivo, sendo utilizada na interpretação de leis e solução de casos jurídicos. Dessa forma, pode-se perceber a importância que o acesso preciso às informações jurídicas deve assumir no âmbito de um órgão julgador, cujo papel é também divulgar seus feitos jurídicos à comunidade, obedecendo, neste caso, ao princípio jurídico da publicidade, também sendo sua obrigação prover mecanismos, ou mesmo ferramentas que permitam acesso a esses feitos.

Pontes de Miranda (2000) entende que a Lei de Introdução ao Código Civil exprime o que sejam as fontes do direito em nosso ordenamento positivo, já que em seu Artigo 4º diz: "Quando a lei for omissa, o Juiz decidirá o caso de acordo com a analogia, os costumes e os

princípios gerais de direito". Ainda no entendimento de Pontes de Miranda (2000), em decorrência dos casos omissos em Lei aplicam-se as disposições concernentes a casos análogos e, não os havendo, os princípios gerais de direito. Portanto, são nas omissões legais que a jurisprudência é estabelecida e passa a servir de parâmetro para julgamentos de casos semelhantes no futuro. Clovis Bevilacqua (1972) entende que a lei é a forma por excelência do direito, alinhando-se ao pensamento de Pontes de Miranda (2000) quando admite que na omissão legal, o aplicador da lei deve louvar-se na analogia, operação lógica, em virtude da qual o intérprete estende o disposto da lei a casos por ela não previstos. Dessa forma, a jurisprudência torna-se produto jurídico com grande valor agregado, contribuindo deforma direta para o êxito de uma ação judicial.

1.1. Metodologia utilizada para análise e indexação de jurisprudências

Para que as jurisprudências sejam disponibilizadas para consultas Web por meio do seu sítio, a Coordenadoria de Jurisprudências, subordinada à Secretaria Judiciária do Tribunal Regional Eleitoral do Distrito Federal, utiliza um conjunto de normas e procedimentos desenvolvidos pelo Tribunal Superior Eleitoral que foram compiladas e publicadas no *Manual do Analista de Jurisprudência* (TSE, 2004) e são adotados a fim de permitir a correta organização (indexação, catalogação) e armazenamento das decisões pronunciadas pelos órgãos julgadores do Tribunal.

Segundo o TSE (2004), o principal objetivo da indexação é possibilitar a recuperação de documentos a partir da descrição de seu conteúdo temático, a fim de responder, de maneira eficiente e econômica, às necessidades de informação do usuário. O acórdão é o documento jurisprudencial típico, porém no âmbito da Justiça Eleitoral também são tratados como documentos de jurisprudência as sentenças e as resoluções, estes últimos, contendo decisões tipicamente administrativas. Entretanto, somente os acórdãos e resoluções são indexados, catalogados e armazenados na base de jurisprudência utilizando-se a mesma metodologia em ambos os documentos.

É função do indexador estabelecer vínculo entre os elementos contidos no texto de uma decisão judicial. O conteúdo de uma jurisprudência deve conter o fato ocorrido, o direito discutido, o posicionamento judicial e o fundamento desse posicionamento. Um documento de jurisprudência, em sua estrutura, deve possuir: elementos descritivos e elementos temáticos. De acordo com o manual do analista de jurisprudência, os elementos descritivos são aqueles que permitem identificar fisicamente um documento, diferenciando-o dos demais; já os elementos temáticos estão relacionados ao conteúdo das decisões e constituem os seus

requisitos essenciais. São eles: ementa, relatório, motivação (fundamentação) e dispositivo. A ementa é a síntese do acórdão; o relatório narra e descreve os fatos do processo, reportando-se ao direito que está sendo discutido pelas partes; a motivação ou fundamentação resulta da análise feita pelo juiz sobre as questões de fato e de direito expostas no relatório; e o dispositivo é a parte final do acórdão e caracteriza a manifestação ou posicionamento do Judiciário.

A seguir será apresentado um exemplo da aplicação da metodologia utilizada na análise e indexação de jurisprudências:

1.2 Comentários ao Acórdão

A partir da publicação de uma jurisprudência como, por exemplo, o acórdão nº 2686 de 11/12/2007 constante do Anexo B, o analista de jurisprudência do Tribunal Regional Eleitoral do Distrito Federal, utilizando-se do formulário de indexação apresentado no Anexo A, inicia a análise do documento jurídico em questão extraindo dele o seguinte:

Enunciado:

1. Fato (F) (*situação concreta que deu origem à questão sub judice*): prática, *in thesi,* do tipo penal previsto no art. 347 do Código Eleitoral.

2. Instituto Jurídico (IJ) (é o direito que se discute no âmbito da situação fática): habeas corpus com pedido de concessão de liminar.

3. Entendimento (E) (é o elo, positivo ou negativo, que o tribunal estabelece entre fato e o instituto jurídico): ausência do interesse processual.

4. Argumento (A) (conjunto de razões dadas pelo tribunal para sustentar o entendimento): o mérito do pedido não pode ser enfrentado porque ocorreu a perda superveniente do objeto com a prolação de sentença em ação penal, esgotando a jurisdição do primeiro grau.

A partir dos elementos do enunciado obtido em acórdão constante do Anexo B, identificam-se palavras-chave que servirão como pontos de acesso (TAYLOR, 2004) ao acórdão depois de armazenado na base de jurisprudências eleitorais do Tribunal Regional Eleitoral do Distrito Federal.

Indexação:

Ausência, interesse processual (E), *habeas corpus*, concessão, liminar (IJ), prática, tipo penal, art. 347, Código Eleitoral (F), perda, superveniente, objeto, prolação, sentença, esgotamento, jurisdição, primeiro grau (A)

Por fim, descrevem-se as referências utilizadas pelo tribunal para dar embasamento aos argumentos.

Referência Legislativa:

HC, acórdão nº 280.352 de 18/09/2007 – TRE-DF.

1.3 Sistemas de recuperação de jurisprudência eleitoral

A fim de prover serviço de acesso às suas bases de jurisprudência, o Tribunal Regional Eleitoral do DF desenvolveu um sistema de Recuperação de Informação (RI), disponibilizado no seu sitio na Internet, que visa recuperar decisões judiciais (acórdãos, resoluções) que sejam úteis ou relevantes para apoiar as atividades das assessorias dos magistrados, da Procuradoria Regional Eleitoral, da Corregedoria Regional Eleitoral, de Tribunais Regionais Eleitorais de outras unidades da federação, membros do poder legislativo e executivo, juristas, advogados e representantes de partidos políticos.

1.3.1 Arquitetura de software de recuperação de jurisprudência

A gestão eletrônica das jurisprudências no âmbito do Tribunal Regional Eleitoral do Distrito Federal está baseada em uma arquitetura que envolve três softwares que trabalham de forma integrada, o Sistema de Jurisprudência (SJUR), Inteiro Teor de Acórdãos e Resoluções (ITAR) e Sistema de Acompanhamento de Documentos e Processos (SADP), representando o "back-end" do processamento eletrônico da jurisprudência já que são operados apenas por técnicos especializados, no "front-end" está um sistema Web que representa a interface com o público alvo dos processos de busca por jurisprudências eleitorais do Tribunal Regional Eleitoral do Distrito Federal. Os softwares de "back-end" foram desenvolvidos utilizando-se linguagem de programação Delphi e são operacionalizados de acordo com a arquitetura "cliente-servidor", o "front-end" do sistema foi elaborado utilizando-se a linguagem de programação Java, sendo publicada nos sítios da Internet e Intranet do Tribunal Regional Eleitoral do Distrito Federal. Os dados são mantidos em um banco de dados Oracle 10gR2 combinado com banco de dados BRSearch, que permitem a indexação textual de parte de seus conteúdos (Figura 1).

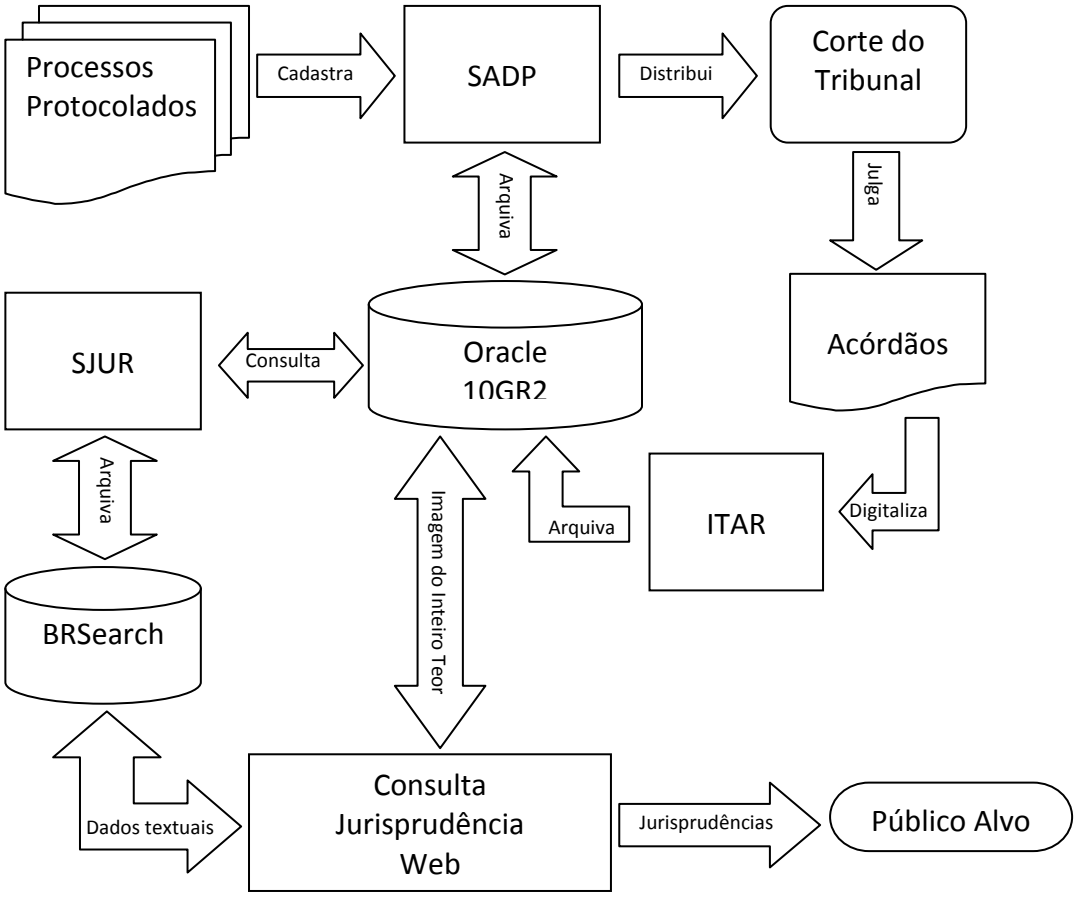

Figura 1 – Arquitetura da gestão eletrônica de jurisprudência do Tribunal Regional Eleitoral do Distrito Federal

1.3.2 Etapas do processamento de Acórdãos

Um processo ao ser protocolado no Tribunal Regional Eleitoral do Distrito Federal é cadastrado no SADP, onde receberá um número que permitirá aos órgãos internos do Tribunal, envolvidos em seu trâmite, acompanhar seu andamento.

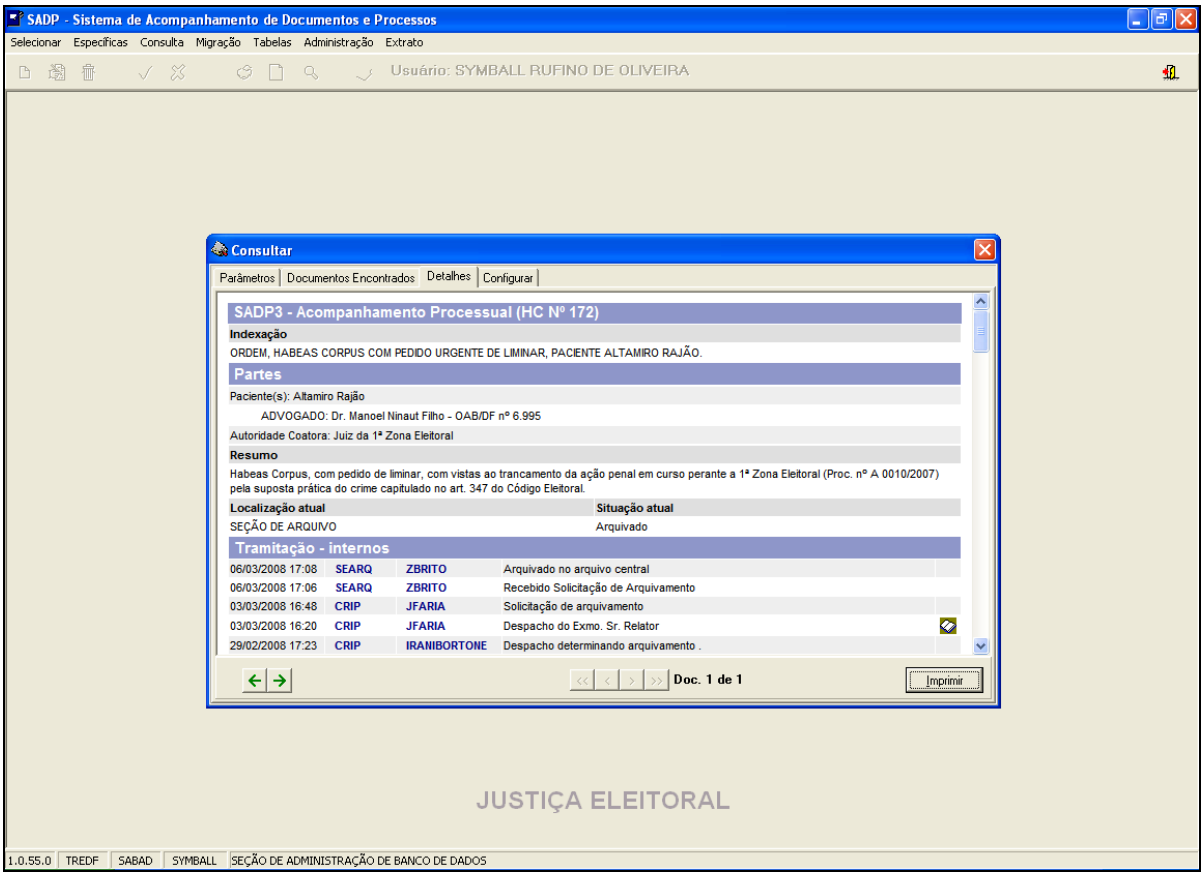

Figura 2 – Tela de cadastramento de processo no Sistema de Acompanhamento de Documentos e Processos

Após tramitação os processos são encaminhados para Secretaria Judiciária, que ainda utilizando o Sistema de Acompanhamento de Documentos e Processos, faz a distribuição dos processos de forma automática para o juiz relator. Depois de julgado o inteiro teor do Acórdão é digitalizado e armazenado por meio do Inteiro Teor de Acórdãos e Resoluções.

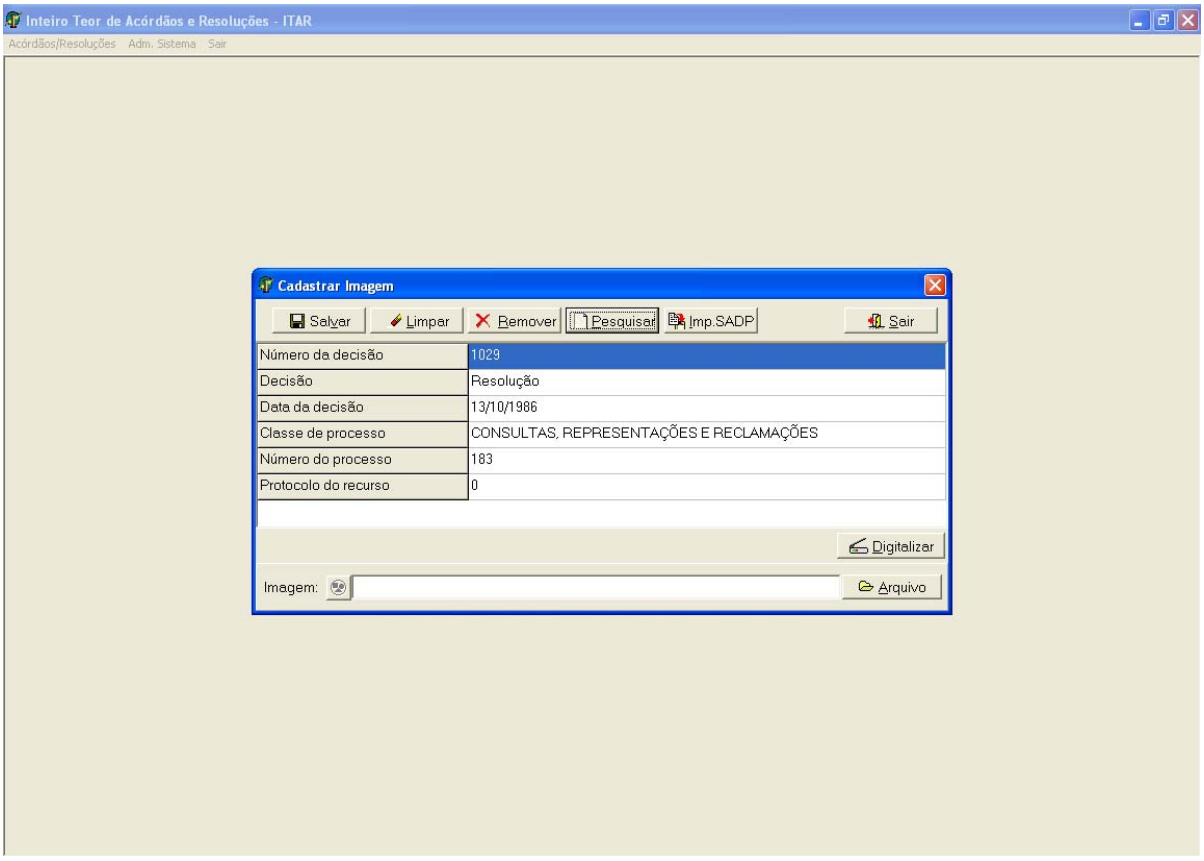

Figura 3 – Tela de cadastro no Inteiro Teor de Acórdãos e Resoluções

A decisão do colegiado é publicada no diário da justiça, em seguida comparada com os registros de jurisprudências existentes no Sistema de Jurisprudências.

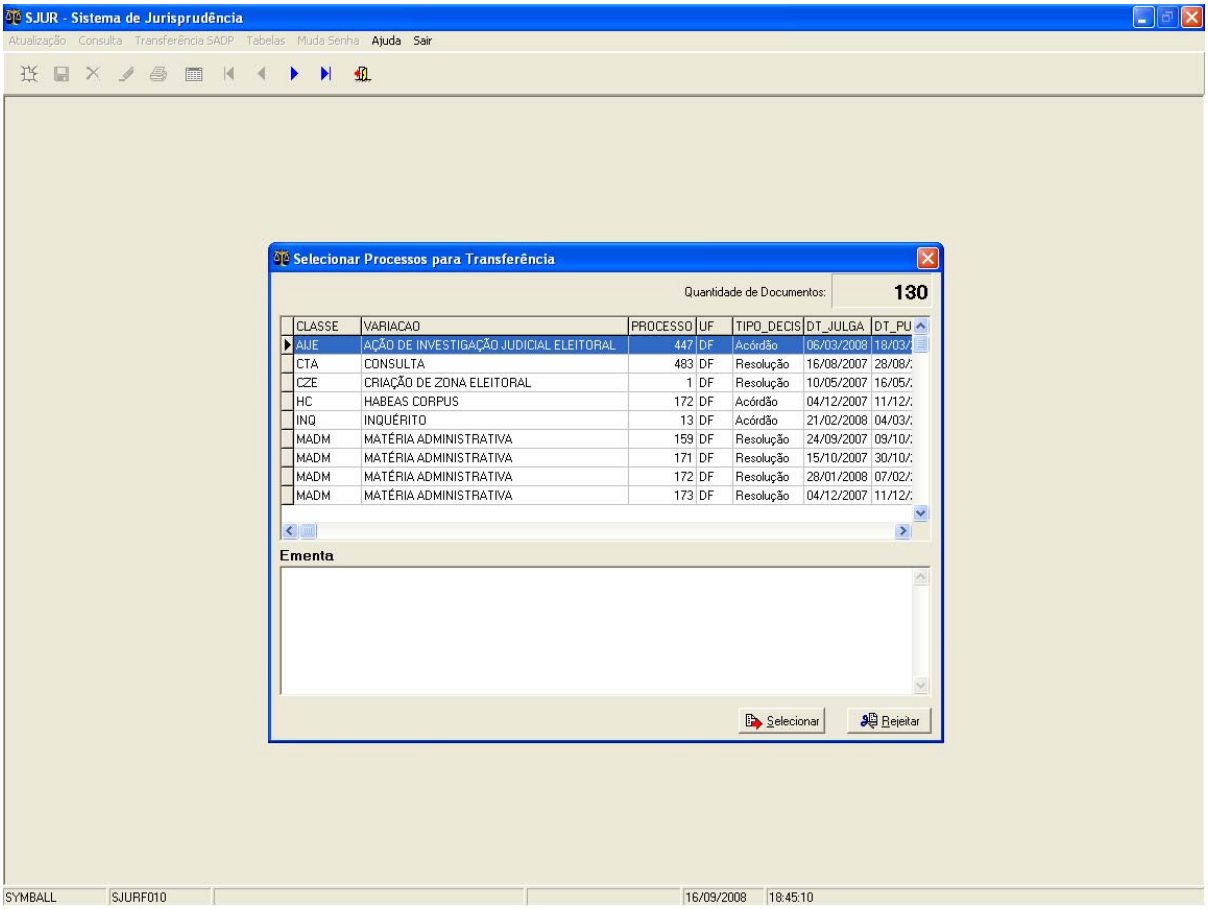

Figura 4 – Tela de consulta do Sistema de Jurisprudências

Havendo coincidência de identidade, o acórdão analisado é expurgado (descartado como jurisprudência), caso contrário, ele é indexado, cadastrado no Sistema de Jurisprudências e armazenado na base de dados para futuras consultas nos sítios da Web.

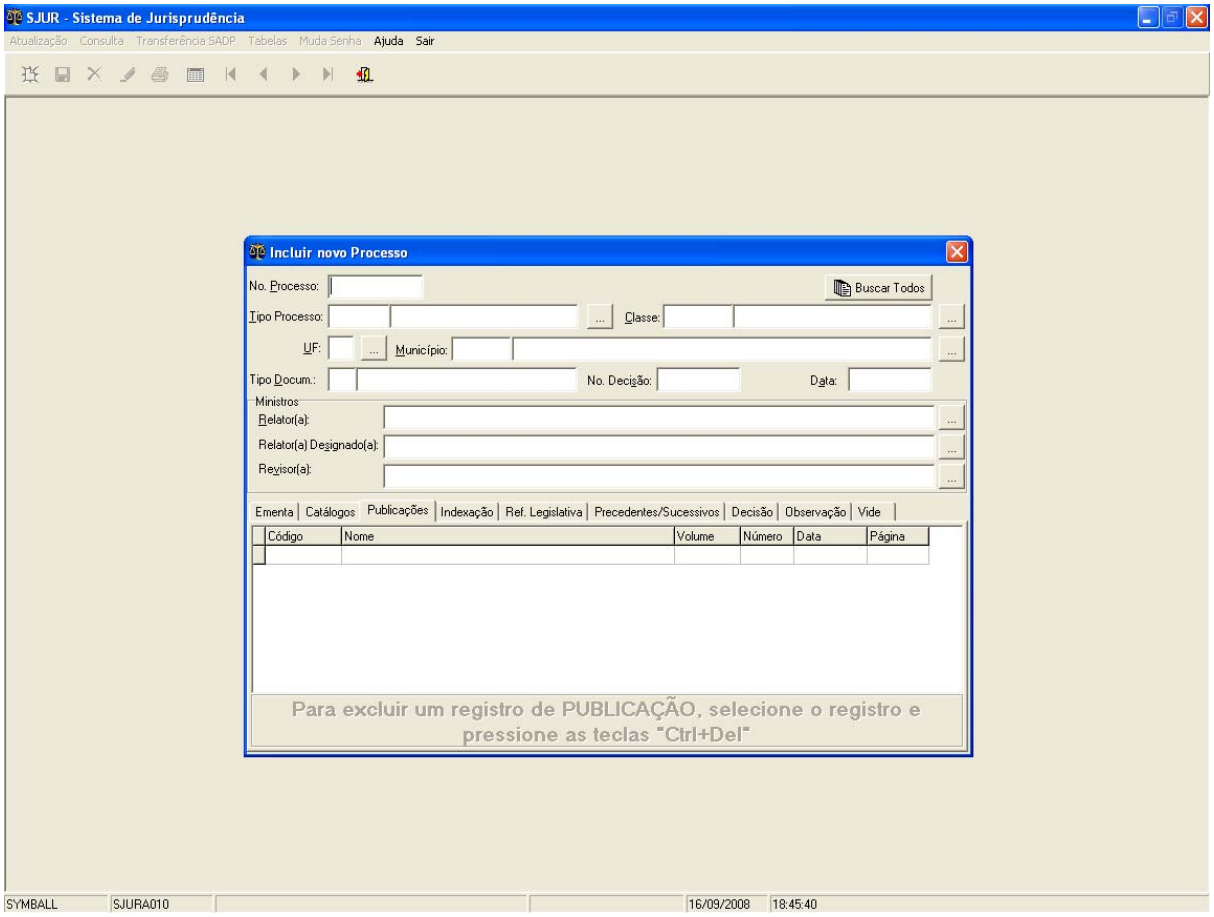

Figura 5 – Tela de cadastro do Sistema de Jurisprudências

1.3.3 Recuperação de jurisprudência eleitoral

Segundo dados fornecidos pela Secretaria Judiciária do Tribunal Regional Eleitoral do Distrito Federal, Setor responsável pela análise, indexação, armazenamento e divulgação das jurisprudências eleitorais firmadas no âmbito do Tribunal, são armazenados, em média, quatrocentos e cinqüenta acórdãos em seus bancos de dados textuais em anos não eleitorais, já nos anos eleitorais esse número pode passar de um mil. Atualmente, existem mais de dez mil jurisprudências eleitorais armazenadas nos banco de dados textuais que estão sob a responsabilidade da Secretaria Judiciária. A maior preocupação daquela Secretaria tem sido prover meios necessários ao acesso preciso a esse acervo. Essa preocupação ganha vulto na análise da afirmação de Câmara Júnior (2007) de que várias pesquisas demonstram que a percepção de qualidade dos tribunais está diretamente ligada à qualidade e facilidade das pesquisas jurisprudenciais.

Portanto, o Tribunal Superior Eleitoral desenvolveu um sistema que, integrado com os tribunais regionais eleitorais, constitui a pesquisa nacional de jurisprudências eleitorais. Por meio dessa pesquisa é possível consultar simultaneamente as bases de jurisprudências de

todos os tribunais eleitorais. Para tanto, foi criado um sistema de recuperação de jurisprudência eleitoral com interface Web e que se encontra disponível nos sítios da internet e intranet de todos os tribunais eleitorais.

Nesse sistema são utilizados dois tipos de registro de informação diferentes, a ementa que é uma estrutura textual obtida a partir da elaboração de um resumo do texto do acórdão, descrevendo seu conteúdo, e os registros que fazem a descrição física dos acórdãos, tais como: tipo de decisão, número do processo, data da decisão, data da publicação, UF de origem, tipo de processo, nome do relator, etc. A consulta é realizada com base na expressão de busca, expressa em linguagem natural, que é comparada com o texto da ementa armazenado no banco de dados, dessa forma são recuperadas todas as jurisprudências que em suas ementas existam palavras expressas na busca; além disso, podem ser informados, utilizando-se a interface do sistema de recuperação de informação, os valores dos campos de descrição física para que possam funcionar como filtros de pesquisa, limitando o resultado da busca. As jurisprudências que atendam aos requisitos da busca são listadas sob a forma de páginas de resumo contendo um link para o inteiro teor do acórdão não possuindo nenhum critério de ordenação ou semelhança, ou seja, a jurisprudência que é apresentada no topo da lista não representa necessariamente a jurisprudência julgada mais relevante pelo sistema.

Figura 6 – Tela de consulta de jurisprudência no sitio do Tribunal Regional Eleitoral do Distrito Federal

O sistema possui os seguintes tipos de pesquisas por jurisprudências disponíveis: pesquisa simples e pesquisa avançada. Na pesquisa simples o usuário poderá selecionar o(s) tribunal (ais) cuja base de dados deseja consultar, o tipo de decisão deseja consultar, que são: acórdãos, resoluções, decisões monocráticas, decisões sem resoluções; e poderá informar, em linguagem natural, palavras que considere chaves para o sucesso do resultado da busca. A pesquisa avançada permite selecionar todas as opções da pesquisa simples, além de permitir: dados da identificação do documento (número do processo, tipo de processo, número da decisão, nome do relator, etc.) e referência legislativa (normas e artigos).

Na pesquisa por jurisprudências o usuário pode utilizar o campo busca livre que permite realizar um processo de busca mediante o emprego de expressões formadas por palavras-chave. Para esse caso, o sistema possui procedimentos que permitam busca por adjacência (ADJ), utilizado quando se deseja recuperar jurisprudências onde os dois termos

informados são adjacentes entre si, na ordem em que foram digitados. Exemplo: propaganda ADJ partidária; por proximidade (PROX), utilizando quando se deseja recuperar jurisprudências onde os dois termos digitados são adjacentes entre si, em qualquer ordem. Exemplo: antecipação PROX tutela; por termos da busca no mesmo campo (MESMO), utilizando quando se deseja recuperar jurisprudências onde os dois termos digitados pertençam ao mesmo campo da jurisprudência, por exemplo, ao campo ementa. Exemplo: inegibilidade MESMO parentesco; e pelo uso de operadores lógicos matemáticos (E, OU, NÃO). Operador E, usado quando se deseja recuperar jurisprudências que tenham ambos os termos. Exemplo: inegibilidade E prefeito; operador OU, usado quando se deseja recuperar jurisprudências que contenham pelo menos um dos termos. Exemplo: crime OU delito; Operador NÃO, usado quando se deseja recuperar jurisprudências que não contenham o termo utilizado. Exemplo: NÃO crime.

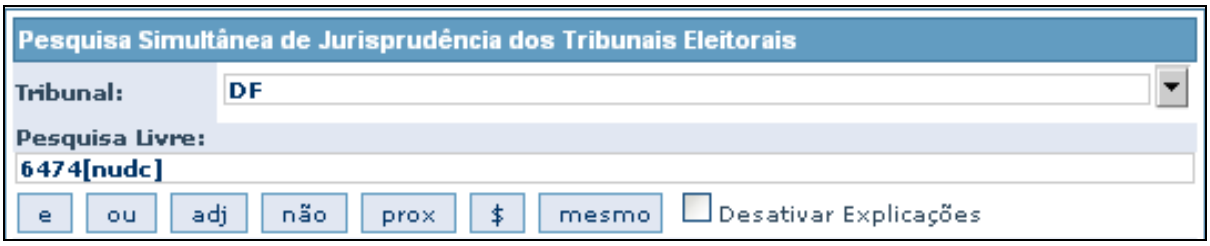

Figura 7 – Tela de consulta pelo campo busca livre

Esses termos podem ser combinados em uma mesma expressão de busca. A busca livre irá localizar a ocorrência das palavras-chave em todos os campos do banco de dados, mas poderá ser limitada a somente alguns campos mediante seleção prévia dos campos que serão verificados.

Exemplo de pesquisa livre em campos especificados por meio de busca livre: usuário procura apenas por Resoluções. Campo a ser pesquisado: Tipo de Decisão. Sigla do campo: tdde. Expressão de busca: resolução [tdde]

Quando nenhum campo é selecionado, a busca ocorre comparando-se a expressão da busca com os termos de indexação dos documentos de jurisprudência armazenados. Portanto, a busca por jurisprudências eleitorais que se utiliza do campo busca livre apresenta muitas possibilidades para sua elaboração, por isso, segundo Lancaster (1993), vão exigir um planejamento acurado da estratégia de busca, visando a uma recuperação de informação consoante com as necessidades informacionais dos usuários.

Considerando-se o perfil do público alvo do sistema de recuperação de jurisprudência no Tribunal Regional Eleitoral do Distrito Federal, como visto antes, advogados, magistrados, assessores, a necessidade de se ter bom conhecimento do assunto

para que se obtenha sucesso na construção da estratégia de busca não configuraria um problema aparente. Pode-se esperar que tenham dificuldade na utilização de parâmetros lógicos na construção da consulta (query), mas não parece ser esse o maior problema encontrado por profissionais que buscam jurisprudências eleitorais. A dificuldade está inserida ainda na baixa precisão do resultado da busca, conforme preconizado por Maron e Blair (1985).

2. O Raciocínio Baseado em Casos

2.1 Histórico

Na tentativa de criar modelos cognitivos de solução de problemas e do aprendizado de situações com base em memórias episódicas, Schank e Abelson (1977) desenvolveram, na década de 70, estudo sobre Memória Dinâmica. O estudo propõe que o nosso conhecimento fica gravado na memória na forma de roteiros[3] que servem como plano de ação para o celebro na resolução de problemas. Como exemplo pode-se citar o roteiro do restaurante de Schank, que mostra o conjunto de ações (roteiro) que normalmente realizamos quando vamos ao restaurante: entrar, escolher mesa, sentar-se, esperar pelo garçom, aceitar o cardápio, ler o conteúdo do cardápio, fazer o pedido, comer, pedir a conta, pagar, ir embora (SCHANK, 1982). Este estudo é considerado por muitos pesquisadores como uma das principais origens do raciocínio baseado em casos.

Schank (1982) continuou explorando os padrões de situações e influenciado pelas Teorias da Formação de Conceito, da Resolução de Problema e da Aprendizagem Experimental, desenvolveu um novo estudo que considerava que as lembranças de casos passados estavam estritamente relacionadas à resolução de problemas e ao aprendizado. Com esse estudo, Schank (1982) consolidou a teoria da Memória Dinâmica que estruturava e representava o conhecimento (roteiros) sob a forma de Pacotes de Organização de Memória (MOP – Memory Organization Packages). Para Schank (1982), "as estruturas de conhecimentos organizadas como roteiros na memória poderiam ser alteradas por novas experiências vividas pelo indivíduo, promovendo, dessa forma, a aprendizagem" (a criação de novos roteiros).

Aamodt e Plaz (1994) consideram também importante para a evolução do raciocínio baseado em casos os trabalhos realizados por Wittgenstein (1953) quando observou que conceitos naturais, como cadeira e mesa, são polimórficos, podendo ser classificados apenas

[3] Em ingles: *scripts*

como conjunto de instâncias que na percepção dos autores devem ser entendidas como casos que possuem similaridades.

No Brasil, as pesquisas têm sido direcionadas à aplicação da tecnologia raciocínio baseado em casos no desenvolvimento de soluções inteligentes para a gerência do conhecimento em aplicações do domínio de engenharia de software e banco de dados inteligentes. Grande parte dessas pesquisas é realizada na Universidade Federal de Santa Catarina (UFSC) e Universidade do Vale do Itajaí (UNIVALI), por meio do grupo de Gerência do Conhecimento Baseado em Casos e pelo grupo de Bancos de Dados Inteligentes, criado em 1993, vinculado ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande do Sul (UFRGS). Esses grupos de pesquisa produziram trabalhos como Um estudo sobre o Raciocínio Baseado em Casos (Abel, 1996) que demonstra o interesse dos pesquisadores pela disciplina que começava a surgir no Brasil. Bueno (1999) e Weber (1998) publicaram trabalhos que apontam para uma nova aplicação do raciocínio baseado em casos no campo das pesquisas brasileiras, a recuperação de informação jurídica. Alimentando-se dessas fontes, Braga Junior (2001) propõe um estudo da aplicação do raciocínio baseado em casos à recuperação da informação jurídica construindo um modelo cujo pressuposto é a melhora na qualidade do resultado das buscas por informações jurídicas.

2.2 Definição

O Raciocínio Baseado em Casos (RBC[4]), segundo Harmon e King (1988), é uma das áreas de estudo da Inteligência Artificial que se ocupa de sistemas que usem o conhecimento simbólico para simular o comportamento dos especialistas humanos. Os estudos do raciocínio baseado em casos foram influenciados por várias outras áreas como ciências cognitivas, sistemas baseados em conhecimento, aprendizagem de máquinas, redes neurais. Há nessa disciplina pontos comuns com campos da ciência como reconhecimento de padrões, incerteza e estatística. Para Riesbeck e Schank (1989) A idéia básica do enfoque do raciocínio baseado em casos é resolver um novo problema relembrando uma situação anterior similar, dessa forma, reutilizando informação e conhecimento daquela situação.

Gentner (1983) entende o raciocínio baseado em casos como um novo ramo de solução de problemas baseado em experiências passadas. Riesbeck e Schank (1989) classificam o raciocínio baseado em casos como sistema de solução de problemas baseado em conhecimento. Leake (1996) descreve o raciocínio baseado em casos como abordagem

[4] Do inglês, *Case-Based Reasoning* (CBR)

baseada em lembranças. De acordo com Hoeschl et al. (2000), O raciocínio baseado em casos é um tipo de raciocínio que busca soluções para um determinado problema mediante a análise comparativa entre a realidade vivida e outra semelhante apresentada. Watson (1997), Aadmodt e Plaza (1994) consideram o raciocínio baseado em casos como um paradigma de resolução de problemas a partir da identificação da obtenção de um caso anterior similar. Wangenheim e Wangenheim (2003) definem o raciocínio baseado em casos como uma abordagem para solução de problemas e para o aprendizado construídas com base em experiências passadas, constituindo-se em um modelo cognitivo para se entender alguns aspectos do pensamento e comportamento humano.

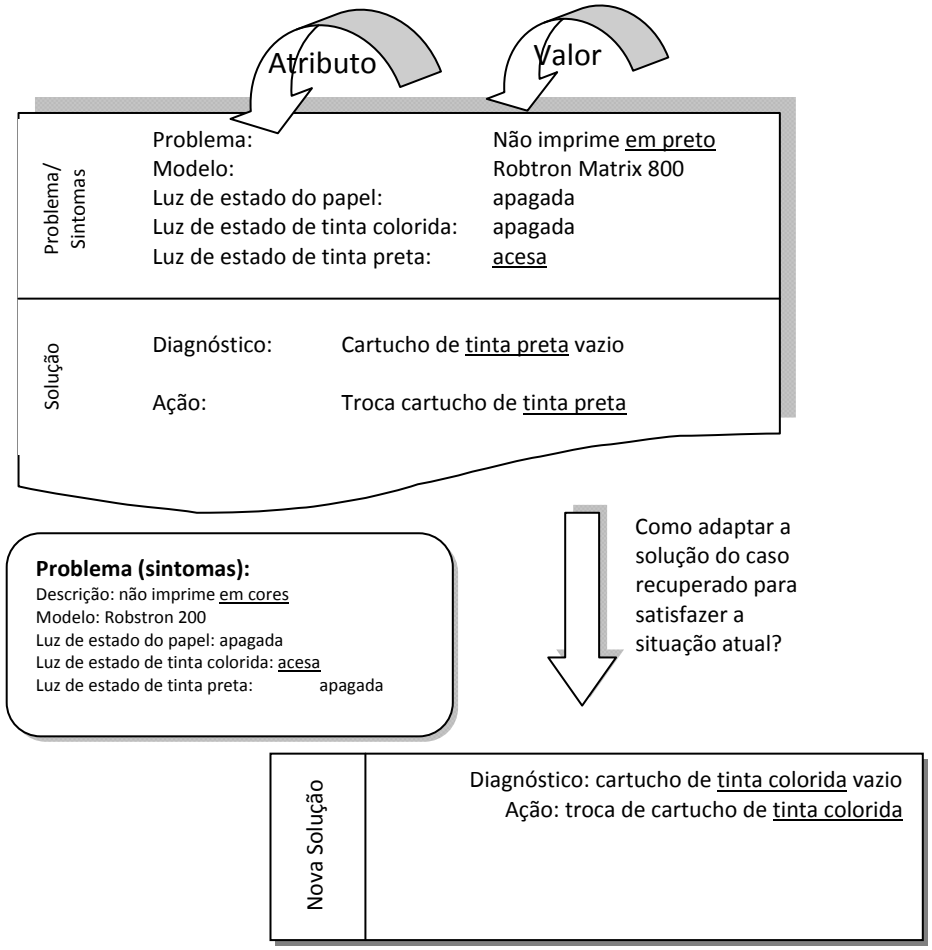

Figura 8 – Exemplo solução de um problema atual com base em um caso passado

Fonte: Wangenheim e Wangenheim (2003)

2.3 Elementos do raciocínio baseado em casos

Os elementos básicos do raciocínio baseado em casos são: a representação do conhecimento, medida de similaridade, adaptação e aprendizado (veja figura 9).

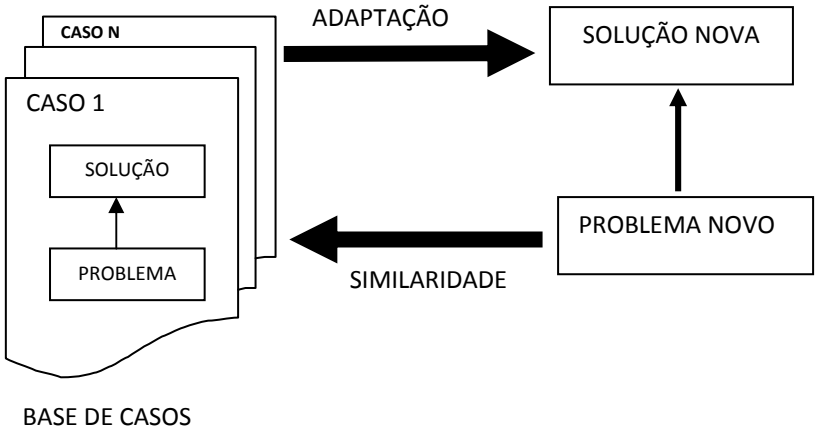

Figura 9 – Modelo básico do enfoque Raciocínio Baseado em Casos

Fonte: Wangenheim e Wangenheim (2003)

Em raciocínio baseado em casos a representação do conhecimento é feita principalmente em forma de casos que descrevem experiências concretas. Um caso é parte do conhecimento representado por um conjunto de experiências passadas acumuladas por um indivíduo (Watson, 1997). A medida de similaridade é central para aplicação do raciocínio baseado em casos na recuperação de informação. Pode-se entender similaridade entre um caso novo e um caso passado como sendo a utilidade desse caso para a condução da solução daquele outro. Formalmente, a medida de similaridade entre um conjunto de casos novos **N** e um conjunto de casos passados **P** pertencente ao universo **U** é uma função do tipo: **sim(x,y):(NxP→[0,1])**, tal que **x Є N** e **y Є P**. Se **sim(x,y)=1**, então a similaridade entre o caso x (novo) e o caso y (passado) é máxima e se **sim(x,y)=0** não há similaridade entre x e y. Nos casos em que **0 < sim(x,y) < 1**, a adaptação do caso passado (y) para solução do caso novo (x) passa ser um elemento necessário porque situações passadas dificilmente serão idênticas ao problema atual, por isso a necessidade de adaptá-la para utilização na solução do novo caso. O aprendizado determina a capacidade do modelo de se manter atualizado, já que sempre que houver sucesso na solução de um problema com base em um caso passado, deverá ser possível aplicar a mesma solução na ocorrência de um caso novo similar no futuro.

2.4 O Ciclo do raciocínio baseado em casos

Dependendo da abordagem dada para aplicação do raciocínio baseado em casos, pode-se ter diferentes tarefas compondo seu ciclo. LeaKe (1996) divide o sistema do raciocínio baseado em casos em duas abordagens: interpretativa e de resolução de problemas. Abordagens interpretativas têm foco na interpretação do novo caso, sem se preocupar com a solução do problema. Nesta situação os casos passados servem como base para o

entendimento do problema atual. Na abordagem de resolução de problemas, cujo foco está alinhado com esse trabalho, Aamodt e Plaza (1994) propõem uma arquitetura do ciclo do raciocínio baseado em casos conhecida como 4R, em referência às iniciais do nome d cada uma das etapas do ciclo. Esta arquitetura engloba um ciclo de raciocínio contínuo composto por quatro tarefas principais:

- Recuperar casos similares ao problema a partir da base de casos;
- Reutilizar a informação e o conhecimento contidos no caso recuperado, adaptados ou não, para contribuir na solução do problema;
- Revisar a solução proposta; avaliar a solução encontrada;
- Reter informações e conhecimentos revisados obtidos no caso passado que contribuíram para a construção da solução do problema. Esta tarefa caracteriza a aprendizagem.

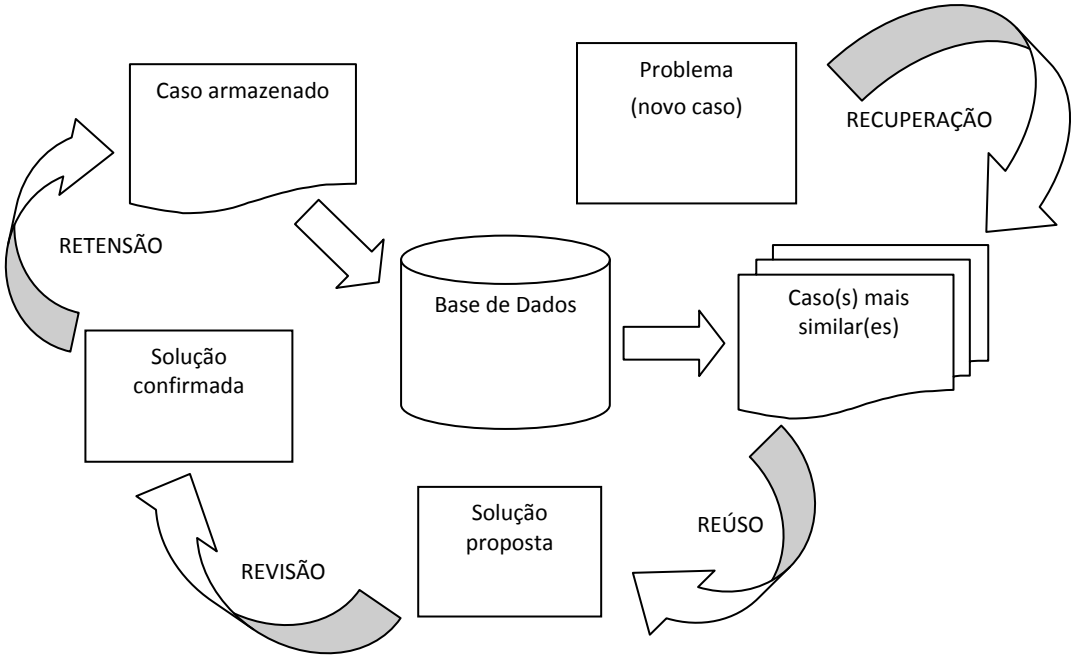

Figura 10 – Ciclo do raciocínio baseado em casos

Fonte: Aamodt e Plaza (1994)

Este ciclo raramente ocorre sem intervenção humana. Tarefas como a revisão de casos e a retenção de novos casos são deixadas a cargo da capacidade discricionária dos seres humanos.

2.5 O raciocínio baseado em casos aplicado à recuperação de informação Jurídica

O uso de métodos computacionais baseados em conhecimento para o raciocínio legal teve início no ano de 1983 em um grupo de pesquisa da Universidade de Amherst, Massachusetts, coordenado pela pesquisadora Edwina Rissland (Wangenheim e Wangenheim, 2003). O sistema legal orientado à *common law* dos EUA é fortemente enfocado em precedente da legislação, utiliza primariamente a jurisprudência em seus argumentos jurídicos e busca julgar o caso atual utilizando como base casos anteriores similares. Segundo Will (1910), nos sistemas de *common law*, o direito é criado ou aperfeiçoado pelos juízes: uma decisão a ser tomada num caso depende das decisões adotadas para casos anteriores e afeta o direito a ser aplicado a casos futuros. Por ser o sistema legal americano alicerçado em casos jurídicos passados (*common Law)*, e considerando o volume existente de jurisprudências arquivadas nos tribunais daquele país, o estudo da aplicação do raciocínio baseado em casos na recuperação de casos jurídicos realizado pelo grupo de pesquisa da Universidade de Amherst no ano de 1983 apresentou bons resultados no meio acadêmico dando visibilidade científica à pesquisa do raciocínio baseado em casos como solução para a recuperação de jurisprudências.

Mesmo em sistemas legais orientados pela civil law, como no caso brasileiro, as pesquisas da aplicação do raciocínio baseado em casos na recuperação de jurisprudências ganharam importância. O sistema *civil law*, de origem romano-germânico, é aquele cujo direito positivo é baseado na norma legal, ou seja, nesse sistema o juiz deverá julgar os casos segundo a lei e conforme a sua consciência (Lima, 1986). No sistema jurídico baseado na *civil law*, embora fundamentado na norma legal, os julgadores não devem deixar de conhecer também a jurisprudência, já que ela deve ser utilizada na interpretação de leis e solução de casos jurídicos.

A Universidade de Santa Catarina vem desenvolvendo os principais trabalhos científicos na área de aplicação do raciocínio baseado em casos à recuperação de informação jurídica. Esses estudos são desenvolvidos dentro do grupo de pesquisa chamado Gerência do Conhecimento Baseado em Casos dirigido pela pesquisadora Christiane Gresse Von Wangenheim (1999, 2003). Um dos trabalhos mais marcantes da recuperação de informação jurídica utilizando o raciocínio baseado em casos é a tese de doutorado de Rosina Weber Lee (1998) intitulada Pesquisa Jurisprudencial Inteligente. Nesse trabalho, Weber-Lee (1998) propõe o desenvolvimento de um sistema para pesquisa jurisprudencial que seja mais eficiente que os sistemas de banco de dados de texto disponíveis no mercado em termos da

utilidade das decisões judiciais recuperadas. Para tanto, apresenta o sistema Prudentia, com a intenção de demonstrar que com o emprego da tecnologia de raciocínio baseado em casos, é possível alcançar um desempenho melhor comparado com o resultado obtido pelos sistemas de banco de dados de texto existentes. Merece destaque também a dissertação de mestrado de Tânia C. D'Agustini Bueno com o título Recuperação da Informação Jurídica em Sistemas Baseados em Casos. Bueno (1999), nessa dissertação, apresenta elementos que complementam o sistema Prudentia ao buscar estabelecer um modelo de recuperação de jurisprudência utilizando o raciocínio baseado em casos. Outro estudo relevante para fundamentar a aplicação jurídica do raciocínio baseado em casos foi publicado no 2ª conferência internacional de raciocínio baseado em casos. Neste estudo, chamado *A Large Case-Based Reasoner for Legal Cases*, Weber, Bueno et al. (1997) apresentam uma metodologia baseada em conhecimento que permite converter decisões em uma base textual em casos jurídicos que podem ser armazenados em uma base de casos, a fim de permitir a construção de um mecanismo de recuperação da informação mais eficiente e com melhores resultados nas buscas por informações jurídicas. Com o trabalho: Proposta de Modelo raciocínio baseado em casos para a Recuperação Inteligente de Jurisprudência na Justiça Federal, Mario Sena Braga Júnior (2001) pretende estabelecer um modelo para construção de sistemas de recuperação de jurisprudências no âmbito do Tribunal Regional Federal – TRF.

É razoável admitir que o esforço científico para construção de tal sistema de recuperação de jurisprudência deve encontrar respaldo nos benefícios que pode trazer para a comunidade jurídica no que se refere à melhoria da precisão na recuperação de informações. É nesse sentido que esse trabalho é desenvolvido, visando dar respaldo científico às pesquisas que apontam para o surgimento de uma nova geração de sistemas de recuperação de informação jurídica.

3. Sistemas de Recuperação da Informação

A preocupação científica de desenvolver soluções para o problema evidenciado pela rápida expansão do volume de informações ocorrida no pós-guerra dos anos 50, sobretudo na produção de material bibliográfico produzido pela comunidade científica, deu origem ao termo Recuperação da Informação (RI), criado por Calvin Mooers, em 1951. Para Mooers (1951), "a recuperação de informação engloba os aspectos intelectuais da descrição da informação e de sua especificação para a busca, bem como qualquer sistema, técnica ou máquina que são utilizados para realizar a operação". A partir dessa definição e, conseqüentemente, da experiência vivenciada à época pelos profissionais da informação

vinculados às bibliotecas e centros de informação, cresce a preocupação de cientistas da informação em elaborar técnicas de organização da informação, estratégias de busca e de construção de mecanismos de busca, visando atender com maior rapidez e precisão às necessidades informacionais expressas pelos usuários desses sistemas de recuperação de informação.

3.1 Evolução histórica

Lesk (1995) apresenta a evolução no tempo da recuperação da informação e identifica uma primeira fase, nos anos 50, na qual a recuperação de informação era feita basicamente por meio do uso de referências bibliográficas, tendo o processo de indexação todo feito de forma manual.

Na década de 70, com a evolução e popularização do uso de dispositivos de informática, tais como memória e disco, o processo de recuperação de informação iniciou sua fase de automação. Aspectos do seu armazenamento e indexação, influenciados pelas tecnologias da informação, caracterizaram a introdução do uso de computadores na organização e recuperação da informação. Surgem, então, os Sistemas de Recuperação de Informação (SRI), impulsionados pela introdução da indexação automática e pela criação de algoritmos de busca (LESK, 1995).

Lancaster e Fayen (1973), em sua obra clássica *Information retrieval on-line*, definiram os sistemas de recuperação de informação como interfaces entre um conjunto específico de usuários e o conjunto de recursos informacionais a ele disponíveis, tendo como objetivo atender às suas necessidades de informação pré-definidas. Tais sistemas permitem aos seus usuários acessar, por meio de um computador, de modo direto, uma base de dados de documentos. Esses sistemas foram utilizados inicialmente por grandes bibliotecas que possuíam sistemas de recuperação de informação em linha ou *Online Public Access Catalog* (*OPAC).* Os OPACs permitiam ao usuário recuperar informações acessando-as por meio de um computador localizado na rede local da biblioteca ou mesmo remotamente, a partir de um computador doméstico (LANCASTER e FAYEN, 1973).

Swanson (1963) considera que um SRI deve ser utilizado quando há muitos recursos informacionais a serem consultados, situação em que uma consulta individual de cada recurso informacional seria inviável, sendo, na sua visão, esse o maior problema da recuperação de informação, ou seja, um sistema de recuperação de informação, em um grande universo de informações, deve retornar prioritariamente os documentos de interesse (relevantes) do usuário.

Na década de 90, quando a interconexão entre computadores era plenamente possível, os sistemas de recuperação da informação passaram a ter como desafio atender às necessidades de informação dos usuários que consultavam as grandes bases de dados textuais existentes na Web. A busca nesse repositório universal de conhecimento e cultura humano, que permite um compartilhamento sem precedentes de informações em uma escala jamais vista, tornou-se um desafio para os SRI (BAEZA-YATES e RIBEIRO-NETO, 1999). O fenômeno de expansão da Web e o vertiginoso crescimento das publicações eletrônicas direcionaram os esforços dos pesquisadores da Ciência da Informação na construção de modelos de SRI. O objetivo desses pesquisadores é atender melhor aos anseios dos usuários que cada vez mais vivem a sensação de estar se afogando em um mar de informações.

3.2 Os modelos de sistemas de recuperação da informação

Os SRI tradicionais adotam estratégia de busca de informação pautada em uma consulta (*query*) formulada pelo usuário para descrever a sua necessidade de informação. Essa consulta sumariza todas as expressões que caracterizam o objeto sobre o qual se busca uma informação. Nesse modelo, as consultas são construídas pelo usuário por meio de linguagem natural. A consulta fornecida pelo usuário é comparada com termos de indexação extraídos dos documentos, de forma manual ou automática, considerados relevantes na representatividade desse documento durante o processo de indexação (RIJSBERGEN, 1979).

Baeza-Yates e Ribeiro Neto (1999) definem um modelo de recuperação da informação como um modelo quádruplo na forma [**D**, **Q**, ***F*, *R*(q$_i$, d$_j$)**], onde:

(1) **D** = { **d**$_j$ | **d**$_j$ ϵ **I** } é um conjunto composto por termos (**d**) que representam o conteúdo dos documentos de uma coleção e que pertencem ao índice (**I**) responsável pela descrição de conteúdo desses documentos.

(2) **Q** = { **q**$_j$ | **q**$_j$ ϵ **NI** } é um conjunto composto por termos de uma consulta (**q**) responsáveis pela representação das necessidades de informações do usuário (**NI**). Essa representação é chamada de consulta (*query*).

(3) **F** é um framework para modelar a representação dos documentos, das consultas, e os seus relacionamentos.

(4) **R**(qi, dj) é a função de ordenação (ranking) que associa um valor real entre os termos da consulta (**q**) e os termos (**d**) pertencentes ao índice (**I**) que representam os documentos da coleção, permitindo definir uma ordem entre os documentos recuperados e os termos da consulta.

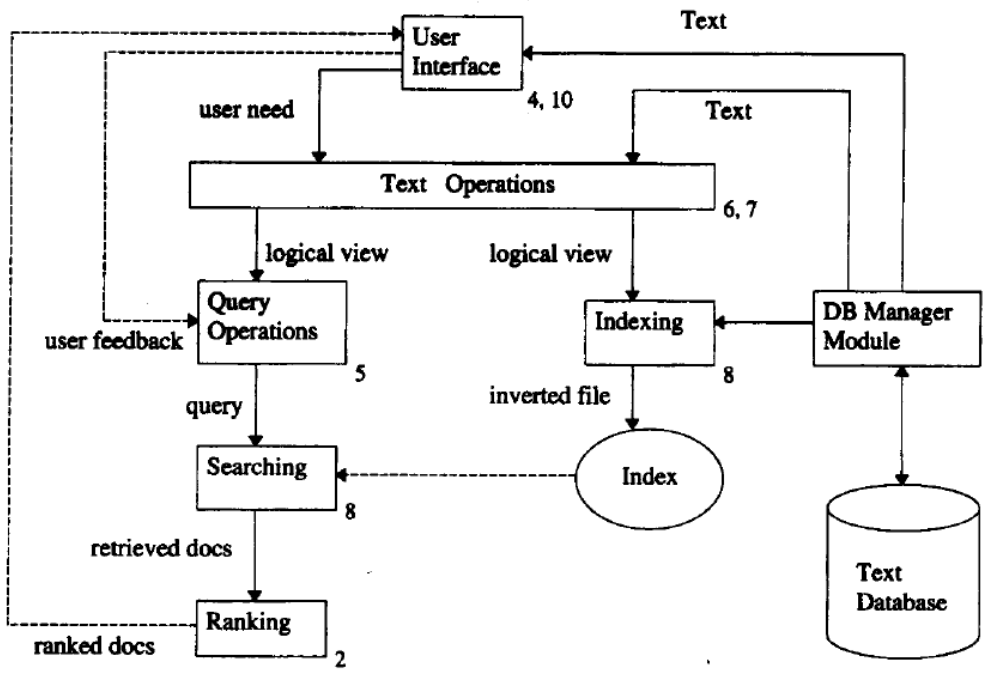

Figura 11 – Processo de recuperação da informação

Fonte: Beza-Yate e Ribeiro-Neto (1999)

A forma de ordenação do resultado de uma busca definida por uma função **R**(qi, dj) determina o modelo de recuperação de informação a ser utilizado pelo usuário, ou seja, para cada forma de ordenamento dos resultados de uma busca haverá um modelo de recuperação de informação adequado para ela. Essa ordenação determinará o grau de relevância atribuída a cada documento recuperado em relação à necessidade de informação do usuário transcrita na forma de uma consulta.

Rijsbergen (1979) esclarece que intelectualmente é possível para uma pessoa estabelecer o grau de relevância de um documento, dada a descrição de uma necessidade informacional. Já para um computador fazê-lo, será necessário construir um modelo no qual a decisão de relevância dos documentos recuperados possa ser quantificada.

Para Ferneda (2003), a eficiência de um sistema de recuperação de informação está diretamente relacionada com o modelo que ele utiliza. Um modelo, por sua vez, influencia diretamente o modo de operação desse sistema.

Baeza-Yates e Ribeiro-Neto (1999) esclarecem que o problema central da recuperação da informação é ser capaz de prever se um documento é ou não relevante para uma determinada consulta descrita por um usuário.

Ferneda (2003) classifica os modelos de sistemas de recuperação de informação em modelos quantitativos e modelos dinâmicos. Os modelos quantitativos são baseados nas disciplinas como a lógica, a estatística e a teoria de conjuntos. Nesses modelos, a

representação dos documentos é feita pela associação de termos de indexação e seus respectivos pesos aos documentos da coleção. São modelos quantitativos: o booleano, o vetorial, o probabilístico, o *fuzzy* e o booleano estendido. Os modelos dinâmicos têm como principal característica a participação do usuário na descrição dos documentos da coleção. Os usuários interferem diretamente na representação dos documentos e podem adaptar os documentos aos seus interesses informacionais, de acordo com os resultados da busca e da atribuição de relevância dada aos documentos recuperados (*relevance feedback*). Esses modelos são baseados na disciplina da inteligência artificial, mais especificamente em sistemas especialistas. São modelos dinâmicos: os sistemas especialistas, a rede semântica, a rede neural e os algoritmos genéticos.

Baeza-Yates e Ribeiro-Neto (1999) propõem uma taxonomia onde os modelos de SRI classificados como clássicos estão organizados em: modelo booleano, modelo vetorial e modelo probabilístico, tendo cada um deles uma extensão atualizada com modelos baseados no paradigma clássico, conforme figura 12.

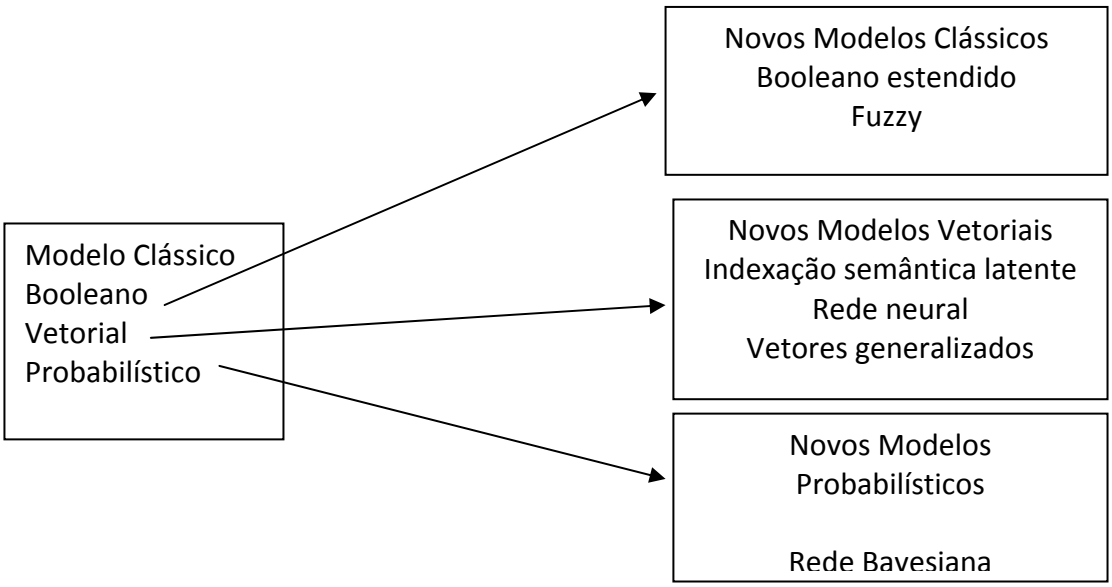

Figura 12 – Taxonomia dos modelos de recuperação da informação

Fonte: Adaptado de Beza-Yate e Ribeiro (1999)

Apesar de apresentarem classificações diferentes, os modelos de recuperação de informação apresentados encontram grande área de interseção na suas definições. Portanto, por possuírem maior relevância para essa pesquisa, os modelos booleanos e dinâmicos serão apresentados com maior grau de aprofundamento.

3.2.1 Modelo Booleano

Este modelo está baseado em um sistema binário no qual existem dois valores possíveis para qualquer símbolo algébrico: verdadeiro ou falso, 0 ou 1. Construída a partir dos estudos de lógica desenvolvidos por Aristóteles, a álgebra booleana e a teoria dos conjuntos formam a base do modelo booleano de recuperação da informação. Neste modelo, uma consulta é representada por uma expressão lógica formada a partir dos seus termos, da forma: **Q** = (**q**1 **and q**2) **or not q**3. A função de ordenação (*ranking*) **R** retorna os documentos que possuem combinações dos termos que satisfazem à construção lógica da consulta. O resultado de **R** considera relevantes os documentos D cujos termos índices satisfaçam à consulta Q, neste caso os termos da consulta **q** terão peso 1. Os termos que não aparecem na consulta terão peso 0. Ou seja, tipicamente, no modelo lógico não há classificação de documentos, já que o cálculo da similaridade (S) de uma consulta em relação a um documento será: S ($\mathbf{q}_i$, $\mathbf{d}_j$) = 1 ou 0.

São operadores lógicos desse modelo: AND, OR, NOT (E, OU, NÃO), descritos anteriormente, e apresentam como resultado os documentos cuja representação satisfaça às restrições lógicas da consulta formulada (**Q**).

AND é um operador que forma uma *expressão conjuntiva* e, quando aplicado a um par de termos **q**1 e **q**2 de uma consulta, na forma do enunciado (**q**1 AND **q**2), irá recuperar apenas documentos indexados por ambos os termos, considerando irrelevantes os documentos indexados por apenas um dos termos da consulta. Analisando esse operador sob a luz da teoria dos conjuntos, pode-se deduzir que a expressão conjuntiva (**q**1 AND **q**2) equivale à intersecção entre o conjunto dos documentos indexados pelo termo **q**1 e o conjunto dos documentos indexados pelo termo **q**2.

OR é um operador que forma uma *expressão disjuntiva* que em uma consulta na forma do enunciado (**q**1 OR **q**2) recupera os documentos indexados pelos termos **q**1 ou **q**2. A expressão disjuntiva (**q**1 OR **q**2) equivale à união entre o conjunto dos documentos indexados pelo termo **q**1 e **q**2.

NOT é um operador de *negação lógica* que, em geral, complementa as expressões disjuntivas e conjuntivas, eliminando do resultado os termos referenciados. Uma consulta na forma do enunciado (**q**1 and NOT **q**2) produz como resultado um conjunto de documentos indexados por **q**1 e que não são indexados por **q**2. Na teoria dos conjuntos é equivalente à diferença entre o conjunto dos documentos que possuem o termo **q1** e os que possuem o termo **q**2.

Operadores de proximidades são aqueles utilizados para busca no modelo lógico cujos documentos que compõem o *corpus* são armazenados sob a forma de texto completo ou parcial. Os principais operadores de proximidade são os de adjacência (ADJ) e os de aproximação (NEAR). A forma geral de uso desses operadores é $\mathbf{q}_1$ Operador n unidades de $\mathbf{q}_2$, onde **n** é a distância numérica entre os termos da expressão considerando como medida as unidades. As unidades podem ser expressas em forma de palavras, linhas ou parágrafos. Portanto, uma expressão lógica do tipo $\mathbf{q}_1$ ADJ5 $\mathbf{q}_2$ recupera todos os documentos que possuam os termos índices $\mathbf{q}_1$ distante no máximo 5 palavras de $\mathbf{q}_2$, nessa mesma ordem dos termos, ou seja, uma expressão do tipo $\mathbf{q}_2$ ADJ5 $\mathbf{q}_1$ produz um resultado completamente diferente da consulta anterior. Já quando utilizamos o operador NEAR, a ordem dos termos na expressão lógica não altera o resultado, ou seja, uma consulta formulada por uma expressão lógica do tipo $\mathbf{q}_1$ NEAR5 $\mathbf{q}_2$, recupera todos os documentos que possuem os termos índices $\mathbf{q}_1$ distante no máximo 5 palavras de $\mathbf{q}_2$, independente da ordem dos termos na expressão lógica. Portanto,

$\mathbf{q}_1$ NEAR5 $\mathbf{q}_2 = \mathbf{q}_2$ NEAR5 $\mathbf{q}_1$.

A combinação entre operadores lógicos booleanos e os operadores lógicos de proximidade permite melhorar os aspectos de restritividade e abrangência das consultas nos modelos booleanos. Contudo, Blair (1990) não acredita que essa combinação possa trazer vantagens que superem as limitações do modelo booleano.

Para Sparck-Jones (1997), os principais problemas do modelo booleano são:

(a) normalmente, o usuário não possui treinamento apropriado, tendo dificuldade em formular consultas usando operadores lógicos;

(b) há pequeno controle sobre o tamanho da saída produzida por uma determinada consulta;

(c) a recuperação lógica resulta em uma simples partição da coleção de documentos em dois subconjuntos discretos: os registros que satisfazem a consulta e os que não a satisfazem.

O sistema de recuperação de informação de jurisprudência do Tribunal Regional Eleitoral do Distrito Federal é baseado no modelo lógico e, portanto apresenta os problemas descritos por Sparck-Jones (1997). Sendo considerado o problema (c) um fator determinante para as dificuldades encontradas no sistema atual de recuperação de jurisprudência, na medida em que o modelo lógico considera a correspondência exata entre os termos da consulta e os termos de indexação dos documentos da coleção. Por outro lado, no sistema de recuperação de jurisprudência a consulta deve ser formada a partir de uma expressão (texto linguagem

natural) que determina uma correspondência inexata entre os termos da consulta e os termos descritores dos documentos. Por isso, avaliar um novo modelo de recuperação de informação jurídica no âmbito do Tribunal Regional Eleitoral do Distrito Federal pode contribuir para melhorar o grau de relevância dos documentos recuperados pelas buscas realizadas por usuários do seu serviço de consulta a jurisprudência eleitoral.

3.2.2 Modelo Dinâmico

Esses modelos apresentam como principais características a participação do usuário na definição da representação dos documentos, permitindo uma adaptação dos documentos recuperados aos interesses dos usuários (Ferneda, 2003). Rijsbergen (1979), Lancaster e Fayen (1973, p.36), Baeza-Yates e Ribeiro-Neto (1999) definem como *relevance feedback* o processo de interferência do usuário na representação dos documentos relevantes para satisfazer suas necessidades de informação durante o processo de busca. Os modelos que mais se adéquam ao *relevance feedback* são aqueles classificados por Ferneda (2003) como dinâmicos e enunciados por Rijsbergen (1979), Baeza-Yates e Ribeiro-Neto (1999) como modelo alternativo que representa evolução dos modelos clássicos, principalmente o modelo probabilístico e o algébrico que dão origem respectivamente, à aplicação das teorias de rede bayesiana (*bayesian network*) e dos modelos de indexação semântica latente e rede neural na recuperação da informação. Esses são modelos cuja teoria é tratada no universo da inteligência artificial, mais especificamente pelos sistemas especialistas.

Sistemas especialistas são sistemas computacionais dedicados a uma determinada área do conhecimento que procura reduzir o conhecimento de um especialista em um determinado assunto a um conjunto de regras que regem as relações de tomada de decisão e resolução de problemas relacionados com um domínio específico. São estruturados nas seguintes partes:

- Base de conhecimento
    - Conjunto de regras
    - Memória de trabalho
- Motor de inferência

Alguns modelos de representação do conhecimento são utilizados para que se possa construir a base de conhecimento. Os principais modelos de representação do conhecimento são:

- As redes semânticas: Proposta por Quillian em 1968 é formada a partir de nodos e arcos que interconectam os nodos estabelecendo uma relação semântica entre os nodos interconectados. As redes semânticas foram desenvolvidas objetivando a implementação computacional eficiente das possibilidades de relações (interconexões) existentes entre os termos de um domínio que possam ser criadas pelos mecanismos de cognição humanos na solução de problemas. (SOWA, 2002).
- Os frames**:** Minsky (1975) introduziu o modelo de frames para a representação do conhecimento. Um frame é uma coleção de atributos, chamados de slots, e valores, que descrevem alguma entidade do mundo (RICH e KNIGHT, 1993). Os frames integram conhecimento declarativo sobre objetos, eventos e conhecimento procedimental relacionados a como recuperar informações ou calcular valores. Os atributos também apresentam propriedades, que dizem respeito ao tipo de valores e às restrições de numero que podem ser associadas a cada atributo. Essas propriedades são chamadas *facetas*. Assim como nas redes semânticas, uma das características nos frames é a possibilidade de que sejam criados novos subtipos de objetos que herdem todas as propriedades da classe original. Essa herança é bastante usada tanto para a representação do conhecimento como para a utilização de mecanismos de inferência.
- Pacotes de Organização de Memória: Schank (1982) consolidou a teoria da Memória Dinâmica que estruturava e representava o conhecimento sob a forma de Pacotes de Organização de Memória (MOP – Memory Organization Packages). MOPs são usados para representar o conhecimento sobre classes e eventos. (Riesbeck e Schank, 1989). Um pacote de organização de memória possui um conjunto de normas que representam as suas características básicas. As conexões existentes entre esses pacotes de organização determinam quais e quando as informações existentes em memória dinâmica estarão disponíveis. A teoria da memória dinâmica constituiu a raiz para o desenvolvimento da técnica de raciocino baseado em casos (CBR - Case-based resoning). "O raciocínio baseado em casos é uma abordagem para solução de problemas e para o aprendizado com base em experiências passadas." (WANGENHEIM e WANGENHEIM, 2003, p.1). Abel (1995), Bueno (1999), Weber-Lee (1998) e

Braga Júnior (2001) preconizam o raciocínio baseado em casos como modelo de recuperação da informação aplicado à área temática do direito.

Existem outras técnicas de representação do conhecimento, mas a essa pesquisa interessa apenas aquelas que dizem respeito ao problema estudado, dessa forma vale ressaltar o modelo de motor de inferência que utiliza o raciocínio baseado em casos na construção de sistemas de recuperação da informação. O motor em foco difere dos demais motores de inferência construídos sob a influência da disciplina da inteligência artificial para sistemas de recuperação da informação por se basear em casos concretos passados armazenados em sua base de conhecimento (base de casos passados) e não em regras de reducionistas de um determinado domínio, como ocorre nos modelos de recuperação de informação baseados em sistemas especialistas.

Wangenheim e Wangenheim (2003, p.2) apresentam argumento que dão validade ao modelo de recuperação de informação baseado em casos quando esclarecem que

> "Com seu enfoque na utilização de experiências, o raciocínio baseado em casos diferencia-se radicalmente de outras metodologias para desenvolvimento de programas e sistemas da área da inteligência artificial. Ao contrário de enfoques tradicionais para encontrar uma solução para um problema em inteligência artificial, em que se descreve conhecimento genérico na forma de regras, quadros, roteiros, etc., no raciocínio baseado em casos é o conhecimento específico, na forma de exemplos concretos, que se encontra no centro do processo de solução de um problema."

Portanto, essa visão abre espaço para sustentação de novos modelos dinâmicos de recuperação de informação pautados na inteligência artificial. Nesse sentido, as soluções reducionistas, onde o processo de resolução de um problema é feito por meio de um sistema baseado em regras, como nos sistemas especialistas tradicionais, evolui para um processo de solução de problemas no qual o enfoque está baseado em casos concretos.

Nesse capítulo foram apresentados os dois modelos de recuperação de informação mais relevantes a essa pesquisa, já que o modelo lógico serviu de base para construção do sistema de recuperação de informação atualmente utilizado pelo Tribunal Regional Eleitoral do Distrito Federal, conforme apresentado no referencial teórico dessa dissertação. Por conseqüência, o sistema atual de recuperação de informação do Tribunal Regional Eleitoral do Distrito Federal apresenta todos os problemas apontados para o modelo lógico de recuperação da informação. Paralelamente, é apresentado o modelo de recuperação de informação baseado em casos como uma alternativa, entre os modelos dinâmicos, para solucionar os problemas dos sistemas baseados no modelo lógico, apontados por

pesquisadores da área e descritos nessa pesquisa. Baeza-Yates e Ribeiro-Neto (1999), ressaltam a necessidade de testar o desempenho de novos modelos de recuperação da informação, objetivo para o qual essa pesquisa procura contribuir.

4. Avaliação de Sistemas de recuperação de informação

Na medida em que se popularizou o uso do computador em ambiente empresarial, surgiram novos sistemas de recuperação de informação utilizando-se de diferentes métodos de ranking e indexação existentes. Tal fenômeno determinou a importância de identificar quais sistemas eram capazes de recuperar um maior número de documentos relevantes ou, ainda, qual deles tinha melhor precisão no atendimento às necessidades de informação expressas previamente pelo usuário. Nesse contexto surge a necessidade de se avaliar esses sistemas.

Rijsbergen (1979) afirma que na literatura aparecem constantemente novas abordagens metodológicas para o tema, entretanto optou por abordar apenas os métodos convencionais. O texto desse autor é considerado por diversos pesquisadores como um dos pioneiros na área. Apesar de ter sido editado há quase 40 anos, o seu conteúdo ainda é considerado atual e relevante.

De acordo com Baeza-Yates e Ribeiro Neto (1999) o tipo de avaliação a ser considerado depende dos objetivos do sistema de recuperação. Afirmam que o primeiro tipo de avaliação a ser considerada é a análise funcional, onde as funcionalidades específicas do sistema são testadas uma a uma. Após essa fase, realiza-se, então, a análise de desempenho do sistema na qual as medidas mais comuns como tempo e espaço são avaliadas, dentre outras.

Muitos esforços vêm sendo desenvolvidos com o objetivo de buscar alternativas para a solução do problema da avaliação dos sistemas de recuperação de informação, entretanto, os profissionais que atuam na área acreditam que esta solução ainda está longe de ser alcançada.

Harter e Hert (1997) consideram o final da década de 90 como um marco histórico da recuperação de informação em função da rápida mudança nesses sistemas provocada pela influência de outros campos de pesquisa como a lingüística e a inteligência artificial. Essas áreas contribuem para a proliferação de visões alternativas ao processo de avaliação de sistemas de recuperação de informação e conduzem a pesquisa a novos problemas, tais como: o que deve ser entendido por avaliação? Que componentes devem ser considerados na avaliação de um sistema de recuperação de informação? Os usuários, suas cognições e seu ambiente de trabalho devem ser incluídos na avaliação? Como avaliar sistemas de recuperação de informação que atendem a usuários reais, utilizando-se um modelo baseado em sistemas experimentais?

4.1. Histórico e estudos de laboratório

De acordo com Ingwersen (2002), desde os anos 60 uma quantidade razoável de pesquisas na área de recuperação de informação foi relacionada com estudos de desempenho, e os métodos mais comumente aplicados foram os testes de laboratório realizados em bases de dados desenvolvidas com este propósito. O pioneiro desses testes foi desenvolvido por Cyril W. Cleverdon (1964 apud LANCASTER e FAYEN, 1973, p. 125), que realizou o primeiro experimento de avaliação de recuperação de informação de dados do catálogo aerodinâmico da College Aeronautics de Cranfield na Inglaterra, quando definiu critérios para a avaliação de linguagens documentárias:

(a) Cobertura: extensão da coleção utilizada pelo sistema de recuperação para realização da busca;

(b) Revocação: proporção entre documentos relevantes recuperados e documentos relevantes existentes na coleção pesquisada;

(c) Precisão: proporção entre os documentos relevantes recuperados e o total de documentos obtidos em uma busca;

(d) Tempo de resposta: intervalo médio entre o momento de realização da busca e a resposta obtida;

(e) Esforço do usuário: esforço realizado pelo usuário na obtenção de respostas precisas para sua busca;

(f) Forma da resposta (output): adequação da forma de apresentação do resultado;

De acordo com Ingwersen (2002) o princípio é científico, ou seja, para cada coleção testada, um número fixo de perguntas é utilizado e um número total de documentos relevantes objetivamente para cada questão é conhecido pelo pesquisador. Variáveis nesses testes são também as técnicas específicas de recuperação de informação ou os métodos específicos de representação, por isso estudos comparativos de análise de desempenho podem ser desenvolvidos. As coleções de teste são geralmente pequenas, variando entre 3.000 a 20.000 itens, valores que estão distantes de sistemas de grande escala com mais de 9.000.000 de registros de documentos.

Ainda de acordo com Ingwersen (2002) os experimentos no laboratório não contavam com a participação de humanos, e esse é o motivo principal para considerar que esses experimentos não poderiam ser considerados representativos das situações de um usuário em suas atividades cotidianas de recuperação da informação. Abordagens mais recentes para o planejamento de sistemas de recuperação de informações naturalmente levam

em conta os efeitos cognitivos da dinâmica interativa e do feedback sobre o intervalo entre o problema do usuário, sua solicitação e a estrutura da pergunta. As perguntas fixas não são mais constantes, mas se tornam variáveis, com perda do conhecimento do número total de textos relevantes como conseqüência. As coleções de testes e os métodos comparativos são conseqüentemente somente operacionais em relação à abordagem tradicional.

De acordo com Baeza-Yates e Ribeiro-Neto (1999) em testes de laboratório, as avaliações realizadas são totalmente diferentes das situações da vida real. Os autores afirmam ainda que os primeiros testes voltavam-se para pesquisas *in batch*[5], e que, a partir da década de 90, houve uma maior atenção para a avaliação de experimentos reais. Apesar dessa tendência, as pesquisas em laboratório ainda continuaram devido à possibilidade de serem repetidas e à escalabilidade proporcionada pelo ambiente fechado de um laboratório.

4.2 Avaliação

De acordo com Lancaster (2004), num sentido geral, a avaliação é o ato de medir o valor de uma atividade ou objeto. Num sentido mais específico, a avaliação é um ramo da pesquisa – a aplicação do método científico para determinar, por exemplo, a qualidade do desempenho de um programa. A avaliação tem como objetivo reunir dados úteis para subsidiar atividades que buscam solucionar problemas ou apoiar a tomada de decisão.

Avaliação, de acordo com Hernom e McClure (1990), constitui-se em um processo de identificação e coleta de dados a respeito de um serviço ou atividade específica, estabelecida por um critério pelo qual o sucesso pode ser estimado, determinando tanto a qualidade do serviço ou atividade quanto o seu grau de acerto na busca dos objetivos ou metas previamente definidas.

4.3 Relevância

Relevância é um conceito inerentemente subjetivo, pois está relacionado diretamente com a satisfação das necessidades humanas, e com o julgamento dos usuários sobre o quão bem um documento recuperado satisfaz as suas necessidades de informação. Os humanos discordam quanto à relevância de um documento para satisfazer a uma determinada solicitação, pois o que pode ser relevante para um indivíduo pode não ser para outro, mesmo que a pergunta seja a mesma. O que pode ser relevante para um usuário em um dia pode não ser mais no dia seguinte. Um documento pode ser relevante se recuperado de uma

[5] Consultas realizadas utilizando-se um grupo de perguntas submetidas a uma base de dados de conteúdo estático

determinada coleção e não quando recuperado em outra coleção. A ordem de recuperação também afeta a relevância, pois o segundo documento recuperado será menos relevante que o primeiro, caso esse tenha atendido à necessidade do usuário. Assim, a relevância depende da pergunta, da coleção, do contexto, das necessidades pessoais dos usuários, de suas preferências, conhecimento, linguagem etc. (GREENGRASS, 2000).

Greengrass (2000) destaca que há diferença entre a relevância de um tópico para atender a uma determinada pergunta e a usabilidade para o usuário que fez a pergunta, ou seja, relevância está relacionada com a pergunta do usuário e a pertinência está relacionada com a necessidade do usuário.

De acordo com Ingwersen (2002), relevância é definida como a medida ou grau de correspondência ou utilidade existente entre um texto ou documento e uma pergunta ou uma solicitação de informação determinada por um indivíduo.

Saracevic (1975) discute as bases para entender relevância na recuperação de informação e formulou seguinte expressão: Relevância é o (a) A de um(a) B entre um(a) C e um(a) D conforme determinado por um(a) por E. Nesta expressão, cada incógnita pode ser substituída por um dos termos presentes no quadro a seguir:

A	B	C	D	E
medida grau dimensão estimativa avaliação relação	correspondência utilidade conexão satisfação ajuste cruzamento	documento artigo forma textual referência informação oferecida fato	pergunta requisição necessidade do usuário ponto de vista	pessoa julgador usuário solicitante especialista de informação

Quadro 1 – Combinação das expressões de relevância de Saracevic

Fonte: Saracevic (1975)

Rijsbergen (1990 apud INGWERSEN, 2002) enfatiza que se um documento contém informação sobre X então será relevante para X. O processo de localizar documentos relevantes é inerentemente incerto e também altamente dependente do contexto (*aboutness*[6]). Esse problema de relevância refere se ao texto ou a imagem com a tematicidade (*aboutness*) X, como exemplo, Rijsbergen (1979) cita um trecho do livro de Mark Twain, onde o General Lee está sentado num cavalo, nesse contexto, a informação X pode ser: Lee gosta de cavalgar, mas esse fato somente pode ser estabelecido por inferência, através da análise do contexto, agregando valores semânticos e por meio do conhecimento sobre o autor. Por isso, relevância é um valor de natureza pragmática, ligado ao problema de espaço e estado do conhecimento

[6] Termo utilizado em lingüística, ciência da informação e biblioteca como sinônimo de assunto (discurso)

individual do usuário. Isso produz um problema metodológico em relação aos modelos de avaliação para recuperação de informação como um todo (INGWERSEN, 2002).

Uma característica fundamental e intrínseca da recuperação de informação é a efetividade – dentro dos limites das teorias apresentadas e dos modelos de recuperação de informação – que está longe de 100%. Um problema que se deve ter em mente é que o pesquisador não tem como saber que documentos eram relevantes e não foram recuperados, e nunca saberá. É essa a incerteza da recuperação enfatizada por Rijsbergen (1979) e que é inerente do ponto de vista cognitivo, que deve constituir apenas do conjunto original de princípios da ciência da informação (INGWERSEN, 2002).

Esse é essencialmente um problema crucial para a recuperação de informação que não pode ser observado por meio de instrumentos, mas que pode ser detectado indiretamente, ou seja, em recuperação de informação podemos observar o que estamos recuperando, mas todos os experimentos e observações nos informam que alguma informação relevante potencial para uma dada pergunta fica fora do sistema de informação, são informações que não podemos alcançar. A proporção de revocação nos diz sobre o tamanho do buraco negro, mas não sua natureza. A partir do ponto de vista cognitivo o buraco negro se manifesta de duas formas: primeiro como informação potencial a qual explicitamente contém tópicos de assunto relevantes em um dado momento, mas não é recuperado devido à deficiência, por exemplo, de um tesauro apropriado; segundo como aquela informação potencial a qual está implicitamente presente no documento, ou seja, aquela que Rijsbergen (1979) chamou de informação X que não está diretamente contida no documento, mas por meio da percepção pode ser "lida dentro dela" pela interpretação do recipiente (INGWERSEN, 2002).

4.3.1 Relevância lógica

Rijsbergen (1979) argumenta que existe um conceito de relevância que pode ser considerado objetivo, é o conceito denominado relevância lógica, onde a relevância é definida em termos de conseqüência lógica. Explica que, em termos gerais, é pouco utilizado, porém de grande importância para o desenvolvimento de sistemas do tipo pergunta e resposta. A relevância lógica é mais facilmente explicada, onde as perguntas são representadas por um conjunto de sentenças, nesse caso por dois tipos sim (p) e não (não-p). As questões devem ser apresentadas na forma afirmativa e não interrogativa. Ex.: O hidrogênio é um elemento halogênio. O hidrogênio não é um elemento halogênio. Se as duas afirmativas que representam a questão são denominadas relação de componente, então o subconjunto do

conjunto de sentenças armazenadas é o conjunto de premissas para a relação de componente, se e somente se, a afirmativa do componente é uma conseqüência lógica desse subconjunto.

O conjunto de premissas mínimo para a afirmativa de componente é um que é tão pequeno quanto possível no sentido de que se seus membros forem apagados, a relação de componente não deve mais ser um conjunto resultante da conseqüência lógica.

4.3.2 Relevância Situacional

A relevância situacional provê uma forma mais específica para o usuário julgar a relevância à medida que examinam os resultados intermediários da busca (MARCHIONINI, 1992). De forma prática, a relevância situacional trata a avaliação da recuperação de informação sob o ponto de vista das ações realizadas pelo usuário no processo de busca. Essas ações incluem: terminar a busca porque o objetivo foi alcançado; examinar o documento mais detalhadamente; registrar a existência e localização do documento e continuar examinando outros resultados para depois voltar ao anterior para exame mais detalhado; examinar mais detalhadamente outras implicações do documento para a continuação da busca; continuar examinando outros resultados nessa interação; formular nova questão ou redefinir o problema; rejeitar o documento por completo e continuar examinando os resultados e rejeitar o documento e parar a busca de informação sem terminar a tarefa.

Dessa forma, a relevância situacional funciona como critério para computar medidas de performance, como precisão e revocação, em um processo de avaliação de sistemas de recuperação de informação centrados no usuário.

Segundo Ingwersen (2002), a teoria cognitiva da recuperação da informação, como denomina a recuperação da informação interativa, isto é, baseada na relevância situacional, amplia a base do modelo tradicional de recuperação da informação, agregando, a este, aspectos que influem no estado cognitivo do usuário, como suas herança cultural, os seus objetivos pretendidos, suas áreas de interesse, entre outros aspectos.

4.4 Paradigmas da avaliação de sistemas de recuperação de informação

Segundo Harter e Hert (1997), estudos de vários autores afirmam de forma persuasiva que o modelo de avaliação de Cranfield (CLEVERDON, et al. 1966), no senso Kuhniano (KHUN, 1962-1970; MASTERMAM, 1970), transformou-se em um paradigma para as pesquisas de avaliação de sistemas de recuperação de informação. Ellis (1992, apud CAPURRO 2003) classificou o modelo de Cranfield como paradigma físico devido a sua relação íntima com a teoria da informação de Claude Shannon e Warren Weaver (1949-1972).

Em essência esse paradigma estabelece um modelo experimental onde os principais componentes são a coleção de documentos. Em Cranfield foram 1400 documentos do catálogo de dados de engenharia aeronáutica, um conjunto de consultas (queries) e a definição de documentos julgados relevantes para responderem às perguntas selecionadas. Nesse paradigma o julgamento de relevância era binário, ou seja, cada documento recuperado por uma consulta (query) era julgado como relevante ou não relevante. A partir daí, para cada ação de recuperação (consulta, coleção de documentos, documentos recuperados e julgamento de relevância) os seguintes resultados eram obtidos:

- Documentos relevantes recuperados (a);
- Documentos não relevantes recuperados (b);
- Documentos relevantes não recuperados (c);
- Documentos não relevantes não recuperados (d).

Dessa forma, o desempenho de um sistema de recuperação de informação (neste paradigma o sistema de recuperação de informação é considerado como sistema de indexação), para uma dada ação de recuperação, pode ser avaliado examinando-se o resultado produzido por ele e computando-se medidas baseadas nesse resultado. Neste paradigma as principais medidas de desempenho são a precisão (Precision) = a/(a+b) e a revocação (Recall) = a/(a+c)

Depois de estabelecido o paradigma físico da avaliação de sistemas de recuperação de informação, muitos trabalhos discutiram o modelo tradicional de avaliação desses sistemas, considerando a forma como esse modelo foi estabelecido. O principal trabalho, denominado *Information Retrieval Experiment* foi publicado por Sparck Jones (1981). Além deste, Harter e Hert (1997) sugerem duas publicações que apresentam estudos que colocam em questão os métodos de avaliação propostos pelo paradigma físico, o *Information Processing and Management*, por Harman (1992) e o *Journal of the American Society for Information Science*, por Tague-Sutcliffe (1996). Essas publicações deram início à fase pré-paradigmática (Kuhn, 1962-1970) do modelo cognitivo. Recentes trabalhos contrastam o paradigma tradicional com o cognitivo, também conhecido como paradigma comportamental, alternativo, ou orientado ao usuário. Entre eles, merecem destaque as publicações de Dervin e Nilan (1986), Ellis (1996) e Ingwersen (1992).

4.5 Critérios de avaliação

Rijsbergen (1979) apresenta três perspectivas para análise do problema da avaliação do desempenho dos sistemas de recuperação de informação na forma de três perguntas: Por que avaliar? O que avaliar? Como avaliar?

a) Por que avaliar?

Quanto a esta perspectiva o autor evidencia que a sua resposta está focada em dois aspectos. Primeiramente, os aspectos sociais, que são geralmente intangíveis, devido à dificuldade de mensurar os benefícios obtidos da utilização de um sistema de recuperação de informação, pois os resultados são difíceis de serem interpretados. Em outros sistemas de recuperação de informação o benefício pode ser mais mensurável que em outros. Já os aspectos econômicos avaliam o montante de recursos a ser utilizado, o custo de aquisição de equipamentos, computadores etc. são relativamente fáceis de serem estimados e mensurados, porém estimar os custos referentes aos esforços dos indivíduos é mais difícil, ainda mais que os custos podem ou não depender do usuário individual.

Isso torna claro que na avaliação de sistemas de recuperação de informação a preocupação está centrada no fornecimento de dados ao usuário, assim os usuários podem tomar a decisão de querer o sistema ou pagar por ele. Esses métodos são utilizados de forma comparativa para medir quando alterações no sistema trarão aperfeiçoamentos ou melhorias de desempenho, ou seja, quando uma solicitação é feita para uma estratégia particular de pesquisa, o padrão de medida da avaliação pode ser aplicado para determinar quando a solicitação é válida.

b) O que avaliar?

Para responder essa pergunta faz-se necessário retomar o conceito de sistema de recuperação de informação, ilustrado na figura a seguir.

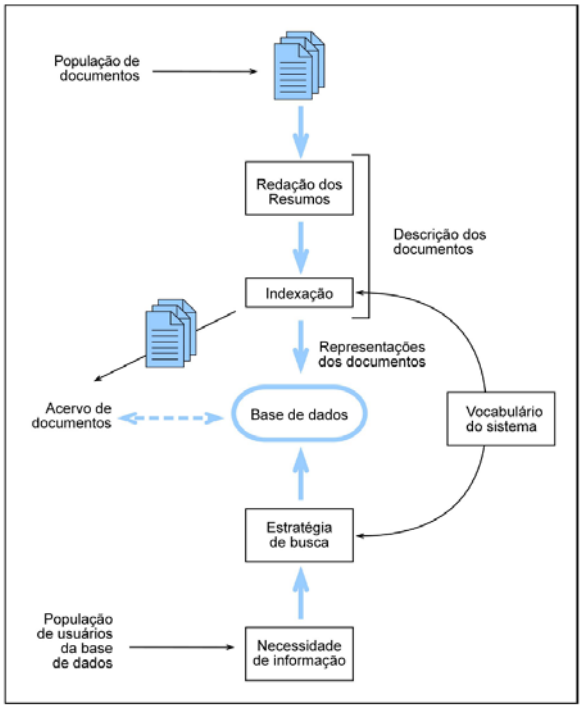

Figura 13 – Processo geral de recuperação da informação

Fonte: Adaptado Beza-Yate e Ribeiro-Neto (1999)

Na figura 13 identifica-se, claramente, os elementos que compõem esse sistema. Destacam-se: a descrição (organização) dos documentos e armazenamento em base de dados; a construção de vocabulário (controlado) dos termos que compõem o tema da base de dados (em se tratando de bases de dados multidisciplinar, caso mais comum na Web, essa etapa não compõe o sistema); a estratégia de busca é outro componente importante da recuperação da informação, nessa etapa os usuários representam, geralmente em linguagem natural, as suas necessidades de informação; a interação do usuário com o sistema de recuperação de informação realizada por meio de uma interface gráfica e, ainda, o motor de busca desenvolvido que realiza a tarefa de recuperação dos itens armazenados nas bases de dados.

Estes componentes podem ser vistos como subsistemas do processo de recuperação da informação e que, portanto, podem ser avaliados separadamente, dependendo do objetivo da avaliação.

Aires (2002) propõe como objetivo da avaliação de sistemas de recuperação de informação uma análise da efetividade sob a seguinte ótica:

- Fazer uma avaliação do sistema
- Fazer uma avaliação dos usuários
- Fazer uma avaliação do sistema pensando nos usuários

Avaliar a recuperação sob a ótica do sistema requer analisar a forma de descrição (indexação) dos documentos, o motor de busca (algoritmos de recuperação de informação) e a estratégia utilizada. Geralmente essa análise por meio de comparações dos resultados de um conjunto de consultas submetidas a diferentes algoritmos de recuperação utilizando estratégias semelhantes, de forma similar aos testes realizados em Cranfield, e, mais atualmente, nas TREC (*Text REtrieval Conferences*). As principais medidas de efetividade são: a precisão e a revocação (ambas serão discutidas no tópico "como avaliar"). Fazer uma avaliação do sistema de recuperação de informação pode ser importante na busca da melhoria da qualidade técnica do sistema. A principal crítica à avaliação sob essa ótica é que não inclui os aspectos intrínsecos do usuário do sistema como um componente participante da avaliação.

A partir da abordagem orientada ao usuário articulada por Brenda Dervin e Michael Nilan muitos estudos foram promovidos num esforço para incorporar o usuário no processo de avaliação do sistema de recuperação de informação, considerando a construção de um modelo de avaliação sob a perspectiva do usurário. Os principais objetivos dessa avaliação são responder às seguintes questões: após um processo de busca: as necessidades do usuário foram atendidas? A informação recuperada é útil ou não? A interface é amigável? Para avaliar o grau de efetividade das respostas e essas perguntas Harter e Hert (1997) sugerem dois conjuntos básicos de medidas mensuráveis qualitativamente: medidas ligadas à percepção e atitude do usuário; medidas que capturem a interação do usuário com o sistema. No primeiro conjunto estão as medidas de satisfação, que requerem um maior aprofundamento dos estudos, uma vez que não possuem, ainda, uma forma de medida bem estabelecida, e a medida de utilidade que segundo Belkin e Vickery (1985) é uma medida que determina que um sistema de informação deve ser avaliado com base no quão proveitoso é, para o usuário, o resultado produzido pelo sistema. Fazem parte do segundo conjunto de medidas de avaliação do usuário a informatividade definida por Tague-Sutcliffe (1996) como a chave para avaliar o valor de um sistema de informação, considerando a informação dependente da conceitualização e entendimento por parte do ser humano; outra medida é a usabilidade de um sistema de recuperação de informação que avalia a interação homem-máquina. Sweeney et al. (1993) definem a usabilidade como medida que avalia a capacidade de um recurso de tecnologia da informação permitir ao usuário interagir de forma eficiente e satisfatória com sistema de informação para realização de uma determinada tarefa em um determinado ambiente e a um custo acessível. A principal vantagem da avaliação sob os aspectos do usuário é a capacidade de se conhecer as suas necessidades de informação de forma a permitir a construção de novos sistemas de Ri que possuam interfaces de interação gráficas mais

amigáveis e outros recursos que possam ajudar aos usurários a encontrar documentos relevantes. Mas, apesar das vantagens desse tipo de avaliação de sistemas de Ri, eles são, em geral, focados em públicos específicos, não sendo relacionados a públicos grandes e diferenciados.

A avaliação do sistema com foco nos usuários tem como propósito geral aumentar as chances de um determinado sistema de recuperação de informação ser adotado e utilizado. Só é possível realizá-la com avaliação de sistemas de informação associada à análise de usuários e, ainda, unindo-se as pesquisas de laboratório às situações reais, contextos, indivíduos e organizações. Nessa avaliação a performance é medida através de *benchmarks* [7]e considera as características do ambiente onde é realizada a avaliação,o que restringe os resultados obtidos por esse tipo de avaliação apenas ao ambiente em que a performance foi medida. Não há muitos estudos relacionados a avaliação de performance e, por conseqüência, não há uma técnica de avaliação que proporcione uma avaliação completa do sistema com foco no usuário. Não existem *benchmarks* para sistemas de recuperação de informação web e nem para sistemas em diferentes línguas. Para realizar essa avaliação a medida utilizada deve verificar como toda esta informação, associada ao sistema de recuperação de informação, afeta o trabalho, lazer, sociedade e cultura do usurário.

c) Como avaliar?

A forma de avaliação de um sistema de recuperação de informação inicia-se com a discussão dos fatores que afetam o julgamento de desempenho de um sistema. Para isso é necessário considerar certos critérios pelos quais a performance de um sistema é julgada, ou seja, é necessário conhecer algumas necessidades dos usuários em relação ao sistema de recuperação de informação. Esses critérios são aqueles parâmetros que refletem a habilidade do sistema em satisfazer a necessidade de informação do usuário.

4.6 Medidas de avaliação

Como foi visto no tópico anterior, o uso das medidas de avaliação do sistema depende de qual componente do sistema se está avaliando. A avaliação do sistema utiliza métricas quantitativas, por outro lado, a avaliação centrada no usuário do sistema utiliza métricas qualitativas. Lancaster e Fayen (1973, appud Cleverdon 1964) apresentam critérios para avaliação do sistema de recuperação de informação composta por 6 itens mensuráveis quantitativamente:

[7] Um padrão pelo um determinado processo pode ser medido, julgado ou avaliado

- Cobertura da coleção – extensão que o sistema inclui material relevante;
- Tempo de resposta – intervalo de tempo entre a solicitação da informação e a resposta fornecida;
- Forma de apresentação do resultado – forma pela qual o sistema apresenta o resultado de uma busca;
- Esforço empreendido pelo usuário para obter a informação – em especial no trabalho de separação dos itens relevantes recuperados dos itens não relevantes;
- Revocação – proporção entre os materiais relevante recuperado e o total de materiais relevantes existente na coleção, em resposta à solicitação;
- Precisão – proporção entre os materiais recuperados que são relevantes e o total de materiais recuperados em resposta a uma solicitação.

Estes dois últimos, revocação e precisão, são utilizados como medidas de efetividade do sistema, ou seja, a medida da capacidade do sistema de recuperar documentos relevantes, enquanto retém os não relevantes. Quanto mais efetivo for o sistema mais deverá atender às necessidades dos usuários.

A técnica de medir a efetividade do sistema é influenciada pela estratégia de recuperação adotada e pelo formato de saída. Este tema é tratado praticamente em todo o texto de Rijsbergen (1979).

Para Ingwersen (2002) os critérios padrão de avaliação são revocação e precisão. Estas medidas são utilizadas na avaliação de sistemas de recuperação de informação utilizando como base o conceito de relevância do resultado de uma busca. Greengrass (2000) ressalta que outras medidas foram propostas, porém não foram amplamente utilizadas.

4.6.1 Revocação (Recall) e precisão (precision)

Revocação, de acordo com Ingwersen (2002), é definida como a razão entre o número de documentos relevantes recuperados (R), em relação ao número total de documentos relevantes existentes na coleção (C). A revocação, conseqüentemente, só pode ser calculada se o número exato de (C) for conhecido, o que é normalmente improvável. Desse modo, resulta em um nível de incerteza, em adição à incerteza inerente à recuperação de informação em si mesma. Saracevic (1995 apud GREENGRASS, 2000) define revocação como a razão de itens relevantes recuperados em relação a todos os itens relevantes em um arquivo ou coleção, ou a probabilidade de um item ser relevante e ser recuperado.

A seguir, é apresentada a fórmula de revocação definida por Ingwersen (2002):

$$\frac{\text{R – nº de documentos relevantes recuperados}}{\text{C – nº total de documentos relevantes existentes na coleção}}$$

Fórmula do índice de revocação na recuperação de documentos

Fonte: Adaptado de Ingwersen (2002)

De acordo com Cyril W. Cleverdon, (1964 apud LANCASTER e FAYEN, 1973, p. 125) a revocação é a taxa referente à relação entre o número de documentos relevantes recuperados e o número total de documentos relevantes contidos no sistema. Para medi-la é necessário conhecer o número total de documentos relevantes contidos no sistema.

Baeza-Yates e Ribeiro-Neto (1999) apresentam a definição de revocação como a fração de documentos relevantes de um conjunto, do qual foram recuperados.

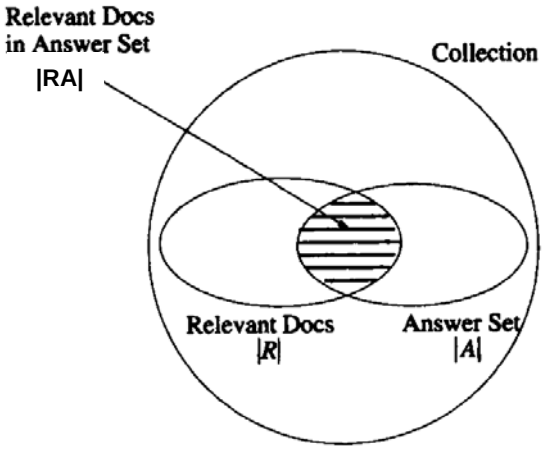

Figura 14 – Representação gráfica da revocação

Fonte: Baeza-Yates e Ribeiro-Neto (1999)

Precisão, de acordo com Ingwersen (2002), é a relação entre o número de documentos relevantes recuperados (R) e o total de documentos recuperados (L). Saracevic (1995 apud GREENGRASS, 2000) define precisão como a razão entre itens relevantes recuperados e todos os itens recuperados, ou a probabilidade deste item ser recuperado ou de ser relevante.

A seguir, é apresentada a fórmula de precisão definida por Ingwersen (2002):

$$\frac{\text{R – nº de documento relevantes recuperados}}{\text{L – nº de documentos recuperados}}$$

Fórmula do índice de precisão na recuperação de documentos

Fonte: Adaptado de Ingwersen (2002)

De acordo com Cleverdon (1964 apud LANCASTER e FAYEN, 1973, p. 125), a precisão é expressa como a taxa entre o número de documentos relevantes recuperados e o número total de documentos recuperados, sendo necessário avaliar a relevância dos documentos recuperados. Esse é um parâmetro fundamental para a avaliação de sistemas de busca. Para avaliar a precisão é importante que o sistema informe o número total de documentos recuperados.

De acordo com Baeza-Yates e Ribeiro Neto (1999) precisão é a fração entre documentos relevantes e o total de documentos recuperados. Afirmam que, para atender às definições apresentadas, é necessário que todos os documentos que respondem à pergunta sejam examinados. Entretanto, o usuário não é apresentado ao conjunto total de documentos que respondem à sua pergunta. Em vez disso, esse conjunto é ordenado de acordo com um nível de relevância, o qual o usuário examina a partir do início da lista.

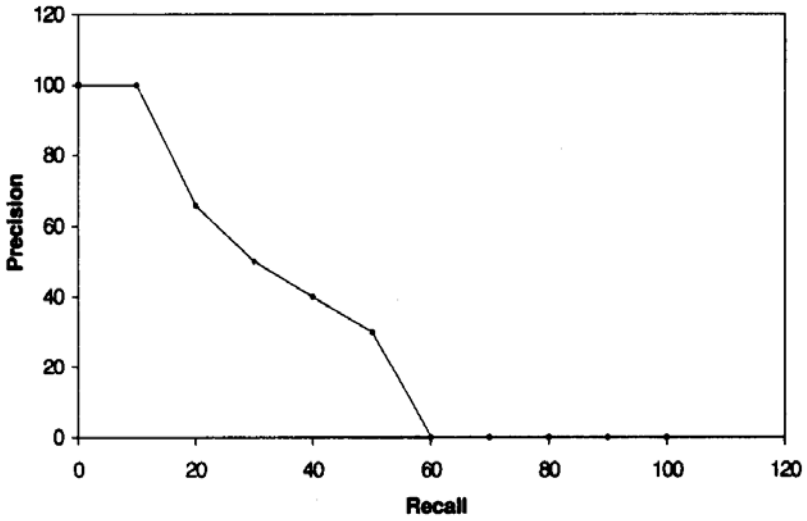

Figura 15 – Gráfico comparativo entre precisão e revocação

Fonte: Baeza-Yates e Ribeiro-Neto (1999)

4.6.2 Técnicas de recuperação de informação

Ingwersen (2002) discute que a definição das medidas de desempenho padrão, precisão e revocação, permitiu comparar as várias técnicas de recuperação de informação. A técnica do uso de probabilidades é uma técnica mais baseada em características, especialmente quando incorpora os pesos. De acordo com Belkin e Croft (1987 apud INGWERSEN, 2002) o uso de termos dependentes para modificar categorias de documentos pode também melhorar o desempenho, mas somente se as dependências são exatamente identificadas pelos usuários ou pelo processamento de linguagem natural. A aplicação

automática de tesauros para expandir as pesquisas somente será realmente efetiva se os termos expandidos e o tipo de tesauro utilizado for rigidamente controlado. Técnicas de *clustering* [8] podem alcançar níveis de desempenho semelhantes às técnicas de recuperação de informação baseada em características, mas tende a ser melhor para resultados de alta precisão. Para certos tipos de perguntas o *clustering* funciona melhor.

4.6.3 Problemas da revocação e precisão

Greengrass (2000) argumenta que não há um ponto de equilíbrio ótimo entre precisão e revocação, ou melhor, que alguém que recupera todos os documentos de uma coleção, estará certo de que recuperou todos os documentos relevantes da coleção. Nessas condições a revocação será perfeita. Por outro lado, em situações comuns onde apenas uma pequena proporção de documentos na coleção é relevante para responder a uma dada pergunta, a recuperação de tudo dará uma baixa precisão, perto de zero. A afirmativa usual é a de que o usuário deseja a melhor combinação de precisão e revocação, ou seja, recuperar todos os documentos relevantes e nenhum documento não relevante.

Entretanto, a afirmação acima pode ter diversas objeções, é freqüentemente o caso de usuário que deseja somente um subconjunto pequeno de um conjunto maior de documentos relevantes. Os documentos relevantes podem conter uma quantidade considerável de redundância; alguns deles apenas serão suficientes para atender ao que o usuário deseja saber. Caso o usuário esteja procurando evidência para dar suporte à hipótese, ou para reduzir a incerteza sobre uma hipótese, alguns documentos podem fornecer evidências suficientes para esse propósito. Há casos em que o usuário deseja apenas se manter atualizado sobre um determinado tópico, como por exemplo, um advogado que está interessado no último documento legal ou estatuto, e não nos documentos que o precederam.

Apesar das medidas de revocação e precisão serem amplamente utilizadas, Baeza-Yates e Ribeiro-Neto (1999) ressaltam que, diante de reflexões mais cuidadosas, essas medidas apresentam diversos problemas.

A estimativa do máximo de revocação para uma determinada pergunta requer um conhecimento detalhado de todos os documentos que compõem a coleção. Quando se trata de grandes coleções esse conhecimento é praticamente impossível, e geralmente não está

[8] Técnica de mineração de textos para fazer agrupamentos automáticos palavras segundo seu grau de semelhança.

disponível, o que implica em que a medida de revocação não poderá ser mensurada com de forma precisa.

Revocação e precisão são medidas relacionadas e que consideram diferentes aspectos de um conjunto de documentos recuperados. Em diversas situações o uso de uma medida que combine revocação e precisão pode ser apropriado.

As medidas de efetividade, como a revocação e a precisão, requererem um conjunto de perguntas processadas *in batch*. Por outro lado, com os sistemas modernos, a interatividade é o aspecto chave do processo de recuperação.

4.6.4 Outras medidas de desempenho do sistema

Apesar da popularidade das medidas de revocação e precisão há outras que são apropriadas para avaliação de desempenho de sistemas de recuperação. Baeza-Yates e Ribeiro-Neto (1999) apresentam medidas como a expectativa de duração de uma busca, a satisfação do usuários considerando o grau de relevância dos documentos obtidos no resultado da busca, a frustração do usuário considerando o grau de documentos não relevantes recuperados, entre outras.

Lancaster e Fayen (1973), fazem o estudo de medidas como: tempo de resposta, esforço do usuário, forma de resposta e diferença simétrica normalizada, entretanto estas e outras medidas de avaliação de desempenho de sistema de recuperação de informação não serão abordadas por não fazerem parte do escopo da pesquisa.

4.7. Coleções de referência

A pesquisa sobre avaliação de recuperação de informação tem sido freqüentemente criticada sob dois aspectos: a falta de uma estrutura formal que caracterize o conceito de relevância como fundamentação básica, além da ausência de testes modelo e *benchmark*. A primeira dessas críticas é difícil de ser resolvida, devido ao grau de subjetividade psicológica relacionada à decisão quanto à relevância de um determinado documento, que caracteriza a informação em oposição ao dado recuperado. Em relação à segunda crítica foram realizados diversos experimentos desde a década de 60, entretanto, diversas pesquisas foram realizadas com coleções de teste pequenas, o que não refletia os problemas vivenciados por grandes coleções e bases de dados. Essas pesquisas eram realizadas por grupos distintos que conduziam pesquisas com foco em diferentes aspectos da recuperação, o que dificultava a comparação de resultados e o *benchmark*.

4.7.1 Coleção Storage and Information Retrieval System

Maron e Blair (1985) realizaram o primeiro experimento de recuperação de informação em base de dados textuais de larga escala. Sua experiência sucedeu alguns estudos com base de dados de texto completo como as realizadas por Swanson (1960, apud MARON e BLAIR) e Salton (1970, apud MARON e BLAIR). Estes estudos avaliavam o resultado da busca por texto completo por meio do uso de computador, obtendo resultados muito animadores, mas utilizam um conjunto muito pequeno de documentos em texto completo.

Para realização do experimento foi utilizado um sistema da IBM de recuperação de informação em base de dados de texto completo chamado STAIRS que era o acrônimo de STorage And Information Retrieval System, um sistema de recuperação muito rápido e de grande capacidade de armazenamento. Os estudos empíricos realizados com o STAIRS revelaram que o grau de efetividade dos sistemas de recuperação baseados em palavras chaves era surpreendentemente baixo. Maron e Blair (1985) apresentaram suporte teórico para explicar que o desempenho tão pobre do Storage and Information Retrieval System não deveria ser visto como surpresa, além de explicar como seu estudo não era inconsistente, considerando-se os resultados animadores de pesquisas realizadas anteriormente. Eles afirmavam que seu trabalho não visava apenas criticar o sistema Storage and Information Retrieval System, mas tecer uma crítica ao princípio utilizado na construção dele e de outros sistemas de recuperação de texto completo.

Estudos anteriores mostram que a indexação de documentos por termos (índices) que representem o conteúdo do documento apresenta problemas, em função da inexistência de regras claras para que determinar a forma como o especialista deve selecionar os termos que devem representar o documento e que diferentes especialistas selecionam diferentes termos para representar um mesmo documento.

Ao constatar que o problema que ocorria com a indexação de referências bibliográficas também ocorreria com os textos completos armazenados nas bases de dados, Maron e Blair (1985) decidiram conduzir um experimento que pudesse avaliar de forma precisa o grau de efetividade da recuperação de informação em sistema onde a organização dos textos completos era realizada por meio de termos (índices). Nesse experimento a precisão e a revocação foram utilizadas como medida de efetividade. A relevância foi considerada de forma binária, ou seja, um documento recuperado pode ser relevante para a estratégia de busca adotada ou pode ser considerado não relevante. A figura 16 ilustra como foram definidas as medidas de efetividade dos resultados da busca.

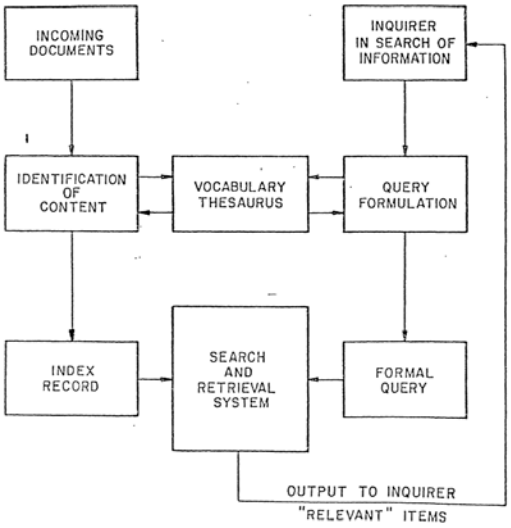

Figura 16 – Esquema de medida de efetividade dos resultados

Fonte: Maron e Blair (1985)

À época, o sistema Storage and Information Retrieval System era considerado o estado da arte em sistemas de recuperação de informação em base de dados textual. Com 40.000 documentos de um escritório de advocacia, este sistema possuía interface que permitia a formulação de consultas a partir de uma simples palavra ou de complexas combinações booleana. As consultas poderiam definir termos que seriam procurados em qualquer parte dos documentos. A recuperação também poderia ser feita por campos de identificação dos documentos, além do corpo do texto do documento, tais como: autor, data, número do documento. O sistema dispunha de função de *ranking* [9]que permitia ordenar o conjunto de documentos recuperados de forma ascendente ou descendente numérica ou alfabética.

O experimento foi conduzido da seguinte forma: dois advogados do escritório que atuavam na defesa de uma grande corporação formulavam suas estratégias de busca por documentos relevantes para defesa e as passavam a seus estagiários que eram pessoas familiarizadas com o sistema Storage and Information Retrieval System para que realizarem as consultas formuladas. Os estagiários poderiam realizar diversas consultas até que julgassem que os documentos recuperados seriam relevantes para atender à necessidade de informação formulada pelos advogados. Os advogados recebiam cópias dos documentos recuperados e organizavam-nos de acordo com seus critérios de relevância (vital, satisfatório, relativamente relevante ou irrelevante). A partir daí os advogados poderiam reformular suas estratégias de busca e solicitar aos estagiários novas consultas no sistema, até que eles estejam

[9] Processo de classificação de itens individuais na escala ordinal de números

satisfeitos com o resultado das suas consultas. A avaliação da revocação e a precisão foi realizada a partir desse conjunto de documentos relevantes (vital, satisfatório ou relativamente relevante) organizados pelos advogados. Para esclarecer o processo de recuperação de informação realizado pelos advogados e estagiários Maron e Blair (1985) formularam o seguinte diagrama:

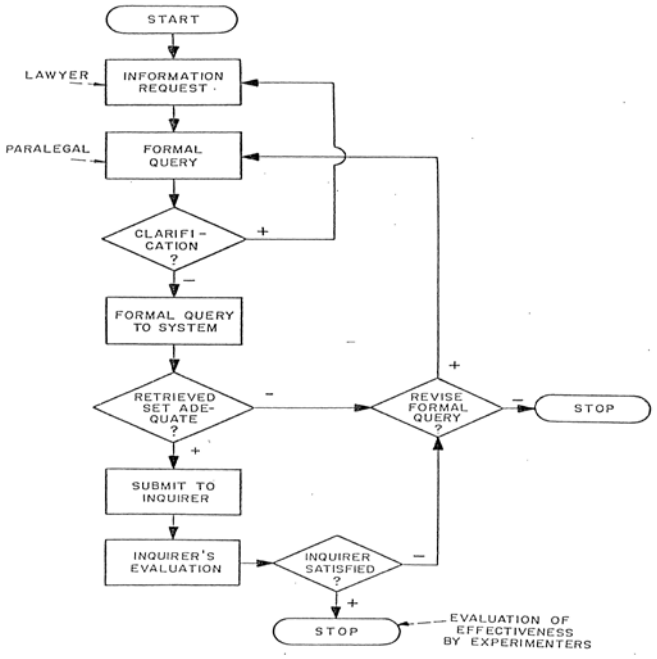

Figura 17 – Representação esquemática do processo de recuperação de informação

Fonte: Maron e Blair (1985)

Como resultado, o estudo revelou que apenas um em cinco documentos relevantes puderam ser recuperados pelo sistema Storage and Information Retrieval System. Essa conclusão movimentou a comunidade científica na promoção de esforços na busca de novas soluções para a recuperação de informação em base de dados de texto completo com vistas a superar a precisão de 20% revelada pela pesquisa de Maron e Blair (1985).

4.7.2 Coleção Text Retrieval Conference

A conferência de recuperação de texto (Text REtrieval Conference) criada nos anos 90 como uma forma de promover o debate científico entre as organizações privadas, as universidades e o governo norte americano sobre os métodos de recuperação de informação é a principal fonte de produção científica na disciplina de avaliação de sistemas de recuperação da informação.

No início dos anos 90, o governo dos Estados Unidos solicitou ao *National Institute of Standards and Technology* (NIST) a construção de uma grande coleção de documentos

para uso em avaliação de tecnologia de recuperação de texto que seria desenvolvida como parte do Projeto Tipster da *Defense Advanced Research Projects Agency* (DARPA). Esse projeto mostrou o interesse do governo norte americano em desenvolver o estado da arte nas tecnologias de processamento de textos por meio da cooperação entre pesquisadores e desenvolvedores no âmbito do governo federal, indústrias e universidades. Por falta de recursos o projeto foi encerrado em 1998. A DARPA e a *Central Intelligence Agency* (CIA) juntaram fundos e, em associação com o NIST e com o *Naval Warfare Systems Center*, retomaram o projeto Tipster.

No esforço de desenvolver o processamento eficiente de documentos o projeto Tipster manteve foco em três principais tecnologias:

- Detecção de documentos: capacidade de encontrar documentos que contenham o tipo de informação que o usuário deseja.
- Extração de informação: capacidade de localizar uma informação específica dentro de um texto.
- Resumo automático: capacidade de condensar o tamanho de um documento expressando sua idéia central.

As pesquisas no projeto Tipster estão focadas em diversas áreas envolvendo as suas três principais tecnologias. Para permitir a organização da produção intelectual dessas pesquisas, o projeto cria um conjunto de conferências. Entre elas a *Text REtrieval Conference*, realizada pela primeira vez em novembro de 1992. A Text Retrieval Conference é uma conferência de avaliação de recuperação de informação na forma textual, cujo propósito é disponibilizar técnicas apropriadas de avaliação de sistema de recuperação de informação utilizados pela indústria e organizações acadêmicas, incluindo o desenvolvimento de novas técnicas de avaliação aplicáveis aos modelos de recuperação da informação atuais (VOORHEES, 2007). A conferência de recuperação textual consiste numa série de *workshops* com o objetivo de apoiar a pesquisa no âmbito da comunidade de pesquisadores da recuperação de informação por meio do fornecimento da infra-estrutura necessária para a avaliação em larga escala de metodologias de recuperação de texto. Especificamente, as conferências têm as seguintes metas: encorajar a pesquisa em recuperação de informações baseada em grandes coleções; aumentar a comunicação entre a indústria, a universidade, e o governo por meio da criação de um fórum aberto para o intercâmbio de idéias de pesquisa; acelerar o processo de transformação de tecnologias a partir dos laboratórios de pesquisa em produtos comerciais, por meio do aperfeiçoamento das metodologias de recuperação aplicado

a problemas reais; e ampliar a disponibilidade de técnicas apropriadas de avaliação para uso pela indústria e pela academia, incluindo o desenvolvimento de técnicas de avaliação mais aplicáveis aos sistemas atuais (VOORHEES e HARMAN, 2005).

Atualmente, a Conferência de Recuperação de Informação Textual é patrocinada pelo NIST e pelo Departamento de Defesa dos Estados Unidos, sendo supervisionada por um comitê formado por representantes do governo, da indústria e da universidade. Para cada Conferência de Recuperação de Informação Textual o NIST fornece um conjunto de documentos e perguntas para teste com os quais os participantes deverão utilizar seus próprios sistemas de recuperação de dados e retornar ao NIST uma lista ordenada dos documentos recuperados. Os participantes da Conferência de Recuperação de Informação Textual utilizam uma grande variedade de modelos de recuperação, incluindo métodos que adotam tesauro automatizado, sistemas de peso, cálculos de similaridade, técnicas de linguagem natural, *feedback* [10]quanto à relevância, dentre outras técnicas. O NIST compatibiliza os resultados individuais, julga os documentos recuperados para correção e avalia os resultados. O ciclo da Conferência de Recuperação de Informação Textual culmina com uma oficina que é um fórum onde os participantes podem compartilhar suas experiências (VOORHEES e HARMAN, 2005).

O interesse na pesquisa de avaliação vem crescendo a cada ano tanto em relação ao número de sistemas participantes quanto ao número de tarefas. Para cada tarefa (*track*) há uma base de documentos e algumas consultas/perguntas que estabelecem o que é a informação procurada e o que constitui um documento relevante (*topics*). Na Conferência de Recuperação de Informação Textual de 2003, por exemplo, participaram noventa e três grupos representando vinte e dois países. As coleções de prova da Conferência de Recuperação de Informação Textual e o software de avaliação estão disponíveis, de forma ampla, à comunidade de pesquisa sobre recuperação da informação, assim diferentes organizações podem avaliar os seus próprios sistemas de recuperação quando desejarem (VOORHEES e HARMAN, 2005). Cada sistema trabalha com a mesma coleção teste que é formada por aproximadamente 426 gigabytes de texto (mais de 25 milhões de documentos) e com um conjunto de perguntas (VOORHEES, 2007). Cada pergunta é a descrição de uma necessidade de informação em linguagem natural, o que na nomenclatura Conferência de Recuperação de Informação Textual é denominado tópico, que pode ser testada com um novo

[10] Processo que permite fornecer dados a uma pessoa ou grupo ajudando-o a melhorar seu desempenho no sentido de atingir seus objetivos.

algoritmo. As coleções teste das Conferência de Recuperação de Informação Textuals são organizadas em três partes:

- O conjunto de documentos
- O conjunto de necessidades de informação, chamado de tópicos
- O julgamento de relevância

4.7.2.1 Conjunto de documentos

É uma coleção de documentos, podendo interpretar-se a palavra "documento" como qualquer unidade de informação, tal como: e-mail, um texto postado em um blog, um caso jurídico. Essa coleção tenta reproduzir o espírito de realismo de uma pesquisa, pois mantém os documentos desse conjunto o mais próximo possível de uma situação real.

4.7.2.2 Conjunto de necessidade de informação

A Conferência de Recuperação de Informação Textual distingue entre a necessidade de informação estipulada e a consulta estruturada submetida ao sistema de recuperação de informação. Na nomenclatura da Conferência de Recuperação de Informação Textual a necessidade de informação declarada é chamada de *topic* e a consulta estruturada é chamada de *query*. Isso equivale a dizer que as *tracks* disponibilizam aos pesquisadores de seu domínio temático os *tópicos*, sendo de responsabilidade do sistema em avaliação a construção das *consultas.* As coleções de teste criam *tópicos* que permitem a utilização de diversos métodos de construção de *consulta*, além de permitir uma declaração clara de qual critério torna um documento relevante. O formato padrão de um *tópico* é: uma identificação do *tópico*, um título, uma descrição e uma narrativa. Esse formato pode variar de acordo com a área temática adotada. Um exemplo de tópico é apresentado na figura 18.

<top>

<num> Number: 168
<title> Topic: Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK).

<narr> Narrative: A relevant document must provide information on the government's responsibility to make AMTRAK an economically viable entity. It could also discuss the privatization of AMTRAK as an alternative to continuing government subsidies. Documents comparing government subsidies given to air and bus transportation with those provided to AMTRAK would also be relevant.

</top>

Figura 18 – Exemplo de representação de um tópico

Fonte: Baeza-Yates e Ribeiro Neto (1999)

4.7.2.3 Julgamento de relevância

Voorhees (2007) define o julgamento de relevância, na visão da Conferência de Recuperação de Informação Textual, como o elemento transformador de um conjunto de documentos e de necessidades de informação em uma coleção de teste. Por meio desse julgamento é possível determinar quais documentos são relevantes e quais são irrelevantes em uma coleção de teste. A Conferência de Recuperação de Informação Textual utiliza freqüentemente o julgamento de relevância binário, ou seja, ou um documento é relevante para um dado tópico ou não é.

A decisão quanto à relevância de um documento que irá compor a coleção de teste é feita em relação ao conjunto de tópicos definidos para tarefa em análise. Com isso, um documento pode ser considerado relevante para um determinado tópico, mas irrelevante para outro. Especialistas no domínio assessoram no julgamento de relevância de um documento em uma coleção. Para definir relevância os assessores assumem que devem escrever um relatório a respeito do assunto objeto do tópico analisado. Se julgarem que alguma informação contida no documento em análise poderá ser utilizada na confecção do relatório, então, esse documento é marcado como relevante para aquele tópico, caso contrário é marcado como irrelevante.

Em função do grande volume de documentos existentes, a coleção de teste utilizada em uma tarefa da Conferência de Recuperação de Informação Textual é criada a partir do julgamento apenas de um subconjunto dos documentos relacionados a um determinado tópico. Por esse motivo, as coleções de teste representam uma abstração útil da coleção de documentos relevantes para os tópicos determinados.

Na Conferência de Recuperação de Informação Textual, o método de *pooling* é a técnica mais utilizada para selecionar documentos que irão compor uma coleção de teste construída a partir do julgamento de relevância feito por especialistas humanos. No método de *pooling* [11]os primeiros documentos recuperados (*top*) como resultado de um conjunto de buscas realizadas sobre uma coleção são combinados visando à formação de um *pool* e apenas os documentos do *pool* [12] serão submetidos ao julgamento de relevância (SPARCK JONES e RIJSBERGEN, 1975).

4.7.2.4 Avaliação do resultado da busca

[11] Processo de coleta e organização que permite compartilhar um conjunto de documentos
[12] Conjunto organizado que pode ser compartilhado

O resultado da busca em uma coleção de teste pode ser avaliado de diferentes formas. Na Conferência de Recuperação de Informação Textual, utiliza-se um pacote para avaliação chamado *trec-eval,* acrônimo de *TREC Avaluation*, escrito por Cris Buckley da Sabir Research (VOORHEES, 2007). Esse pacote utiliza, aproximadamente, 85 índices diferentes de avaliação, entre eles a precisão e a revocação. Para fazer a avaliação dos documentos recuperados é definido um nível de corte do desses documentos. O nível de corte é uma classificação que define o conjunto de documentos recuperados importantes para a avaliação. Por exemplo, o nível de corte dez define que o conjunto de documentos recuperados a serem utilizados na avaliação será dos dez primeiros documentos constantes na lista de recuperação. O valor da precisão ou revocação da recuperação de documentos é calculado pela média obtida a partir da precisão ou revocação de cada *tópico* previamente a partir das necessidades de informação. Historicamente, o peso da precisão ou revocação obtida em cada tópico é igual, não havendo diferenças entre os valores obtidos em cada avaliação. O valor máximo da precisão será 1.0 (100%) quando todos os documentos recuperados são relevantes e de 1.0 para a revocação quando todos os documentos relevantes para o tópico avaliado que existam na coleção de teste sejam recuperados.

4.7.2.5 Legal Track

Uma das tarefas mais recentes da Conferência de Recuperação de Informação Textual, introduzida no ano de 2006, é a *Legal Track*. Essa tarefa tem como propósito avaliar a eficiência de diferentes sistemas de recuperação de informação na busca por acesso a legislações, regulamentos e decisões judiciais (jurisprudências). A *Legal Track* tem como propósito desenvolver tecnologias capazes de atender às necessidades da comunidade jurídica no processo de busca por informações pertinentes a um determinado caso que permita ajudar a ela obter informações legais relacionadas com o contexto de um determinado litígio. Partindo do princípio de que a comunidade jurídica é familiarizada com o uso de expressões booleanas associadas às palavras-chave para construção de suas consultas submetidas a sistemas de recuperação de informação, de acordo com Voorhees (2007), essa tarefa objetiva fazer avaliações de sistemas de recuperação de informação baseados no modelo Booleano, além de outras tecnologias para recuperação de informação jurídica que se mostrem inovadoras.

Essa tarefa é formada por três grupos: *main task, interative task, relevance feedback task*, e o conjunto de documentos usados por todas as tarefas é formado por documentos relacionados a processo envolvendo a empresa americana *US tabacco companies*. Esses

documentos estão armazenados na biblioteca da universidade da Califórnia em São Francisco, EUA. A coleção é composta por aproximadamente sete milhões de documentos.

Nessa tarefa o conjunto de tópicos cujo resultado da busca será avaliado é construído por advogados e refletem o tipo de questionamento que fariam aos sistemas de recuperação jurídica em suas práticas profissionais diárias.

A *main task* realiza buscas em uma coleção de documentos *ad hoc*[13], ou seja, é uma tarefa em que o sistema de recuperação de informação ao realizar a busca o faz em um conjunto de documentos previamente conhecido e selecionados esse fim, ou seja, novas questões representando as necessidades de informação do usuário (tópicos) são continuamente formuladas (dinâmicas) e submetidas ao sistema de recuperação de informação avaliado que possui um conjunto de documentos estático. Essa tarefa assemelha-se ao processo de busca em uma biblioteca utilizando sistema de referências (HARTER e HERT, 1997). O modelo de busca utilizado na *main track* é o booleano.

A *interative task* busca permitir ao usuário o máximo de documentos relevantes quanto possível para um determinado *tópico* a partir da interação efetiva entre o usuário e o sistema de recuperação de informação. A avaliação consiste em até 100 documentos por tópico os quais são pontuados utilizando uma medida onde cada documento relevante obtido equivale a um ponto e cada documento não relevante recuperado vale menos meio ponto.

A *relevance feedback task* visa à obtenção de um valor máximo para a medida de revocação. Os documentos julgados como não relevantes para um determinado tópico são retirados da coleção de teste e um novo *pool* de documentos é formado e uma nova busca é realizada.

A *Legal Task* é hoje o principal projeto de pesquisa no campo da avaliação de sistemas de recuperação de informação para busca de informação jurídica.

4.7.2.6 Críticas aos métodos de avaliação utilizados na Text Retrieval Conference

Na visão de Harter e Hert (1997), os métodos de avaliação utilizados na Conferência de Recuperação de Informação Textual podem ser questionados sob vários aspectos, entretanto, consideram como aspecto principal a forte influência do paradigma de Cranfield no processo de avaliação cujo julgamento é baseado na noção de relevância.

Outros comentários ou questionamentos à metodologia de avaliação utilizada na Conferência de Recuperação de Informação Textual surgem em função, principalmente, das

[13] Expressão latina que quer dizer "com este objetivo". Geralmente significa uma solução designada para um problema ou tarefa específicos, que não pode ser aplicada em outros casos.

medidas utilizadas na avaliação dos sistemas de recuperação da informação. A maior parte delas está relacionada com:

i) Sua utilização em ambientes de laboratório e não em ambientes reais (SPARCK JONES, 1997);
ii) Credibilidade a dar aos julgamentos de relevância, já que este é um conceito subjetivo (BEAULIEU et al., 1996);
iii) Representatividade do conjunto de consultas e de documentos que costumam ser voltados para tópicos de ciência e tecnologia (SPARCK JONES, 1997);
iv) Ausência de usuários finais participando do processo de busca (SPARCK JONES, 1997);
v) O uso insatisfatório da revocação em banco de dados de larga escala devido à incapacidade de se ter conhecimento prévio de todos os documentos da coleção (BELKIN et al., 1987);

Apesar destas questões, os trabalhos realizados na Conferência de Recuperação de Informação Textual constituem-se em importante fonte de produção da pesquisa científica em avaliação de recuperação de informação (HARTER e HERT, 1997). E que tem apontado como melhorar a eficiência dos sistemas de recuperação de informação por meio de avaliação de novas técnicas e indicação de novos rumos para a pesquisa (HAWKING et al., 1999).

4.7.3 Iniciativas de criação de coleções de teste no Brasil

No Brasil o principal trabalho de criação de coleções visando o estudo de avaliação de sistemas de recuperação da informação foi realizado pelo Núcleo Interinstitucional de Lingüística Computacional (NILC) da Universidade de São Paulo, disponível no endereço eletrônico: <http://www.nilc.icmc.usp.br>. As iniciativas em língua portuguesa estão disponíveis na página do centro de recursos para o processamento computacional da língua portuguesa, denominado de linguateca, disponível no seguinte endereço eletrônico: <http://acdc.linguateca.pt/aval_conjunta/Merlin/ColecoesTeste.html>.

Nesse trabalho serão apresentadas duas iniciativas de criação de coleções de teste no Brasil que foram consideradas relevantes para ampliação do entendimento do tema pesquisado.

4.7.3.1 Coleção .gov.br

Esta coleção de teste, proposta pelo NILC, foi criada com páginas do domínio .gov.br retiradas do portal do e-gov.

As páginas do Portal de Serviços e Informações do Governo Brasileiro (e-gov: http://www.e.gov.br/) constituem uma fonte de informação segura para solucionar dúvidas do dia a dia, por exemplo, como proceder para conseguir um alvará, quais são os direitos dos portadores de deficiência, como obter carteira de identidade e outras informações para públicos específicos como pesquisadores e pequenos agricultores: visando solucionar questões como: acompanhar um pedido de bolsa de mestrado; custos de máquinas agrícolas, previsão de safra e preços de terra, entre outras. São muitas as perguntas de domínios específicos para as quais os documentos realmente relevantes são os documentos sobre serviços/informações do domínio.gov.br.

Assim, esta coleção de teste foi criada com páginas do domínio .gov.br e propiciou uma avaliação de como as máquinas de busca estão lidando com o tipo de requisições que seriam mais bem atendidas através de respostas presentes no domínio .gov.br.

4.7.3.2 Corpus Yes, user!

Este corpus foi criado no âmbito do trabalho de doutorado de Rachel Aires, intitulado "Uso de marcadores estilísticos para a busca na Web em português", Universidade de São Paulo, Brasil, concluído em agosto de 2005.

Aires (2005) parte do pressuposto de que nem todos os usuários de sistemas de recuperação de informação para web estão sempre interessados em encontrar todos os documentos a respeito de um determinado assunto. A autora propõe uma análise qualitativa dos *logs*[14] da máquina de busca *todobr* (www.todobr.com.br), que identificar alguns objetivos dos usuários no momento da busca.

A coleção de documentos foi criada a partir da submissão de consultas a três máquinas de busca na web: www.google.com, www.alltheweb.com e www.altavista.com. Os dez primeiros resultados obtidos em cada máquina de busca para cada consulta submetida foram selecionados para compor o *Corpus Yes, user!*

Após esse trabalho, Aires (2005) obteve 1703 textos extraídos da Web brasileira, totalizando pouco mais de um milhão e oitocentas mil palavras. Esses textos foram classificados conforme pudessem satisfazer as necessidades do usuário expressas em uma lista

[14] Um arquivo usado para registrar, descrever ou indicar eventos selecionados identificados durante a execução de um processo ou atividade

de seis necessidades e uma sétima classificação que corresponde aos textos que não satisfaziam a nenhuma das seis necessidades, que foi chamada de categoria "outros".

As categorias de necessidades identificadas por Aires (2005) formam:

a) Definição ou explicação de como ou porque algo acontece.

Exemplos:

- O que é a aurora boreal?
- Como é formado um arco-íris?
- Por que um avião a jato deixa marcas no céu?
- De onde vem o som de um trovão?

b) Explicação sobre como fazer algo ou como algo é feito.

Exemplos:

- Como instalar Linux em meu computador?
- Como é feito um exame ginecológico?
- Como é feito o azeite de oliva?

c) Um apanhado de informações sobre um determinado assunto

Exemplos:

- Um panorama sobre a literatura americana no século XX.
- Um pouco da história do frevo.

d) Notícias

Exemplo:

- Acompanhar as últimas notícias sobre investigações sobre corrupção.

e) Informações sobre uma pessoa ou empresa/instituição/organização

Exemplos:

- Encontrar uma página pessoal.
- Páginas com informações para contato (com currículo, telefone, endereço).
- Página de apresentação de uma determinada empresa.

f) Serviços Online

Exemplos:

- Lojas virtuais
- Serviço dos correios para acompanhamento de envio de encomendas.

g) Outros

Exemplos:

- Blogues
- Piadas

- Publicidade

A coleção proposta por Aires (2005) encontram-se estruturada na forma de diretórios de necessidades, estando disponíveis no seguinte endereço eletrônico: <http://www.linguateca.pt/Repositorio/YesUser/YesUser.tar.gz>.

MÉTODOS

1. Fundamentos

Flick (2004) afirma que o tipo da pesquisa deve estar amparado na pergunta da pesquisa, já que seu objeto de estudo está contido nessa pergunta. De acordo com Mattar (1999), o tipo da pesquisa pode sofrer diferentes classificações, conforme o foco dado pelo pesquisador na análise do problema, podendo ser a pesquisa analisada quanto:

- À natureza das variáveis pesquisadas;
- À natureza do relacionamento entre as variáveis estudadas;
- Ao objetivo do problema de pesquisa;
- À forma utilizada para a coleta de dados primários;
- Ao escopo e amplitude da pesquisa;
- À dimensão da pesquisa no tempo;
- À possibilidade de controle sobre as variáveis em estudo; e
- Ao ambiente de pesquisa.

Focado na natureza e no relacionamento existente entre as variáveis, esta dissertação adota a pesquisa analítica, explanatória ou causal definida por Collis e Hussey (2005). Tal escolha encontra respaldo na intenção do trabalho em descobrir a relação de causa ou efeito entre os valores dos índices de precisão (variável dependente) obtidos a partir da análise dos resultados de buscas por informações jurídicas em dois diferentes modelos de recuperação de informação jurídica (variável independente).

Quanto à natureza das variáveis, pode-se classificar a pesquisa como quantitativa, conforme definido por Flick (2004), pelo caráter fortemente valorado das variáveis pesquisadas, considerando que os métodos de avaliação de sistemas de recuperação de informação definidos na literatura da Ciência da Informação caracterizam-se pela contagem e classificação dos valores de precisão obtidos no resultado da busca (RIJSBERGEN, 1979).

2. Método Aplicado à Pesquisa

O método é a parte da dissertação que deve explicitar a lógica da ação a ser seguida pelo pesquisador. Deve esclarecer o "caminho que faz" para poder chegar às suas conclusões. É necessário que sua lógica de agir na condução da pesquisa fique explicita. Portanto, tem como missão descrever os principais fenômenos a serem estudados, suas ramificações, inter-relações e a forma de se obter os dados necessários (HÜBNER, 1998, p.41).

O método de procedimento adotado para responder à pergunta da pesquisa (quais os efeitos de um sistema de recuperação de informação baseado em casos sobre a medida de precisão no resultado da busca por informação jurídica?) baseou-se nas experiências de avaliação de sistemas de recuperação de informação desenvolvidas pela *Legal Track* realizadas na *Text REtrieval Conference* realizada no ano de 2007, conforme descrito por Voorhees (2007) e Tomlinson et al. (2007). Para essa pesquisa o método de avaliação da Conferência de Recuperação de Informação Textual será aplicado ao caso do processo de recuperação de informação jurídica no âmbito da Justiça Eleitoral do Distrito Federal, tendo como unidade de análise as atividades de recuperação de jurisprudência eleitoral realizados pela Coordenadoria de Registros Partidários e Jurisprudência do Tribunal Regional Eleitoral do Distrito Federal.

O método de avaliação utilizado na Conferência de Recuperação de Informação Textual é constituído de uma base de documentos, que nessa pesquisa será identificada como uma base de casos jurídicos concretos (jurisprudências eleitorais); um grupo de especialistas participantes que nesta pesquisa será formado por servidores lotados na Secretaria Judiciária (SJU) do Tribunal Regional Eleitoral do Distrito Federal; um conjunto de consultas e/ou perguntas que estabelecem o que é a informação procurada, ou seja, quais as reais necessidades de informação dos profissionais da Coordenadoria de Registros Partidários e Jurisprudência Eleitoral (CORPJ) expressas sob a forma de tópicos; um conjunto de critérios que constituem o julgamento de relevância de um documento (caso jurídico concreto) estabelecidos por profissionais do Tribunal Regional Eleitoral do Distrito Federal; um processo de avaliação da medida de precisão obtida pelo sistema de recuperação de informação baseado em casos para cada conjunto de documentos recuperados limitados por um *rank*[15] (nível de corte) daqueles considerados importantes para a avaliação. Nesta pesquisa o *rank* será dos dez primeiros documentos constantes na lista de recuperação. Esta escolha foi baseada na experiência profissional dos participantes da pesquisa em relação à quantidade de documentos normalmente lidos por eles em cada resultado de busca por informações jurídicas. O valor do índice de precisão final será obtido a partir da média aritmética da precisão calculada em cada tópico para cada usuário; por fim, um relatório apresenta as considerações do pesquisador sobre o sistema de recuperação de informação avaliado.

[15] Palavra normalmente relacionada com posição relativa ou forma de ordenação. Neste caso refere-se a posição relativa do ponto de corte estabelecido para o resultado da busca.

2.1 Fases da pesquisa

- A definição da coleção de casos jurídicos (universo);
- A definição do grupo de participantes;
- A definição das necessidades de informação (tópicos);
- A definição da coleção de teste (amostra) baseado nos critérios de julgamento de relevância determinados pelos especialistas participantes;
- A construção do protótipo do modelo de recuperação de informação baseado em casos (instrumento);
- Avaliação do protótipo de recuperação de informação (coleta de dados);
- Apresentação dos resultados e considerações sobre a avaliação realizada (análise de dados);

2.2 Universo

O universo inicial estudado é representado por aproximadamente 4.500 (quatro mil e quinhentos) acórdãos representando as decisões jurídicas do colegiado pleno do Tribunal Regional Eleitoral do Distrito Federal, armazenadas em banco de dados textual. Não são contempladas neste universo as decisões administrativas proferidas pelos membros do Tribunal, já que essas não foram consideradas objeto de estudo por não serem de interesse do processo eleitoral do Distrito Federal.

Em anos eleitorais, há um aumento na quantidade de decisões colegiadas proferidas pelos tribunais eleitorais e, por conseqüência, a procura por jurisprudências eleitorais é mais freqüente. Sendo assim, essa pesquisa optou por fazer um recorte no universo estudado que permita representar o fenômeno da recuperação de informação jurídica no âmbito do Tribunal Regional Eleitoral do Distrito Federal ocorrido em anos de eleições gerais, quando efetivamente há pleito eleitoral. Para representar esse cenário fez-se a escolha pelas jurisprudências publicadas nas últimas eleições gerais realizadas no ano de 2006. Portanto, o universo final estudado será composto por todos os acórdãos produzidos no ano de 2006, em número aproximado de 800 casos jurídicos concretos.

2.3 Especialistas participantes

Para realização das tarefas de julgamento de relevância, definição das necessidades de informação e formulação das consultas a serem submetidas ao sistema de recuperação de

informação jurídica, cinco servidores do quadro permanente de pessoal do Tribunal Regional Eleitoral do Distrito Federal foram convidados a participar dessa pesquisa. O critério de seleção dos especialistas considerou o desempenho de atividades diretamente relacionadas com a organização, armazenamento e recuperação de jurisprudência eleitoral, conforme Quadro 2.

Cargo/Função	Lotação
Secretário Judiciário	Secretaria judiciária
Chefe da Seção de Jurisprudência	Coordenadoria de Registro de Partidos Políticos e Jurisprudência
Técnico Judiciário	Seção de Jurisprudência
Analista Judiciário	Seção de Jurisprudência
Chefe da Seção de Apontamentos	Coordenadoria de Registros e Informações Processuais

Quadro 2 – Relação de servidores participantes

Para melhor entendimento das lotações dos especialistas selecionados, a figura 19 apresenta um organograma da Secretaria Judiciária, obtido a partir do organograma do Tribunal Regional Eleitoral do Distrito Federal publicado em seu Regimento Interno.

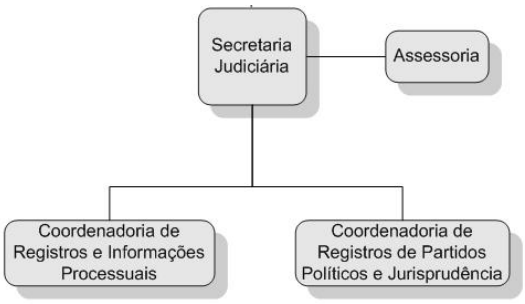

Figura 19 – Organograma da Secretaria Judiciária

Fonte: adaptado de Tribunal Regional Eleitoral do Distrito Federal (2008)

2.4 Necessidades de informação

Na nomenclatura da Conferência de Recuperação de Informação Textual a necessidade de informação declarada é chamada de tópico. Os tópicos permitem a utilização de diversos métodos de construção de *consulta*, além de permitir uma declaração clara de qual critério torna um documento relevante. O formato padrão de um *tópico* é: uma identificação, um título, uma descrição e uma narrativa, conforme Anexo C.

Utilizando o formato padrão de um tópico cada especialista participante da pesquisa formulou cinco necessidades de informação diferentes relacionadas com as suas reais atividades profissionais.

Dessa forma, a pesquisa teve 25 tópicos diferentes, que representam necessidades de informação reais dos especialistas participantes, que foram utilizados na construção da amostra de casos jurídicos concretos e também na avaliação do sistema de recuperação de informação baseado em casos que é objeto de estudo da pesquisa.

2.5 Amostra

A amostra utilizada foi formada por um conjunto de casos jurídicos obtidos a partir do universo pesquisado utilizando-se a técnica de avaliação de relevância chamada método de *pooling*[16] (VOORHEES, 2007), já consagrado nas *Text Retrieval Conferences* e que foi adaptado para o caso em estudo. Para se obter o conjunto de casos jurídicos que representa a amostra desse estudo, foi utilizado o conjunto de 25 tópicos elaborados previamente, em linguagem natural, pelos 5 especialistas participantes da pesquisa, representando um padrão das suas reais necessidades de informação. Esses tópicos foram convertidos em consultas que, em seguida, foram submetidas ao sistema de recuperação da informação atual do Tribunal Regional Eleitoral do Distrito Federal. Do *ranking* gerado no resultado da busca para cada tópico são selecionados os 10 primeiros casos jurídicos possivelmente relevantes (k=10, onde k é o ponto de corte na seleção de casos para formar o conjunto). Os casos jurídicos contidos no conjunto são, então, submetidos à avaliação dos especialistas que decidem sobre a relevância de cada caso jurídico em relação ao tópico utilizado, registrando a identificação dos casos relevantes na ficha padrão do tópico. Dessa forma, um especialista identifica 10 documentos para cada tópico, totalizando uma amostra de 250 casos jurídicos concretos.

O método de *pooling* é baseado em dois pressupostos: primeiro, a grande maioria dos casos jurídicos relevantes para um tópico é coletada na montagem do conjunto de casos; segundo, aqueles documentos que não estão no conjunto de casos jurídicos podem ser considerados como não relevantes. A observação a esses pressupostos garante a exatidão dos resultados obtidos na avaliação dos sistemas de recuperação da informação.

2.6 Instrumento

O principal instrumento utilizado para condução dos procedimentos da pesquisa foi o protótipo de sistema de recuperação de informação jurídica baseado em casos, construído a partir do modelo preconizado por Braga Júnior (2001). Nesse estudo o autor propõe um modelo de organização e recuperação de casos jurídicos onde o cálculo de similaridade é acrescentado ao processo de recuperação de casos jurídicos similares. Braga Júnior (2001)

[16] Método utilizado para formar um conjunto de documentos utilizáveis em um experimento

conclui sua pesquisa propondo a criação de um motor de recuperação de jurisprudência baseado em casos, conforme demonstrado na Figura 20. Resumidamente, o motor de conhecimento é formado por um modelo de organização da informação contida no caso jurídico, pelo processamento da pesquisa textual e pelo uso do cálculo de similaridade utilizado na busca de casos jurídicos similares.

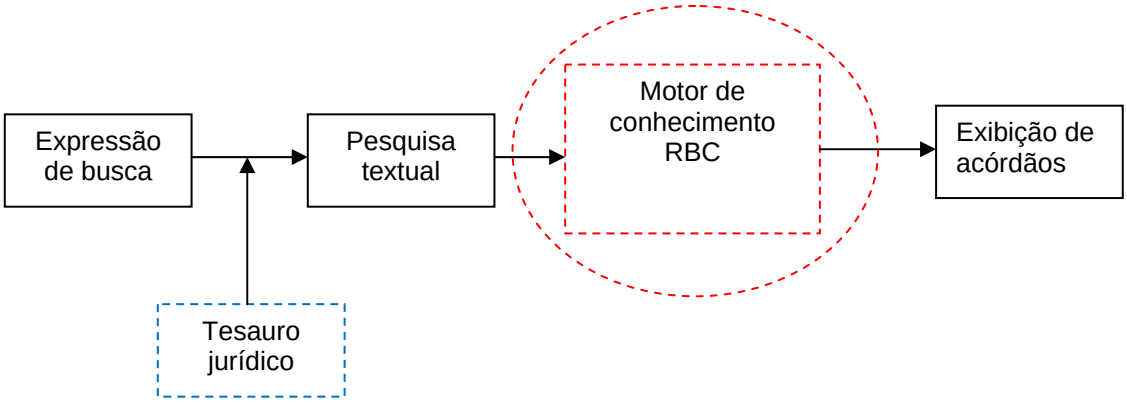

Figura 20 – Motor de Conhecimento baseado em casos para Recuperação de Acórdãos

Fonte: adaptado de Braga Júnior (2001)

2.6.1 Processo de construção do motor recuperação baseado casos

A construção do motor de recuperação proposto por Braga Júnior (2001) foi elaborada em três etapas: a organização da informação contida no acórdão; o processamento da pesquisa textual e o cálculo de similaridade utilizado na identificação de casos jurídicos semelhantes.

1) O modelo de organização da informação contida no acórdão

A dissertação apresentada por Braga Júnior (2001) foi elaborada no seio do Programa de Pós-Graduação em Engenharia da Produção da Universidade de Santa Catarina. Por esse motivo foram necessárias algumas adaptações de conceitos para que se faça uma ponte entre aquela Ciência e a Ciência da Informação, sob a luz da qual foi escrito esse trabalho. Dessa forma, foi possível estabelecer pontos de contato entre as duas disciplinas para que o conhecimento científico desenvolvido naquele estudo possa ser aproveitado nessa pesquisa.

Inicialmente será apresentada a visão da Ciência da Informação quanto à descrição de um documento, em seguida serão apresentados os conceitos relacionados ao modelo descritos com as devidas adaptações.

O processo de organização de informação, segundo Taylor (2004) é formado basicamente pela catalogação, que permite a descrição física do documento e pela

indexação, classificação e elaboração de resumo, que permitem a descrição de conteúdo de um documento. Para fazer a descrição física ou de conteúdo, registros de informação são identificados no documento analisado para representarem seus pontos de acesso. Os pontos de acesso da catalogação representam os aspectos físicos do documento, como: autor, título, editora, etc. Os pontos de acesso da indexação representam os aspectos de conteúdo desse documento, como: termos de indexação, números de classificação e resumos. Os termos da indexação são conceitos-chave obtidos a partir da leitura atenta e detalhada do documento feita por um especialista cujo propósito é representar a idéia principal do tema abordado no documento. Em um processo clássico de recuperação de informação, as palavras de entrada que representam a necessidade de informação do usuário são comparadas aos termos da indexação para seleção dos documentos que atendem ao interesse do usuário.

A estrutura de um acórdão é caracterizada pela ocorrência de dois grupos de elementos: descritivos e temáticos. O primeiro tem o papel de identificar fisicamente do documento, diferenciando-o dos demais. Nele encontramos elementos como: nome do relator, tipo, número e data da decisão. No segundo, estão os elementos que dizem respeito ao conteúdo do acórdão e que de acordo com os Art. 458 e 563 do Código do Processo Civil, são elementos essenciais para um acórdão, são eles: ementa, relatório, motivação (ou fundamentação) e dispositivo.

O modelo de organização proposto por Braga Júnior (2001) para informação contida em acórdão apresenta-se em quatro grupos de pontos de acesso. O primeiro e o segundo são formados por pontos de acesso de catalogação, descrevendo fisicamente os casos jurídicos expressos pelos acórdãos. O primeiro grupo distingue-se do segundo por possuir um único elemento descritivo, o número do processo, cuja característica é a identificação única da ocorrência de um acórdão em uma coleção. Por isso, foi classificado como **ponto de acesso 1**; os demais elementos descritivos são classificados como **ponto de acesso 2**, cuja finalidade para o modelo é servir como filtro na seleção de casos jurídicos, permitindo um refinamento do universo de casos pesquisados. O terceiro grupo é formado pela ementa, que descreve o conteúdo do acórdão por meio da elaboração de resumo. Classificada como **ponto de acesso 3**, é utilizada na pesquisa textual como um filtro que faz a seleção inicial dos casos jurídicos similares. O quarto grupo apresenta as categorias de análise (GUIMARÃES, 1994) formada pelos elementos temáticos: fato, matéria (instituto jurídico), entendimento e argumento, classificados como **ponto de acesso 4**. São empregados

como parâmetros na função de cálculo de similaridade e por esse motivo receberam atribuição de pesos específicos.

Os pesos atribuídos às categorias de análise estão assim distribuídos: 0,35 para o fato, 0,35 para a matéria, 0,20 para entendimento e 0,10 para o argumento, atribuindo um peso total às categorias de análise igual a 1 ou 100%. No caso de não haver termos índices relacionados a uma determinada categoria de análise, o peso dessa categoria será atribuído à categoria fato, sendo essa a única categoria a possuir obrigatoriamente termo índice. De acordo com Braga Júnior (2001), os pesos foram atribuídos em função da experiência do dia-a-dia de atendentes do Tribunal Regional Federal da 1ª Região (TRF1) que auxiliam a pesquisa de jurisprudência no balcão. Eles constataram que os pedidos de consulta se concentram, em sua maioria, no fato e na matéria jurídica e menos no entendimento e argumentos proferidos nos acórdãos.

Ponto de Acesso			**Peso**	**Aplicação**
1	Número do processo		1	Identificação unívoca
2	Nome do relator, número da decisão, data da decisão, tipo do processo, tipo de documento, data da publicação, referência legislativa (sigla, norma, artigo[parágrafo/inciso{letra, item}]		--	Refinamento da pesquisa
3	Ementa		--	Filtro de casos similares
4	Categorias de Análise	Fato	0,35	Cálculo de similaridade
		Matéria	0,35	
		Entendimento	0,20	
		Argumento	0,10	

Quadro 3 – Pesos dos pontos de acesso

Fonte: Adaptado de Braga Júnior (2001)

Os pesos obtidos em pesquisa realizada pelo Tribunal Regional Federal da 1ª Região, mostrados no parágrafo anterior, foram submetidos à avaliação da Seção de Jurisprudência do Tribunal Regional Eleitoral do Distrito Federal, que confirmou que eram pesos correspondentes também à realidade daquele Tribunal.

Com essa forma de descrição dos casos jurídicos, uma busca que contenha a indicação do número do processo dispensa o uso dos demais pontos de acesso, já que retorna uma única ocorrência existente na coleção, por isso possui peso 1. Nas buscas que não contenham a indicação do número do processo, o ponto de acesso 2 serve

como refinamento da pesquisa (filtro de pesquisa) e o ponto de acesso 3 faz a pré-seleção dos casos jurídicos similares identificando no texto da ementa termos de indexação atribuídos às categorias de análise, em seguida, a partir dos casos jurídicos obtidos calcula-se, com base no ponto de acesso 4, o grau de similaridade entre o caso de entrada apresentado pelo usuário e os casos jurídicos previamente filtrados.

A indexação dos casos jurídicos selecionados como parte da coleção de amostra será feita pelos especialistas participantes, em formulário próprio, conforme Anexo E, obedecendo ao modelo de organização de informação jurídica descrito acima. Depois de indexados, os casos jurídicos serão armazenados em base de dados textuais sob a forma de atributo e valor. Os atributos podem ser estruturados em tabelas relacionais ou em paradigmas de orientação a objetos. Nesse trabalho os atributos serão estruturados em tabelas relacionais. Para a tabela de acórdão foi adotada a definição no banco de dados constante no Quadro 4.

Atributo	**Descrição**	**Ponto de acesso**
Número do processo	Número do processo do acórdão. Faz referência à tabela de processos do sistema de acompanhamento de processos. Identificador único.	1
Ano	Ano de autuação do processo	
Nome do relator	Nome do juiz relator do processo. Faz referência à tabela de Juiz do sistema de acompanhamento de processos	2
Decisão	Decisão do recurso. Resultado do julgamento. Possui os valores: qualquer (default), unânime e por maioria.	
Número da decisão	Número do acórdão ou da resolução	2
Data da decisão	Data do julgamento do processo	
Tipo do documento	Tipo de documento jurídico que dá origem à jurisprudência, são eles: Acórdão, Resolução, Decisão sem resolução, Decisão monocrática	
Tipo do processo	Faz referência à tabela de classes do sistema de acompanhamento de documentos e processos (SADP)	
Publicação	Descreve os dados do veículo oficial que publicou a decisão	
Data da publicação	Data da publicação oficial da decisão	
Inteiro teor	Imagem digitalizada o inteiro teor da decisão. Formata padrão GIF.	
Ementa	Resumo dos principais argumentos que fundamentaram a decisão. Texto completo sem divisões.	3

Continuação

Atributo	Descrição	Ponto de acesso
Fato	Categorias de análise obtidas a partir da leitura detalhada do inteiro teor da decisão.	4
Matéria		
Entendimento		
Argumento		
Referência legislativa (sigla, norma, artigo [parágrafo/inciso {letra, item}]	Relaciona os dispositivos legais que dão base à decisão	
Doutrina	Referência bibliográfica a respeito do acórdão, indicado pelo juiz relator.	
Veja também	Referência a outros processos que possuem as mesmas características de acórdão.	
Observação	Referências importantes a respeito do acórdão, como índices de correção, atas, nomes de pessoas etc.	
Indexador	Nome do especialista que fez a indexação.	
Data da indexação	Data em que a decisão foi indexada.	

Quadro 4 – Atributos da tabela de acórdão

Fonte: Adaptada de Braga Júnior (2001)

2) O processamento da pesquisa textual

A interface de recuperação de casos jurídicos oferece duas possibilidades de consultas: por campos de descrição física ou por campos de descrição de conteúdo de um documento jurídico. O processamento da pesquisa textual refere-se apenas a parte da recuperação utilizando campos de descrição de conteúdo.

Nesse processo estão envolvidos a ementa, o fato, a matéria, o entendimento e o argumento. A ementa funciona como um filtro da busca selecionando previamente os casos jurídicos cujos termos da consulta encontrem-se no conteúdo textual da ementa. As categorias de análise são utilizadas para calcular a similaridade entre os casos pré-selecionados pelo texto da ementa e o caso de entrada, construído a partir da expressão de busca do usuário submetida ao tesauro jurídico eleitoral.

Na pesquisa são eliminadas as *stopwords*[17]do texto digitado na expressão de busca e, a seguir, comparam-se as palavras de entrada com os descritores do tesauro jurídico para montagem do caso de entrada. O Quadro 5 mostra um exemplo de caso de

[17] Conhecidas também como Noise Words, são palavras retiradas do texto de uma consulta formulado em linguagem natural por um usuário em uma busca antes do seu processamento.

entrada para a expressão de busca: inelegível por condenação pelo crime de estelionato com decisão transitado em julgado.

Categoria de análise	**Termos índices**	**Pesos**
Fato	Condenação criminal, crime contra patrimônio, trânsito em julgado	0,65
Matéria	Inelegibilidade	0,35
Entendimento		--
Argumento		--

Quadro 5 – Exemplo de caso de entrada

As palavras do texto, juntamente com os descritores associados às respectivas categorias de análise encontradas no tesauro, formam o caso de entrada. Não existindo termos associados para uma determinada categoria de análise, o peso se transfere para a categoria de fato. A partir do caso de entrada montado, faz-se pesquisa textual na ementa. Utilizam-se os termos do caso de entrada como expressão, onde o conector lógico OR é utilizado para ligar todos os termos de uma categoria de análise. Para ligar termos de categorias diferentes usa-se o conector lógico AND. Considerando o caso de entrada de exemplo apresentado anteriormente no Quadro 5, a expressão de busca textual na ementa ficaria dessa forma: [Condenação criminal OR crime contra patrimônio OR trânsito em julgado] AND Inelegibilidade.

Essa consulta obtém todos os casos jurídicos cuja ementa atenda às condições da expressão de busca. A partir dessa lista de casos jurídicos, aplica-se o cálculo de similaridade para identificar a semelhança entre o caso de entrada e os casos pertencentes ao filtro. Os casos cuja similaridade seja igual ou superior a 50% são classificados em ordem decrescente de semelhança e exibidos para consulta. A Figura 21 demonstra a lógica de recuperação prevista para o novo modelo.

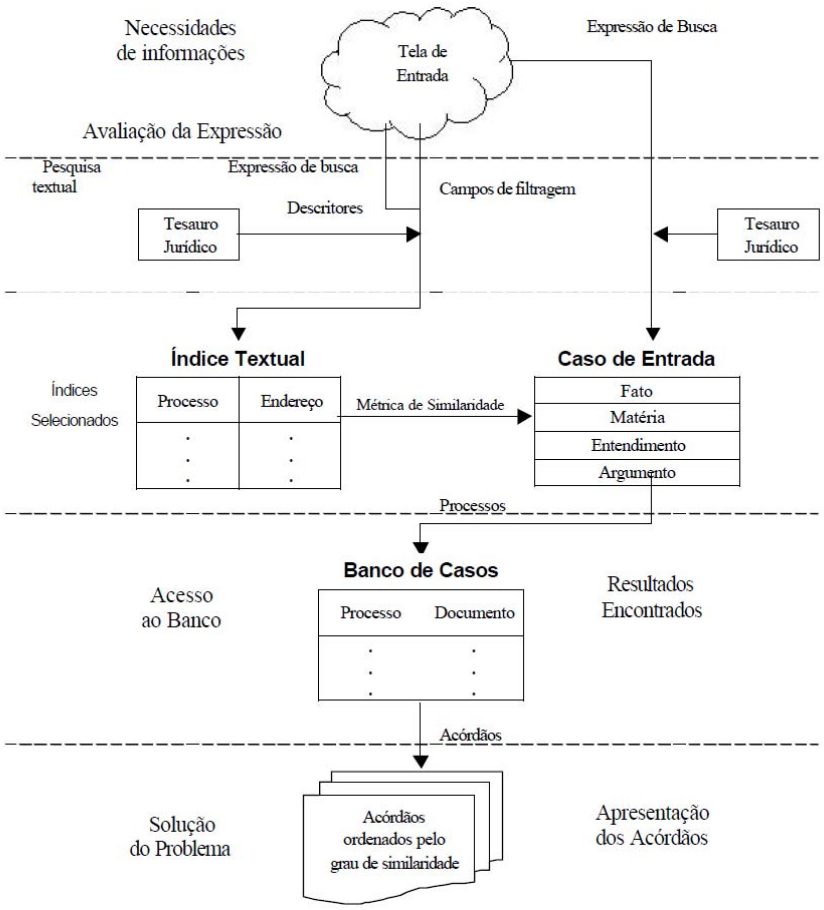

Figura 21 – Processamento da pesquisa textual

Fonte: Braga Júnior (2001)

3) A definição de métrica de similaridade

O objetivo do raciocínio baseado em casos é a reutilização de soluções conhecidas no contexto de um problema novo, de solução ainda desconhecida. Em Função disso, a determinação de casos adequados, que não precisam necessariamente ser idênticos à situação atual, é um dos problemas centrais dessa técnica de inteligência artificial (WANGENHEIM e WANGENHEIM, 2003).

Similaridade não é sinônimo de igualdade no senso lógico, e sim uma noção abstrata da utilidade. Um caso é útil para a solução de um problema se a sua solução é similar à aplicada ao problema atual. Um caso é tão mais útil para solução de um problema, quanto menos for necessário modificar-se aquele caso para adaptá-lo ao problema atual. Identificar o grau de utilidade de um caso passado para solucionar um problema atual é um dos maiores desafios para o uso do raciocínio baseado em casos na recuperação da informação.

Buckhard (1998) afirma que para determinação de similaridade em um sistema de raciocínio baseado em casos, as seguintes premissas têm de ser satisfeitas:

- A similaridade entre a questão atual e o caso implica utilidade;
- A similaridade é baseada em fatos *a priori*;
- Como casos podem ser mais ou menos úteis em relação a uma questão, a similaridade precisa prover uma mediada [0,1] Є ℝ;

Portanto, para utilizar similaridade em um sistema de raciocínio baseado em casos são necessários: a definição das metas de recuperação que se pretende atingir; a identificação de entidades de informação (EI) para determinação da similaridade entre um caso e uma questão; e a definição de um método para decidir se um caso é similar.

a) Metas de recuperação

A utilidade de um caso é um conceito intuitivo sensível ao contexto que depende da aplicação e do objetivo específicos para os quais os casos são recuperados da base de casos. Portanto, o conceito de similaridade deve ser definido em relação às metas de recuperação específicas a serem implementadas por um determinado sistema de raciocínio baseado em casos. Uma meta de recuperação deve apresentar o objeto a ser reutilizado, a finalidade de sua utilização, a tarefa relacionada à reutilização, o ponto de vista específico e o contexto particular (WANGENHEIM e WANGENHEIM, 2003). O Quadro 6 apresenta um gabarito que expressa as metas utilizadas nesse trabalho para a recuperação de casos jurídicos.

Gabarito de metas	**Exemplo**
Recupere <objeto>	Jurisprudência
Para <finalidade>	Suporte à decisão
Relativa ao <processo>	Litígio eleitoral
Do <ponto de vista>	Profissional do direito
No contexto de <ambiente>	Justiça Eleitoral do Distrito Federal

Quadro 6 – Gabarito de metas de recuperação

Fonte: Adaptado de Wangenheim e Wangenheim (2003)

b) Identificação de entidades de informação

Nessa pesquisa, para fins de cálculo de similaridade, as entidades de informação foram identificadas durante o processo de indexação e são representadas pelas categorias de análise formada pelos elementos temáticos: fato, matéria, entendimento e argumento.

c) Método para definição de similaridade

A formalização do conceito de similaridade adotada nesse estudo foi o da similaridade como medida, ou seja, o enfoque desse conceito propõe-se quantificar a extensão da semelhança entre objetos ou fatos.

Portanto, a medida de similaridade é a formalização de um julgamento de semelhança por meio de modelos matemáticos que estabeleçam a medida numérica de distância ou similaridade.

Formalmente, a medida de similaridade entre um conjunto de casos novos N e um conjunto de casos passados P pertencente ao universo U é uma função do tipo: sim(x,y):(NxP→[0,1]), tal que x Є N e y Є P. Se sim(x,y)=1, então a similaridade entre o caso x (novo) e o caso y (passado) é máxima, se sim(x,y)=0 não há similaridade entre x e y, e se $0 < sim(x,y) < 1$, então existe similaridade parcial entre x e y.

Para se chegar ao grau de utilidade de casos jurídicos existentes na coleção de teste (amostra) para resolver o problema apresentado pelo caso de entrada proposto pelo usuário, essa pesquisa utilizou-se de das medidas de similaridades local e global.

- Similaridade local

O primeiro passo foi obter a medida de similaridade local. Essa medida determina a similaridade entre duas entidades de informação, ou seja, a similaridade local está preocupada, por exemplo, com a semelhança entre a categoria de análise fato do caso existente e a categoria de análise fato do caso de entrada. No cálculo de similaridade local o tipo básico utilizado para descrever a entidade da informação é determinante. Como visto anteriormente as categorias de análise são descritas de forma textual, contendo descritores identificados durante o processo de indexação para os casos existentes na base de dados e termos informados pelos usuários e ajustados com o uso do tesauro para o caso de entrada. Por esse motivo, o cálculo de similaridade local foi realizado utilizando a técnica da contagem de palavras. Portanto, a fórmula da similaridade utilizada foi assim formulada:

$$Sim(Q,C) = \left[\frac{QPI}{QPC} \right]$$

Onde:

QPI representa o total de palavras idênticas encontradas na relação entre a categoria de análise do caso de entrada e a categoria de análise do caso recuperado;

QPC representa a quantidade de palavras contidas na categoria de análise do caso recuperado.

Para exemplificar a fórmula da similaridade local, considere o seguinte exemplo:

Expressão de busca: *propaganda eleitoral irregular em praça pública*

O Quadro 7 apresenta o caso de entrada após ajuste.

Categorias de Análise	**Caso de entrada**	**Caso existente**	**Valor da Similaridade local**
Fato	Propaganda eleitoral, irregularidade, patrimônio público	Propaganda eleitoral, bem público	QPI=3 QPC=4 Sim=0,75
Matéria	Crime eleitoral	Crime eleitoral	QPI=2 QPC=2 Sim= 1
Entendimento		Penalidade, multa	QPI=0 QPC=2 Sim=0
Argumento			Sim=0

Quadro 7 – Exemplo de cálculo de similaridade local

Fonte: Adaptado de Wangenheim e Wangenheim (2003)

- Similaridade global

A similaridade global determina a similaridade entre dois casos, permitindo identificar a utilidade de um caso na resolução de determinando problema. Para modelar a similaridade global utilizou-se a técnica do vizinho-mais-próximo (*nearest neighbour*), segunda a qual as ocorrências em uma base de casos podem ser vistas como pontos em um espaço multidimensional. O modelo geométrico utilizado estabelece que a distância espacial entre as respectivas representações dos casos reflete a similaridade entre eles. Definida uma medida de distância, a busca fica reduzida à determinação do vizinho geometricamente mais próximo. O valor de similaridade concreto permite, além de uma estimativa quantitativa do grau de adequação do exemplo de caso y para solução do problema x, formular, em um modelo geométrico para raciocínio baseado em casos, medida de distância (d) entre casos, ou seja, se a sim(x,y)=0,7, então, d(x,y)=0,3. Aquele que prover o menor valor de distância será o vizinho-mais-próximo, ou, ainda, os casos que tiverem maior semelhança terão entre si a menor distância. A figura 22 apresenta

um exemplo de medida de distância entre dois casos passados 1, 2 e um caso atual Q.

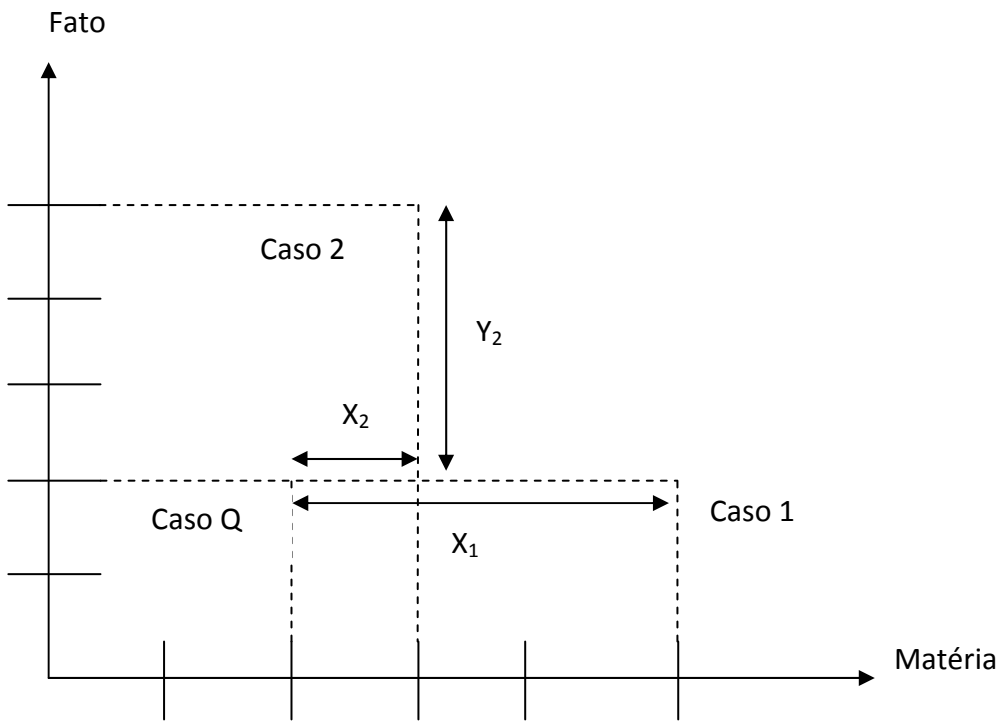

Figura 22 – Exemplo de distância de vizinho-mais-próximo

Fonte: Adaptado de Wangenheim e Wangenheim (2003)

Nesse exemplo, X_1 no eixo Matéria da Figura 22 representa a distância Euclidiana entre o Caso Q e o Caso 1, nesse caso X_1=3 unidades. No eixo Fato da Figura 22, a distância geométrica entre o Caso Q e o Caso 1 é 0. Com isso é possível concluir que não havendo distância entre o Caso Q e o Caso 1 no eixo Fato, nesse eixo os dois casos são similares, ou seja, possuem 100% de similaridade. Nota-se que quanto menor a distância geométrica, maior será a similaridade entre os casos. Se analisarmos a distância entre o Caso Q e o Caso 2 na Figura 22, temos no eixo Matéria a distância geométrica X_2 entre os casos é de 1 unidade, e a distância geométrica Y_2 no eixo Fato entre o Caso Q e o Caso 2 é de 3 unidades. Portanto, a distância geométrica entre Caso atual Q para o Caso passado 1 é d1=X1+0=3+0=3, e a distância geométrica entre Caso atual Q e o Caso passado 2 é d2=X2+Y2=1+3=4. Para o exemplo, o Caso passado 1 é o vizinho-mais-próximo do Caso atual Q, portanto o caso atual Q possui maior grau de similaridade global com o Caso passado 1.

A fórmula para o cálculo de medida de similaridade global utilizado nessa pesquisa foi o seguinte:

$$Sim\ (C_e, C_i) = \sum_{J=1}^{4} (f_j\ (C_{aej}, C_{aij}) x W_j)$$

Onde,

C_e = Caso de entrada
C_i = Caso i da base de casos jurídicos
f_j = Medida de similaridade local entre C_{aej} e C_{aij}
C_{aej} = Categoria de análise j-ésimo do caso de entrada
C_{aij} = Categoria de análise j-ésimo do caso i da base de casos
W_j = Peso da categoria de análise j-ésimo
J = Número de categorias de análise de cada caso
I = Número de casos utilizados no calculo de similaridade

Utilizando o exemplo de cálculo de similaridade local descrito anteriormente no Quadro 7 e os valores de peso das categorias de análise anteriormente apresentados no Quadro 3, temos o seguinte exemplo de cálculo de similaridade global:

$Sim\ (C_e, C_i) = (0,75*0,65)+(1*0,35)= (0,4875)+(0,35)=0,8375$

Onde os valores utilizados na primeira parte do cálculo foram:

0,75 = Similaridade local da categoria fato entre o caso de entrada e o caso existente (Quadro 7).

0,65 = Peso atribuído a categoria fato, sendo 0,35 da própria categoria e 0,20 herdado da categoria entendimento e 0,10 da categoria argumento (Quadro 3), pelo fato dessas categorias de análise possuir similaridade local igual a zero (Quadro 7).

1 = Similaridade local da categoria matéria entre o caso de entrada e o caso existente (Quadro 7).

0,35 = Peso atribuído a categoria matéria (Quadro 3).

Como as categorias entendimento e argumento do caso de entrada em relação ao caso existente tiveram similaridade local igual a zero, não aparecem no cálculo de similaridade global.

Com base no resultado do cálculo de similaridade global entre o caso de entrada e o caso existente, para o exemplo em questão, é possível afirmar que o grau de semelhança entre o caso de entrada e o caso existente é de 83,75%.

4) Ambiente de desenvolvimento

O desenvolvimento do motor de recuperação de informação jurídica que utiliza raciocínio baseado em casos utilizou ambiente operacional composto por: uma

linguagem de programação *PHP*[18], utilizada para construir as interfaces Web do sistema; um servidor de aplicações *TomCat 5.2*, utilizado para disponibilizar a aplicação desenvolvida e para a publicação de suas páginas Web; um banco de dados objeto-relacional Oracle 10g, utilizado na criação e manutenção de tabelas que armazenam os registros de informação identificados como ponto de acesso aos casos jurídicos, além de armazenar, sob a forma de imagem padrão *GIF*[19], os acórdãos digitalizados; uma ferramenta textual chamada *Oracle Text*, integrada ao banco de dados *Oracle 10g*, utilizada na recuperação de informações jurídicas contidas em registros de informação textual, como, por exemplo, o texto da ementa de um acórdão.

2.6.2 Tesauro jurídico eleitoral

Outro instrumento utilizado na pesquisa foi o tesauro da justiça eleitoral. Elaborado pelo Tribunal Superior Eleitoral, esteve disponível para o uso dos especialistas durante o processo de conversão de suas necessidades de informação descritas em linguagem natural em consultas descritas em linguagem controladas.

Para Cavalcanti (1978, p. 89), tesauro é uma "lista estruturada de termos associados, empregada por analistas de informação e indexadores para descrever um documento com a desejada especificidade, em nível de entrada, e para permitir aos pesquisadores a recuperação da informação que procuram".

O tesauro utilizado pelo Tribunal Regional Eleitoral do Distrito Federal, intitulado *Thesaurus*, representa o vocabulário controlado da Justiça Eleitoral por meio de termos organizados em ordem alfabética, sem rigorosa relação de hierarquia entre eles, nas seguintes situações (TSE, 2004):

a) Descritores (termos autorizados): devem ser utilizados na indexação;

b) Não-descritores (termos não autorizados): deve ser substituído pelo termo autorizado correspondente, que lhe é seguido pela anotação "USE";

c) Modificadores: termos de sentido amplo e que não são utilizados isoladamente na indexação, pois não representam sozinhos conceitos para recuperação de informação. São combinados com descritores para esclarecer ou limitar seus significados. Ex: ausência, inexistência, necessidade, etc.

[18] Personal Home Pages – Linguagem de criação páginas Web pessoais
[19] Graphics Interchange Format – Formato padrão de compactação de imagens

Cada descritor do *Thesaurus* é identificado e classificado pela categoria de análise correspondente: fato, matéria, entendimento ou argumento. A classificação é um campo do tesauro associado ao descritor que permite aos especialistas participantes da pesquisa a conversão de suas necessidades de informação em consulta que no modelo de Braga Júnior (2001) é chamada de casos de entrada.

O *Thesaurus* tem como função assegurar a padronização da terminologia utilizada por usuários, documentalistas e indexadores da Justiça Eleitoral (TSE, 2004).

A manutenção de novos descritores é feita pela Comissão de *Thesaurus* e Catálogo do Tribunal Superior Eleitoral levando-se em consideração as sugestões apresentadas pelos indexadores quando, no curso de seu trabalho, encontram a necessidade de representar conceitos existentes em um documento e que não encontrem termos correspondentes no *Thesaurus*.

O *Thesaurus* utiliza um conjunto de anotações que permitiram aos especialistas da pesquisa obter melhor precisão na descrição de suas necessidades de informação. Os termos são:

1. **UP** (Usado Para): precede o termo não autorizado substituído pelo descritor.

Ex: Título de eleitor

UP – Título eleitoral

2. **USE**: indica o escritor que substitui o termo não autorizado.

Ex: Regime democrático

USE – Democracia

As anotações UP e USE têm por finalidade evitar a sinonímia e estabelecer a relação de univocidade entre termo e conceito.

3. **TR** (Termo Relacionado): indica as relações possíveis com outros termos. Objetiva facilitar a compreensão do significado real de um termo, num campo específico, dentro de seu espaço ou ambiente semântico.

Ex: Partido político

TR - Coligação

4. **Nota**: explica os limites da utilização do descritor ou apresenta-lhe informação situacional.

Ex: Apelação em liberdade

Nota – Direito do réu primário e de bons antecedentes de aguardar em liberdade o julgamento da apelação da sentença condenatória.

2.7 Coleta de dados

Os dados da pesquisa foram coletados durante o processo de avaliação do sistema de recuperação de informação baseado em casos. Nesse processo, cada especialista recebeu um conjunto de 5 tópicos distintos que foram transformados por eles em consulta (*query statement*), utilizando-se como técnica de transformação o método manual (VOORHEES, 2007). Para a fase de avaliação do sistema, essa pesquisa adotou a tarefa (*task*) *ad hoc*[20] proposta pela Conferência de Recuperação de Informação Textual, onde as consultas são executadas em uma base de dados cujos casos jurídicos são fixos, determinados pelo tamanho da amostra, ou seja, 250 casos jurídicos concretos.

Para cada consulta submetida ao protótipo do sistema de recuperação de informação jurídica baseado em casos, os 10 primeiros casos apresentados foram avaliados pelo especialista que o julgou como: relevante, não relevante ou não julgado. A categoria não julgado incluiu todos os casos jurídicos presentes no resultado de uma busca, porém o julgamento de relevância não pode ser determinado. Destacam-se como principais motivos para o não julgamento da relevância de um caso jurídico: a falta de certeza do especialista, o tamanho excessivo do acórdão (mais de 100 páginas) e problemas técnicos com a exibição da imagem do documento digitalizado.

Em cada pesquisa, o instrumento utilizado para coletar os dados da avaliação de cada tópico foi preenchido pelo especialista durante o processo de julgamento de relevância. Nesse instrumento são informados a identificação do especialista, o caso de entrada utilizado, a identificação do tópico utilizado, a quantidade de casos jurídicos recuperados, a identificação e julgamento de relevância de cada um dos casos recuperados, conforme modelo apresentado no Anexo D. Os especialistas recorreram ao *Thesaurus* da justiça eleitoral para a elaboração casos de entrada.

Os instrumentos preenchidos pelos especialistas serviram de subsídios para a fase de análise, onde a pesquisa encontrou argumentos para formular as suas conclusões.

2.8 Análise de dados

A análise dos dados coletados corresponde à avaliação de desempenho do sistema de recuperação de casos jurídicos. Para a técnica de avaliação adotado, baseado nas avaliações realizadas na Conferência de Recuperação de Informação Textual, é importante identificar que tipo de tarefa foi utilizada para se obter os dados a serem analisados, já que o tipo de

[20] Expressão latina cuja tradução literal é "para isto" ou "para esta finalidade".

avaliação adotada é definido com base na tarefa realizada. Voorhees (2007) descreve dois tipos distintos: a tarefa em *batch*[21] e a tarefa de seção interativa. Na primeira, a recuperação consiste simplesmente na execução de uma consulta em modo *batch*, ou seja, os especialistas submetem suas consultas ao sistema de recuperação de informação e recebem deles as respostas correspondentes. Na segunda, os usuários especificam suas necessidades de informação por meio de uma série de passos interativos executados no sistema de recuperação de informação onde o resultado da busca pode promover mudanças na consulta que representa a necessidade de informação do usuário. Neste estudo a coleta de dados para avaliação de relevância foi feita utilizando o modo *batch*.

Outra característica relevante para escolha da técnica de avaliação adotada pela pesquisa é quantidade de sistemas de recuperação de informação avaliados. As avaliações podem ser de um sistema individual ou comparadas entre dois ou mais sistema. Para cada caso há um método de avaliação recomendado, de acordo com o que determina a Conferência de Recuperação de Informação Textual. Nessa pesquisa o objeto de estudo é a avaliação de um sistema de recuperação de informação baseado em casos, portanto a técnica de avaliação selecionada foi a indicada para apenas um sistema.

A medida de precisão foi escolhida para ser utilizada na avaliação do sistema. Segundo Baeza-Yates e Ribeiro-Neto (1999), precisão é a fração entre documentos relevantes recuperados e documentos recuperados em um processo de busca, conforme fórmula a seguir:

$$\frac{\text{R – nº de documento relevantes recuperados}}{\text{L – nº de documentos recuperados}}$$

Fórmula da precisão na recuperação de documentos

Fonte: Baeza-Yates e Ribeiro-Neto (1999).

Voorhees (2007) apresenta as principais técnicas destinadas a avaliar sistemas de recuperação de informação individualmente, utilizando a mediada de precisão e cuja tarefa é realiza em modo *batch*. São elas:

- Média principal da precisão (MPP): a idéia dessa técnica é gerar um único valor que represente a média dos valores de precisão obtidos por cada um dos documentos relevantes identificados no resultado da busca. Por exemplo, considere o resultado de uma busca, conforme Quadro 8:

[21] Conjunto ou lote de tarefas

Ordem	Julgamento	Análise	Precisão
1	Relevante	Um documento relevante em um documento recuperado (1/1)	1,00
2	Não relevante		
3	Relevante	Segundo documento relevante em três documentos recuperados (2/3)	0,66
4	Não relevante		
5	Não relevante		
6	Relevante	Terceiro documento relevante em seis documentos recuperados (3/6)	0,50
7	Não relevante		
8	Não relevante		
9	Não relevante		
10	Relevante	Quarto documento relevante em dez documentos recuperados (4/10)	0,40
11	Não relevante		
12	Não relevante		
13	Não relevante		
14	Não relevante		
15	Relevante	Quinto documento relevante em quinze documentos recuperados (5/15)	0,33
Média Principal de Precisão		Soma das precisões de cada documento relevante dividida pelo total de documentos relevantes	0,57

Quadro 8 – Exemplo de cálculo da Média Principal de Precisão

No Quadro 8, as medidas de precisão obtidas para cada documento recuperado, respeitada a sua ordem de apresentação no resultado, foram: 1,00; 0,66; 0,50; 0,40; 0,30; por conseqüência, o cálculo da média principal de precisão para a consulta realizada neste exemplo foi calculada na forma: (1,00+0,66+0,50+0,40+0,33)/5 produzindo um valor médio de precisão de 0,57 ou 57%. Essa medida favorece os sistemas que possuem um bom recurso para ordenação do resultado de uma busca, permitindo exibir os documentos relevantes no topo da lista.

- Precisão-R (Prec-R): nessa técnica o valor da precisão de uma busca é calculado com base em uma definição do número de documentos relevantes (R) existentes no resultado da busca serão considerados para o cálculo de precisão. O valor de R será igual ao número de documentos recuperados que serão analisados quanto a sua relevância. A definição desse valor será influenciada pela freqüência média de documentos lidos por um usuário em cada resultado de busca. Para essa pesquisa, os participantes indicaram como quantidade habitual de documentos lidos em seus resultados de busca os dez primeiros documentos, portanto no cálculo da Precisão-R, R será igual a dez (R=10), ou seja, somente os dez primeiros documentos recuperados serão analisados quanto a sua relevância para o cálculo da precisão. Por exemplo, considere o seguinte resultado de uma busca conforme o Quadro 9:

Ordem	Julgamento	Análise
1	Relevante	Considera para Cálculo
2	Não relevante	
3	Relevante	Considera para Cálculo
4	Não relevante	
5	Não relevante	
6	Relevante	Considera para Cálculo
7	Não relevante	
8	Não relevante	
9	Não relevante	
10	Relevante	Considera para Cálculo
11	Não relevante	
12	Não relevante	
13	Não relevante	
14	Não relevante	
15	Relevante	Não Considera para Cálculo

Quadro 9 – Exemplo de cálculo da Precisão-R

Considerando que para essa consulta somente os 10 primeiros documentos foram avaliados quanto a sua relevância, a Precisão-R dessa busca considera apenas 4 documentos relevantes para o seu cálculo, logo a Precisão-R foi de 0,40. Essa medida é útil para observação do comportamento de sistemas de recuperação de informação em experimentos que avaliam consultas individuais;

- Tabela de dados estatísticos consolidados (TDEC): permite armazenar em uma tabela os valores de medidas de precisão obtidos na execução do conjunto de

> todas as consultas utilizadas no processo de avaliação do sistema, visando prover um resumo estatístico do processo de avaliação. Essa tabela deve incluir: o número de consultas executadas durante a avaliação, número total de documentos recuperados por todas as consultas, número total de documentos relevantes que foram efetivamente recuperados por todas as consultas executadas, número total de documentos relevantes que poderiam ser recuperados por todas as consultas, a medida local de precisão, ou seja, a precisão de um determinado tópico, a medida global de precisão que seria a medida de precisão média de todos os tópicos, entre outros.

Nessa pesquisa utilizou-se como técnica de avaliação a Tabela de Dados Estatísticos Consolidados, conforme modelo apresentado no Anexo F, utilizando a técnica Precisão-R para a avaliação das consultas individuais e para a avaliação do recurso de ordenação do modelo avaliado utilizou-se a Média Principal de Precisão. Portanto, a técnica de análise dos dados ou de avaliação do sistema adotada nesse estudo apropriou-se das três técnicas descritas por Voorhees (2007) e propôs uma técnica híbrida, que integra a melhor característica de cada uma delas.

Por fim, foi elaborada uma conclusão, com base nos dados registrados na Tabela de Dados Estatísticos Consolidados, apresentando as descobertas realizadas a partir da análise desses dados cuja contribuição com o estudo de efetividade dos sistemas de recuperação de informação no domínio jurídico possa ser relevante para fornecer para Ciência da Informação resposta à principal pergunta que suscitou essa pesquisa.

RESULTADOS

Os resultados obtidos pela pesquisa experimental são apresentados a seguir sob a forma de gráficos comparativos das medidas de precisão adotadas neste estudo.

Os dados coletados, conforme representação apresentada nos apêndices de A a K, são resultado de um conjunto de ações de busca, realizadas utilizando-se um protótipo de sistema de recuperação de informação jurídica baseada em casos, refletindo o tipo de questionamento que os participantes da pesquisa fazem em suas práticas profissionais diárias. Os gráficos mostram os resultados obtidos utilizando-se como base os métodos Precisão-R e Média Principal de Precisão, além de exibirem a média de precisão de cada tópico.

1. Especialista 1

A Figura 23 revela que a precisão baseada no método de Média Principal da Precisão foi sempre superior ao método de Precisão-R. Isso pode ser explicado pelo fato do método MPP considerar não apenas a relação entre os documentos recuperados e os documentos recuperados relevantes, mas também a posição relativa do documento relevante no resultado apresentado, ou seja, esse índice valoriza o método de ordenação utilizado. No caso desta pesquisa os resultados são ordenados de forma decrescente de similaridade entre o caso existente e o caso de entrada.

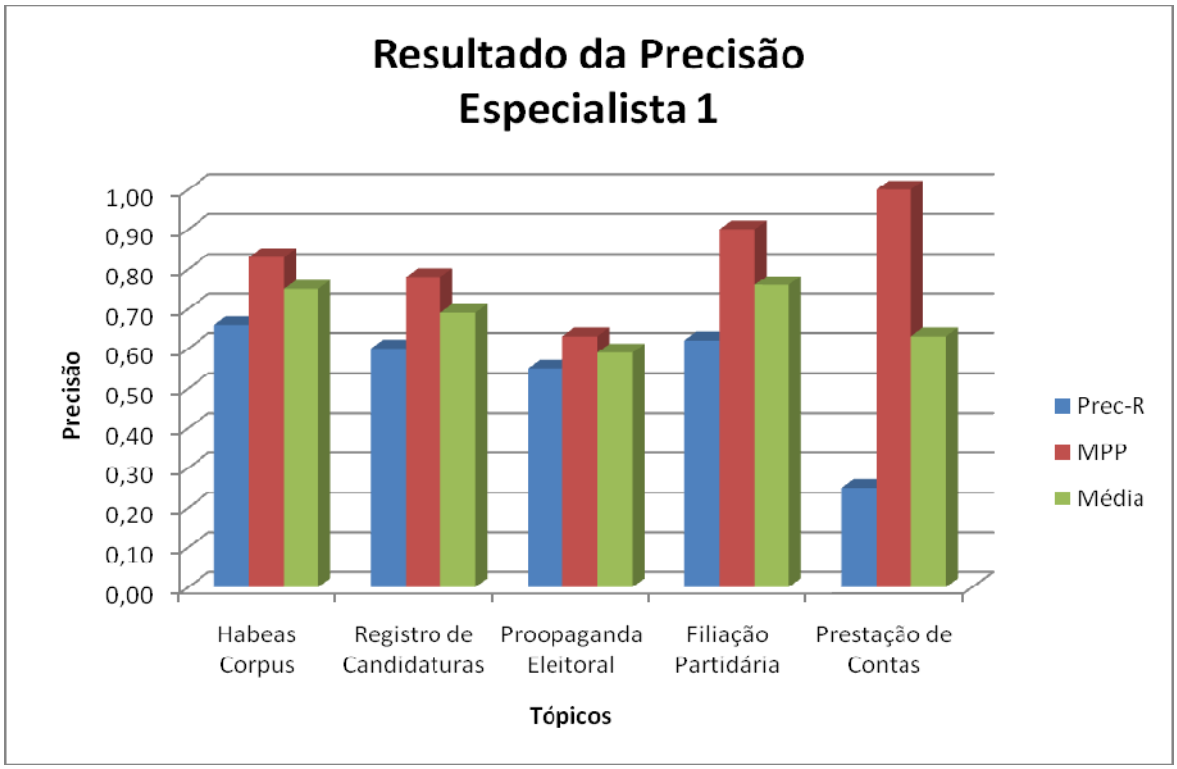

Figura 23 – Gráfico do resultado da precisão para o especialista 1

O tópico *Prestação de Contas*, mostrado na Figura 23, é emblemático para análise da comparação de resultados obtidos nos diferentes métodos de cálculo de precisão. Neste tópico, a Precisão-R obteve um índice de 25% de precisão no resultado da busca, enquanto o resultado, sob a ótica do método Média Principal de Precisão, teve grau de precisão de 100%, demonstrando que este método privilegia os sistemas cujo resultado obtido apresentar no topo da lista os documentos que se infere serem os mais relevantes.

2. Especialista 2

A Figura 24, relativo ao índice de precisão para o especialista 3, confirma a tendência observada no gráfico anterior, quando deixa explícita a superioridade do resultado obtido pelo método Média Principal de Precisão, o que reforça a importância da ordenação do resultado da busca pelo grau de similaridade utilizado no modelo de recuperação de informação jurídica baseada em casos.

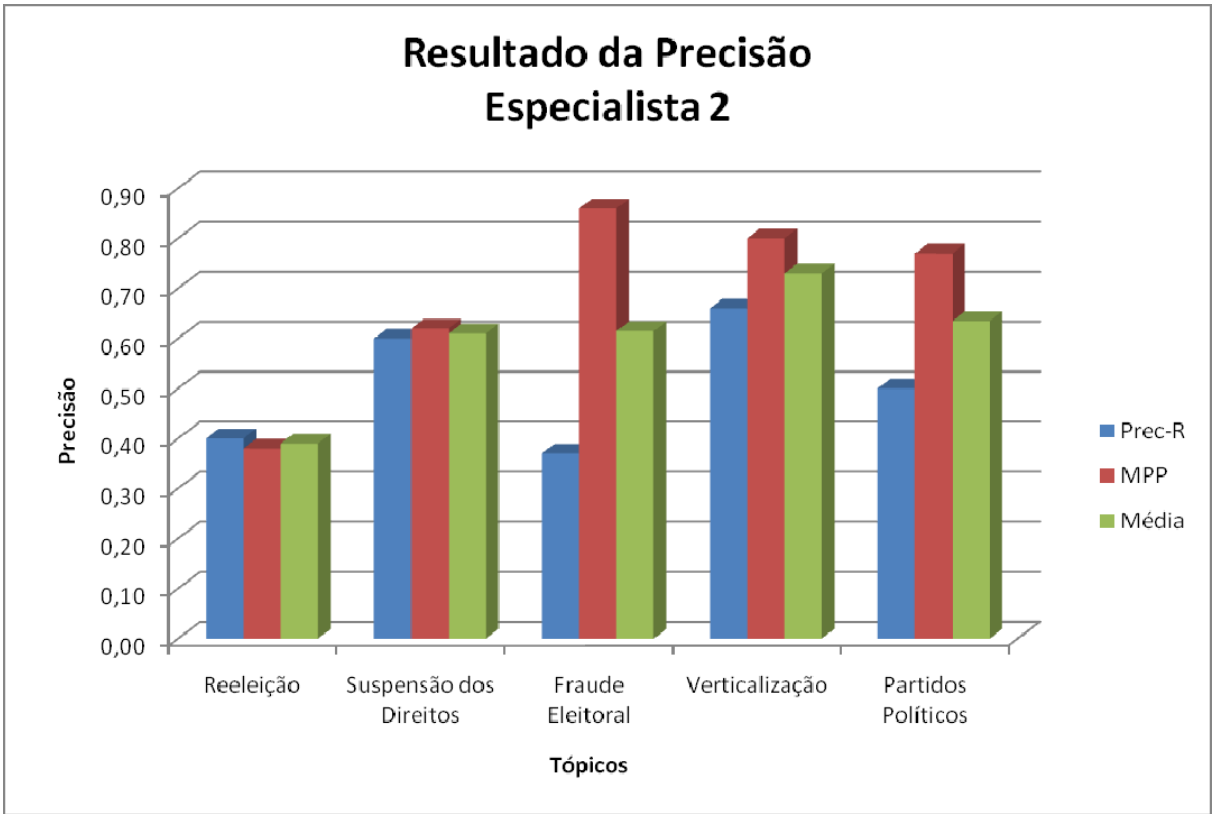

Figura 24 – Gráfico do resultado da precisão para o especialista 2

Na Figura 24, são apresentados os resultados consolidados dos 5 tópicos avaliados pelo especialista 2 quanto à precisão, considerando os métodos Precisão-R e Média Principal de Precisão, além de exibir média de precisão por tópico e média global obtida pelo especialista2.

3. Especialista 3

Observou-se que o índice de precisão obtido pelo especialista 3 na Figura25, apresenta a ocorrência do índice máximo de precisão nos tópicos Cancelamento de Inscrição e Propaganda Eleitoral. Esse fenômeno foi classificado por esta pesquisa como definição insuficiente de documentos relevantes para um tópico. Tal fenômeno descaracteriza a aparente efetividade obtida nessa busca. Por esse motivo e para não comprometer o resultado, os índices de Precisão-R e Média Principal de Precisão desse tópico foram desconsiderados no cálculo da média final da precisão obtida pela pesquisa.

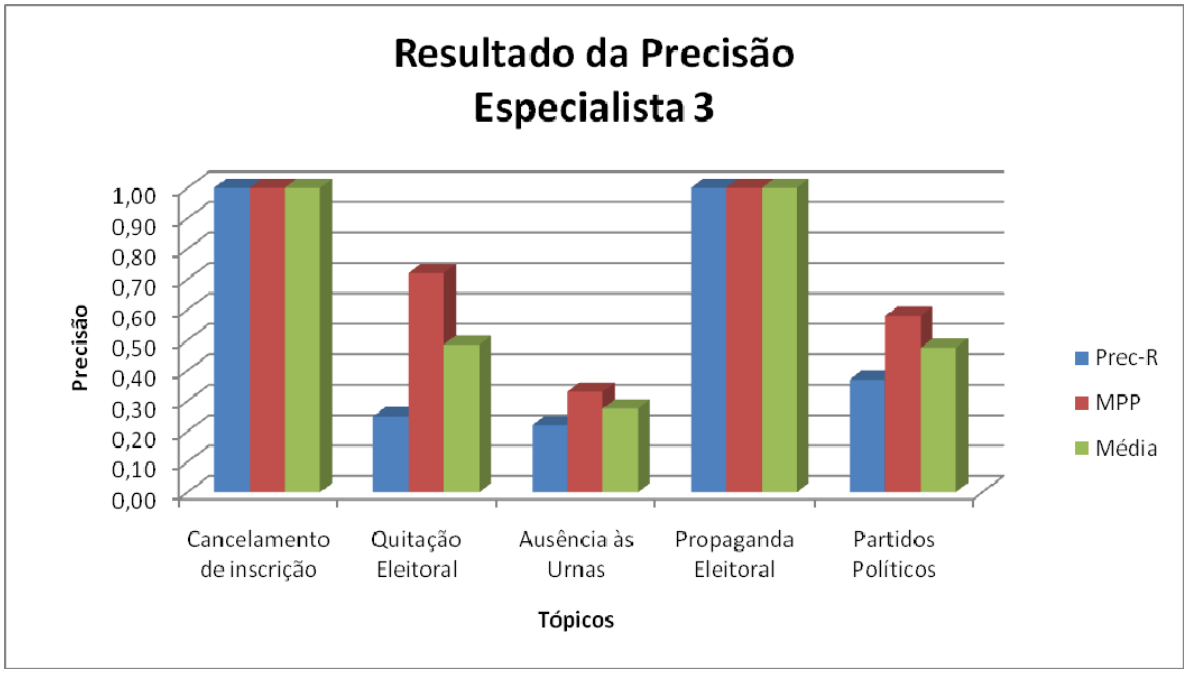

Figura 25 – Gráfico do resultado da precisão para o especialista 3

Na Figura 25, são apresentados os resultados consolidados dos 5 tópicos avaliados pelo especialista 3 quanto à precisão, considerando os métodos Precisão-R e Média Principal de Precisão, além de exibir média de precisão por tópico e média global obtida pelo especialista 3.

4. Especialista 4

Na Figura 26, observou-se valores de precisão coletados a partir dos testes realizados pelo especialista 4, onde é possível perceber que a precisão Precisão-R apresenta-se com valor inferior ou igual ao valor de precisão Média Principal de Precisão em todos os tópicos analisados. Isso caracteriza uma vantagem dos sistemas de recuperação de informação

fortemente focados na ordenação dos seus resultados em relação aos sistemas preocupados apenas em relacionar documentos relevantes em seus resultados.

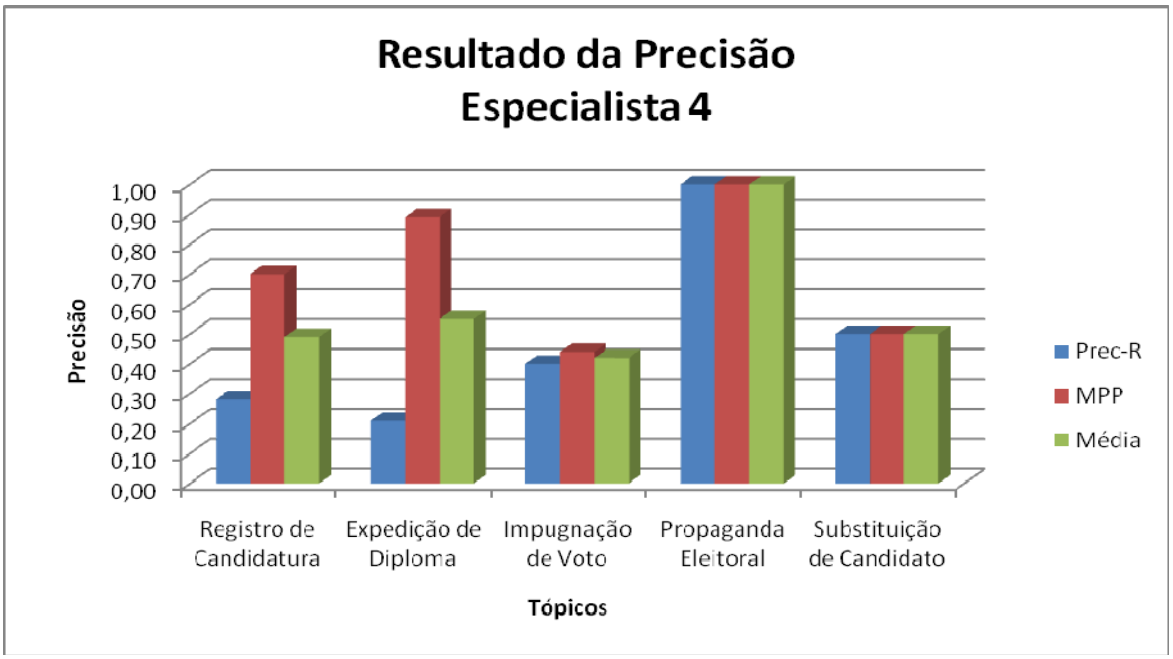

Figura 26 – Gráfico do resultado da precisão para o especialista 4

Na Figura 26, onde os resultados consolidados dos 5 tópicos avaliados pelo especialista 4 são apresentados, ocorre novamente o índice de precisão de 100% no tópico *Propaganda Eleitoral*, tal como ocorrido com o especialista 3. Entretanto, nesse caso, não houve o fenômeno da definição insuficiente de documentos relevantes para um tópico. O que ocorreu foi um incomum caso de precisão máxima, ou seja, todos os cinco documentos recuperados foram considerados pelo especialista 4 como relevantes, contudo essa pesquisa entende que índices de precisão de 100% só são possíveis em testes de laboratório e por esse motivo não considerou o valor de precisão obtido no tópico em questão para apuração do cálculo da média de precisão final.

5. Especialista 5

Ao analisarmos os valores de precisão coletados a partir dos testes realizados pelo especialista 5, representados pela Figura 27, podemos perceber que a combinação de similaridade com relevância permitiu produzir um modelo de recuperação de informação que apresenta uma média geral para o especialista 5 superior a 60% de precisão nos resultados obtidos.

O fenômeno classificado por essa pesquisa como definição insuficiente de documentos relevantes para um tópico, responsável pelo desvio do valor do índice de precisão

para 100% voltou a ocorrer. Agora, envolvendo o tópico *Votação de Conscrito***.** Como definido para o especialista 3, esse resultado será descartado no cálculo da média de precisão final. Também é possível observar a ocorrência da precisão máxima no tópico *Urna Anulada*. Esse tópico recebeu o mesmo tratamento dispensado ao tópico *Propaganda Eleitoral* do especialista 4, tendo, portanto, seus valores de precisão não considerados na média final da precisão.

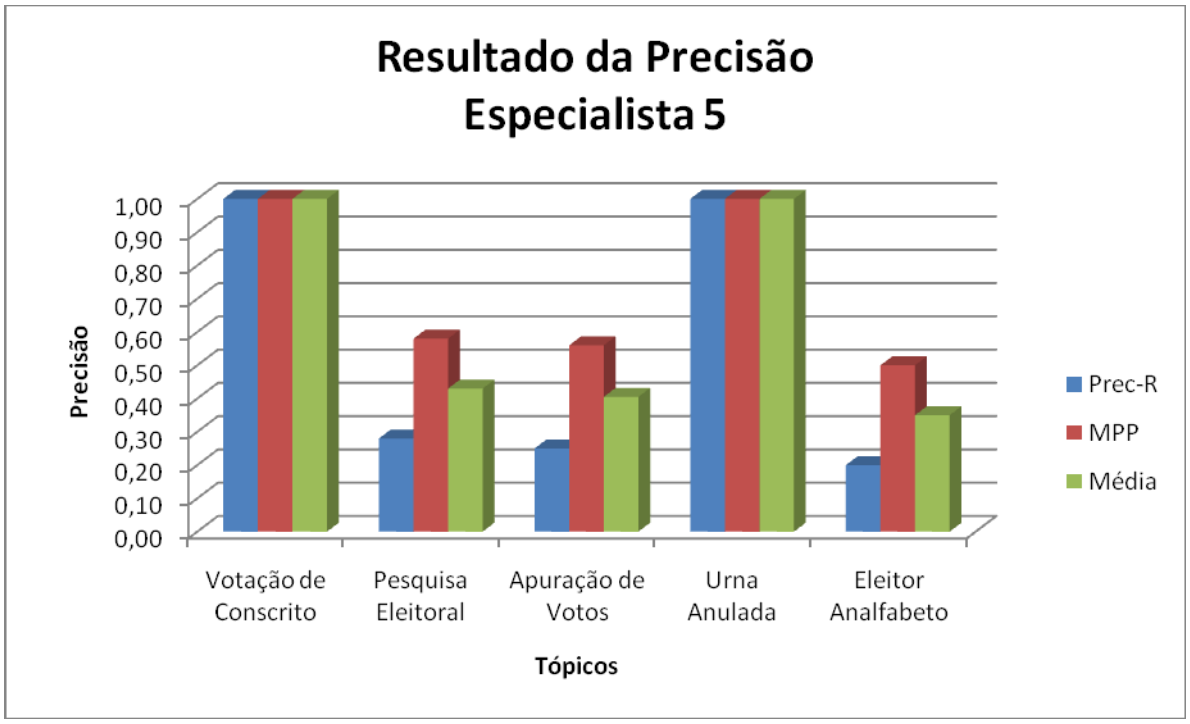

Figura 27 – Gráfico do resultado da precisão para o especialista 5

6. Resultado consolidado

Observando a Figura 28, é possível constatar que o menor valor registrado para a Precisão-R ocorreu no tópico *Eleitor Analfabeto* cuja recuperação foi realizada pelo especialista 5, tendo atingido 20% de precisão no resultado. Ainda observando o método de Precisão-R, os tópicos *Habeas Corpus* utilizado pelo especialista 1 e Verticalização pelo especialista 2, tiveram o maior valor de precisão, 66%. Com isso a média final da Precisão-R foi de 41%.

A Média Principal da Precisão encontrou no tópico *Ausência às Urnas,* utilizado pelo especialista 3, seu menor índice de precisão que foi de 33%, tendo obtido para o especialista 1, no tópico *Prestação de Contas* o seu melhor índice, 100%. Na Figura 28, comparando a Precisão-R com a Média Principal da Precisão para o tópico *Prestação de Contas* percebe-se a importância da organização hierárquica do resultado obedecendo a ordem de similaridade

entre os documentos. Para esse tópico, a Figura 28 apresenta uma Precisão-R de 25% e uma Média Principal da Precisão de 100%, caracterizando a maior distância entre esses dois índices.

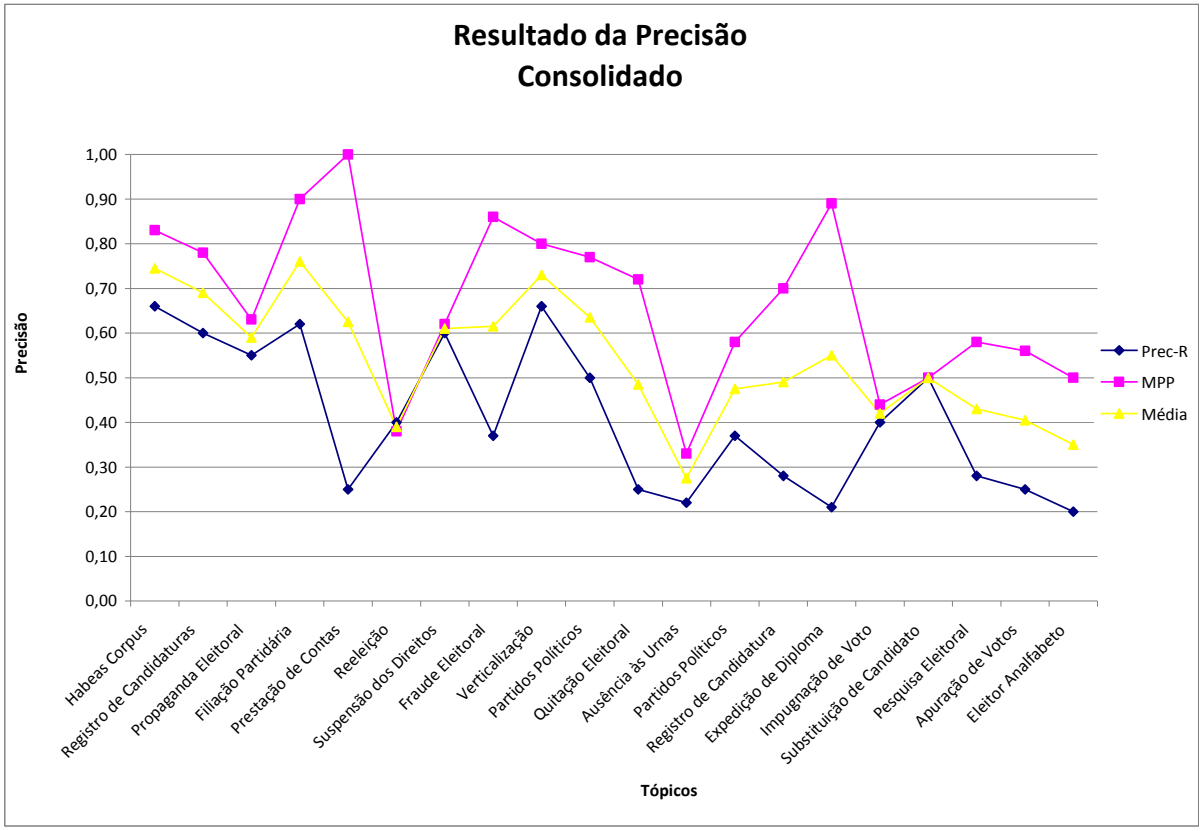

Figura 28 – Gráfico do resultado consolidado da precisão

Observando a Figura 28, é possível constatar que o menor valor registrado para a Precisão-R ocorreu no tópico *Eleitor Analfabeto* cuja recuperação foi realizada pelo especialista 5, tendo atingido 20% de precisão no resultado. Ainda observando o método de Precisão-R, os tópicos *Habeas Corpus* utilizado pelo especialista 1 e *Verticalização* pelo especialista 2, tiveram o maior valor de precisão, 66%. Com isso a média final da Precisão-R foi de 41%.

CONCLUSÃO

O modelo de recuperação baseado em casos utilizado nessa pesquisa apresentou-se como um mecanismo eficiente na recuperação de jurisprudência eleitoral na medida em que o resultado da avaliação da precisão obteve uma média global de 54%. Trata-se de um resultado que supera os 25% apresentados por Maron e Blair (1985) nos estudos da coleção Storage and Information Retrieval System e de 29% apresentados por Voorhees (2007) nas avaliações da *Legal Track* 2007, ambos os resultados referentes ao modelo booleano de recuperação da informação. Essa comparação de resultados remete a pesquisa de volta ao problema que instigou o seu desenvolvimento, ou seja, a pesquisa conseguiu responder positivamente à pergunta que caracterizou o problema esclarecendo que o modelo de recuperação baseado em casos pode produzir melhores índices de precisão se comparados aos do modelo clássico, aprimorando, dessa forma, o sistema de recuperação de jurisprudência eleitoral do Tribunal Regional Eleitoral do Distrito Federal. Os Valores médios da Precisão-R de 41%, combinado com 67% obtido pela Média Principal de Precisão, produziram para o modelo de recuperação baseado em casos uma média global de 54% de precisão. Esse resultado abre espaço para reflexões que permitem formular o argumento de que, no contexto dessa pesquisa, o raciocino baseado em casos aplicado a um modelo de recuperação de informação jurídica pode melhorar o grau de precisão no resultado da busca por informações jurídicas.

Os números obtidos a partir da avaliação do modelo de recuperação baseado em casos mostram que a utilização do método de organização das informações jurídicas baseados nas categorias de análise: fato, matéria, argumento e entendimento, propostas por Guimarães (1994), associado a um motor de busca baseado em casos apresenta indícios de que novos mecanismos construídos nos termos dessa pesquisa possam contribuir para: definição de um novo paradigma para os sistemas de recuperação de informação jurídica; melhorar o índice de precisão no resultado da busca quando comparado ao modelo booleano de recuperação de informação jurídica; construção de um sistema de recuperação de informação jurídica que aproxime o resultado da busca da necessidade de informação dos profissionais do direito. Dessa forma, os resultados obtidos vão de encontro aos objetivos gerais e específicos definidos no início da pesquisa.

Apesar do resultado apresentado por essa pesquisa indicar o raciocínio baseado em casos como uma possibilidade para construção de modelos de sistemas de recuperação de informação jurídica, essa pesquisa não pode ser vista como conclusiva, sendo essa uma característica intrínseca da própria pesquisa científica. Portanto, faz-se necessário que sejam

consideradas as limitações que permearam esse estudo. Entre elas, estão: a realização de experiências de busca por informação jurídica utilizando um ambiente fortemente controlado; a participação apenas de cinco especialistas em direito eleitoral; e a utilização de uma amostra significativa apenas para o estudo das jurisprudências eleitorais do Distrito Federal. Por esse motivo, serão necessárias novas pesquisas que visem ampliar o universo analisado de forma a contemplar a diversidade de jurisprudências produzidas no âmbito das justiças comuns e especializadas brasileiras. Será necessário, ainda, contar com uma maior variação de categorias de participantes na pesquisa, tais como: estagiários e estudantes, advogados, magistrados, assessores, entre outros profissionais do direito, a fim de que se possam consolidar os resultados aqui apresentados.

A metodologia utilizada para avaliação do modelo de recuperação de informação baseado em casos baseou-se nas experiências da *Legal Track* realizada na décima sexta Conferência de Recuperação da Informação (TREC) no ano de 2007, conforme descrito por Voorhees (2007). O uso dos recursos metodológicos da *Legal Track* trouxe maior credibilidade para pesquisa por se tratar de procedimentos e medidas consagradas nas pesquisas de recuperação de informação, contudo duas diferenças fizeram com que esse trabalho apresentasse novas possibilidades. O fato de não ter havido sistemas de recuperação de informação jurídica baseado em casos na relação de sistemas avaliados na Conferência de Recuperação de Informação Textual de 2007 deixou reticente o resultado esperado quando utilizadas as medidas Média Principal de Precisão e Precisão-R. Outro fato relevante no uso da metodologia adotada foi diferença no formato dos documentos utilizados na *Legal Track* e os produzidos pelo Tribunal Regional Eleitoral do Distrito Federal. Na Conferência os documentos avaliados são todos em texto completo armazenados em banco de dados textuais sob a forma não estruturada, já os documentos jurídicos que formam a base de jurisprudência eleitoral do Tribunal do Distrito Federal são semi-estruturados, ou seja, uma parta das informações utilizadas para recuperação de jurisprudências estava armazenada em banco de dados relacionais organizadas sob a forma de atributo-valor, outra parte na forma de texto. Apesar dessas diferenças, a utilização da metodologia baseada no modelo de avaliação da *Legal Track* foi de grande contribuição para essa pesquisa, não exigindo muito esforço para sua adaptação ao modelo de recuperação utilizado.

Trabalhos futuros que podem ser desenvolvidos inspirados nessa pesquisa devem ser focados em dois aspectos: o primeiro diz respeito ao motor de busca. Nesse aspecto, realizar adaptações no modelo avaliado visando automatizar ao processo de organização das jurisprudências por meio do uso da indexação automática pode representar uma importante

contribuição para a evolução dessa pesquisa. Nessa linha, é possível elaborar um estudo que proponha a identificação automática das categorias de análise apresentadas por Guimarães (1994), a partir do texto da jurisprudência. Nesse sentido, o estudo de indexação automática de acórdãos realizado por Câmara Júnior (2007) poderá ser uma boa referência. Um segundo aspecto que pode inspirar novos trabalhos são os métodos de avaliação do modelo de recuperação de informação jurídica. Estudos futuros podem avaliar a necessidade de criação de uma coleção de jurisprudências produzidas pelos principais tribunais brasileiros, nos moldes da Legal Track (VOORHEES, 2007), visando permitir avaliação dos sistemas de recuperação de informação jurídica utilizando um ambiente o mais próximo possível daquele existente nos Tribunais brasileiros. Uma coleção como essa poderá servir como um grande laboratório onde o pesquisador encontraria espaço para: testar novos métodos de avaliação de sistemas de recuperação de informação jurídica; realizar avaliação conjunta de diferentes modelos de sistema de recuperação da informação que estão sendo utilizados pelos órgãos do judiciário brasileiro; identificar as principais diferenças nos métodos utilizados pelo poder judiciário para organizar e armazenar eletronicamente suas decisões; e, finalmente, apoiar o estudo de benchmarking para avaliação de sistemas de recuperação de informação jurídica.

REFERÊNCIAS

AAMODT, A. & PLAZA, E. **Case-Based Reasoning: foundational issues, methodological variations and system approaches**. AI Communications. IOS Press, v. 7, p. 39-59, 1994. Disponível em <http://www.iiia.csic.es/People/enric/AICom_ToC.html>. Acesso em novembro de 2008.

ABEL, Mara. ***Um Estudo sobre Raciocínio Baseado em Casos. Trabalho Individual*** (Pós-Graduação em Ciência da Computação). UFRGS. 1996.

ABEL, Mara; REATEGUI, Eliseo Berni; CASTILHO, José M.V. ***Aquisição, modelagem e processamento de conhecimento utilizando raciocínio baseado em casos***. In: PANEL-95. Proceedings. Canela, 1995.

AIRES, Rachel Virgínia Xavier. **Uso de marcadores estilísticos para a busca na Web em português**. São Paulo, 2005. Tese (Doutorado em Ciências da Computação e Matemática Computacional) - Universidade de São Paulo, USP, 2005.

AIRES, Rachel Virgínia Xavier. **Avaliação de sistemas de recuperação de informação**. 2002. Disponível em: <http://acdc.linguateca.pt/aval_conjunta/Faro2002/HTML/Rachel_Aires/sld001.htm>. Acessado em novembro 2008.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. Addison-Wesley, 1999.

BEAULIEU, M. et al. (1996). Okapi at TREC-5. Maryland. Disponível em: <http://trec.nist.gov/pubs/trec5/t5_proceedings.html>. Acessado em novembro de 2008.

BELKIN, Nicholas J.; CROFT, W. Bruce. **Retrieval Techniques**. In: Williams Martha E., ed. Annual Review of Information Science and Technology: Volume 22. Amsterdam, the Netherlands. 1987.

BELKIN, Nicholas J.; VICKERY, Alina. **Interation in Information Based Systems**. London, UK: Britishn Library. 1985.

BEVILÁQUA, Clóvis. **Teoria Geral do Direito Civil**. 4ª ed. Brasília: Ministério da Justiça, 1972.

BLAIR, D. C. **Language and Representation in Information Retrieval**. Elsevier Science Publishers, 1990.

BRAGA JÚNIOR, Mário de Sena. **Proposta de Modelo de Raciocínio Baseado em Casos para a Recuperação Inteligente de Jurisprudência na Justiça Federal.** Florianópolis, 2001. Dissertação (Mestrado em Engenharia de Produção) - Programa de Pós-Graduação em Engenharia de Produção, UFSC, 2001.

BRÄSCHER, Marisa. **A Ambigüidade na Recuperação da Informação**. DataGramaZero - Revista de Ciência da Informação - v.3 n.1, Fev, 2002.

BRASIL. **Lei de Introdução do Código Civil**. DECRETO-LEI Nº 4.657, DE 4 DE SETEMBRO DE 1942.

BUENO, T. C. D'Agostini, Wangenheim, C. Gresse von, Hoeschl, H. César, Mattos, E. da Silva, Lenz, R. M., Bartsch-Spoerl, B., Burkhard, H.-D.. **Case-Based Reasoning Technology: From Foundations to Applications**. Springer, October 1998.

BUENO, Tânia C. D. et al. **Uso da teoria jurídica para recuperação em amplas bases de textos jurídicos.** Programa de Pós-Graduação em Engenharia de Produção - Universidade Federal de Santa Catarina, Brasil, 1999.

BUENO, Tânia Cristina D' Agostini ; HOESCHL, Hugo Cesar ; BORTOLON, A. ; MATTOS, Eduardo da Silva ; RIBEIRO, M. S. . **Analyzing the use of dynamic weights in legal case based system**. In: Ninth International Conference on ARTIFICIAL INTELLIGENCE and LAW, 2003, Edimburgo. Proceedings of the Conference. New York : ACM, 2003. v. 1. p. 136-141.

BURKHARD, H. D. **Extending some concepts of CBR – Fundations of Case Retrieval Nets**. In M. Lenz et al. (eds.). Case-Based Reasoning Technology from Fundations to Applications, Springer-Verlag, 1998.

CÂMARA JÚNIOR, Auto. **Indexação Automática de Acórdãos por meio de processamento de linguagem natural**. Brasília, 2007. Dissertação (Mestrado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação, UnB, 2007.

CAPURRO, Rafael. **Epistemologia e Ciência da Informação**. V Encontrol Nacional de Pesquisa em Ciência da Informação. Belo Horizonte. 2003. Disponível em: <http://www.enancib.ppgci.ufba.br/artigos/GT1--231.pdf>. Acessado em novembro 2008.

CAVALCANTI, Cordélia R. **Indexação e tesauro: metodologia e técnicas**. Ed. Preliminar. Brasília: Associação de Bibliotecários do Distrito Federal, abr. 1978.

CLEVERDON, Cyril W. **Avaluation of Operation Information Retrieval Systems**. Part 1: Identification of Criteria. Cranfield, England: College of Aeronautics.1964.

CLEVERDON, Cyril W; MILLS, Jack; KEEN, E. Michael. **Factors Determining the Performance of Indexing Systems**. Cranfield, UK: Aslib Cranfield Research Project. College of Aeronautics; 1966.

COLLIS, Hill; HUSSEY, Roger. **Pesquisa em administração**. 2 ed. Porto Alegre: Bookman, 2005.

DERVIN, Brenda; NILAN, Michael. **Information Needs and Users**. In: Williams, Martha E., ed. Annual Review of Information Science and Technology: volume 21. NY. 1986.

ELLIS, David. **Progress and Problems in Information Retrieval**. @nd Edition. Lodon, UK: Library Associations Publishing. 1996.

FERNEDA, Ediberto. **Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**. São Paulo, 2003. Tese (Doutorado em Ciência da Comunicação) – Programa de Pós-Graduação em Ciência da Comunicação da Escola de Comunicação e Artes da Universidade de São Paulo, USP, 2003.

FLICK, U. **Uma introdução à pesquisa qualitativa**. Tradução Sandra Netz. 2. ed. Porto Alegre: Bookman, 2004.

GENTNER, D. **Structure Mapping - A Theoretical Framework for Analogy. Cognitive Science**, Vol. 7, pp. 155-170, 1983.

GREENGRASS, Ed. **Information retrieval: a survey**. 224p. 2000. Disponível em: <http://www.csee.umbc.edu/cadip/readings/IR.report.120600.book.pdf>. Acesso em novembro de. 2008.

GUIMARÃES, José Augusto Chaves. **Análise documentária em jurisprudência: subsídios para uma metodologia de indexação de acórdãos trabalhistas brasileiros.** 1994. Tese (Doutorado em Ciência da Comunicação – área de Biblioteconomia) - Escola de Comunicação e Artes da USP, São Paulo.

HARMAN, Donna K. **Avaluation Issues in Information Retrieval**. Information Processing and Management. 1992.

HARMON, Paul; KING, David. **Sistemas Especialistas**; tradução Antonio Fernandes Carpinteiro. Rio de Janeiro: Campus, 1988.

HARTER, Stephen P., HERT, Carol A. **Avaluation of Information Retrieval System**. Martha E. Williams, ed. Annual review of Information Science and Technology (ARIST), volume 32, 1997.

HAWKING, David; CRASWELL, Nick; HARMAN, Donna. **Results and Challenges in Web Search Evaluation**. 1999. Disponível em: http://www8.org/w8-papers/2c-search-discover/results/results.html. Acessado em novembro de 2008.

HERNOM, Peter; MCCLURE, Charles R. **Avaluation and Library Decision Making**. Norwood, NJ: Ablex Publishing Co.;1990.266p.

HOESCHL, H. C.; BUENO, Tânia Cristina D' Agostini ; DARELLI, L. E. . **Inteligência artificial e direito em Santa Catarina**. 2000. (Programa de rádio ou TV/Mesa redonda).

HÜBNER, Maria Marta. **Guia para elaboração de monografias e projetos de dissertação de mestrado e doutorado**. 1ª Ed. São Paulo. Pioneira Thomson Learning, Mackenzie, 1998.

INGWERSEN, Peter. **Information retrieval interaction**. London: Taylor Graham, 1992, 2002. 246p. Disponível em: <http://www.db.dk/pi/iri>. Acesso em novembro de 2008.

KUHN, Thomas S. **The Structure of Scientific Revolutions**. The uiversity of Chicago Press. 1962/1970

LANCASTER, F.W., FAYEN, E.G. **Information Retrieval On-Line**. Los Angeles, CA: Melville Publishing Co. 1973.

LANCASTER, F. W. **Indexação e resumos: teoria e prática**. Brasília: Briquet de Lemos/Livros, 1993. 347p.

LANCASTER, F.W. **Avaliação de serviços de bibliotecas**. Brasília: Briquet de Lemos, 2004. 356p.

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

LE COADIC, I. F. **A ciência da Informação**. Brasília: Briquet de Lemos, 1996

LESK, M. **The seven ages of information retrieval**. Presented on Conference for the 50th anniversary of "As We May Think", MIT, Cambridge, Massachussets, 1995.

LIMA, Hermes, **Introdução à Ciência do Direito**, Freitas Bastos, 28. ed., 1986.

MARCHIONINI, G. **Interfaces for end-user information seeking**. Journal of the American Society for Information science, New York, v.43, n.2, p.156-63. 1992.

MARINONI, Luiz Guilherme, ARENHART, Sérgio Cruz. **Manual do Processo de Conhecimento: Tutela Jurisdicional Através do Processo de Conhecimento.** Revista dos Tribunais, 2001.

MARON, M. E.; BLAIR, David C. **An Avaluation of Retrieval Effetiveness for a Full-Text Document Retrieval**. Working Paper No. 364. 1985.

MASTERMANN, Margaret. **The nature of a Paradigm**. En: Lakatos, Imre , Musgrave, A. (eds.): Criticisms and the growth of knowledge. Cambridge University Press. 1970.

MATTAR, Fauze Najib. **Pesquisa de marketing: metodologia, planejamento**. 5 ed. – São Paulo: Atlas, 1999. vol. 1 e 2.

MINSKY, Marvin. **A framework for representing knowledge**. In: Winston P. (ed.) The Psychology of computer vision. McGraw-Hill. 1975.

MOOERS, Calvin N. **Zatacoding applied to mechanical organization of knowledge**. American Documentation, v.2, p. 20-32, 1951.

PONTES DE MIRANDA, Francisco Cavalcante. **Sistema da Ciência Positiva do Direito, tomo I**. Atualizado por Vilson Rodrigues Alves. Campinas: Bookseller, 2000, p. 120.

REALE, Miguel. **Fontes e Modelos do Direito: para um Novo Paradigma Hermenêutico**. São Paulo: Saraiva, 1994.

RICH, E., KNIGHT K. **Inteligência Artificial**. São Paulo: Makron Books, 1993.

RIESBECK, Christopher K., SCHANK, Roger C. **Inside case-based reasoning**. Hillsdale, New Jersey: LEA - Lawrence Erlbaum Associates, 1989.

RIJSBERGEN, C.J. Van. **Information Retrieval**. University of Glasgow, 1979. Disponível em: <http://www.dcs.gla.ac.uk/Keith/Preface.html>.Acessado em novembro de 2008.

ROBREDO, Jaime. **Documentação de hoje e de amanhã: uma abordagem revisitada e contemporânea da Ciência da Informação e suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas**. 4.ed. rev. e ampl. Brasília DF: Edição de autor, 2005.

SCHANK, R.C., ABELSON, R. **Scripts, plans, goals, and understanding**. NJ: Erlbaum, Nothvale. 1977.

SCHANK, R. C. Dynamic **Memory, A theory of reminding and learning in computers and people**. New York, Cambridge University Press, 1982.

SARACEVIC, Tefko. **Relevance: A review of and framework for the Thinking on the Notion in information Science**. Jounal of Information Science. 1975.

SHANNON, Claude; WEAVER, Warren. **The mathematical Theory of Comunication**. Urbana, IL: University of Illinois Press. 1949/1972.

SOWA, J. F. **Semantic networks**. 2002.
Disponível em: <http://www.jfsowa.com/pubs/semnet.htm>. Acessado em novembro de 2008.

SPARCK-JONES, K.; Willet, P. **Readings in Information Retrieval**. California: Morgan Kaufmann Publishers, Inc., 1997.

SPARCK JONE, Karen. **Information Retrieval Experiment**. Lodon, Uk: Butterworths. 1981. 352p.

SPARCK JONE, Karen, RIJSBERGEN, C.J. **Report on the need for and provision of an" ideal" information retrieval test collection**. British Library Research and Development Report. 1975.

SVENONIUS, E. **The Intellectual Foundation of Information Organization**. (Ed.); MIT Press. 2000.

SWANSON, Don R. **The Formulation of the Retrieval problem**. New York, NY: McGraw-Hill, 1963.

SWEENEY, M; MAGUIRE, M; SHACKEL, Brain. **Avaluation user-computer Interaction: A Fremawork**. Internation Journal of man-Machine Studies. 1993.

TAGUE-SUTCLIFFE, Jean. Special Topic Issues: **Avaluation of Information Retrieval Systems**. Journal of American Socity for Information Science. 1996.

TAYLOR, Arlene G. **The organization of information**. 2nd ed. Westport, Conn.: Libraries Unlimited, 2004. Chapter 1

TOMLINSON. Stephen. OARD, Douglas W., BARON, Jason R., THOMPSON, Paul. **Overview of the TREC 2007 Legal Track**. 2007. Disponível em: http://trec.nist.gov/pubs/trec16/papers/LEGAL.OVERVIEW.pdf. Acessado em novembro de 2008.

TRIBUNAL REGIONAL ELEITORAL DO DISTRITO FEDERAL. **Regimento Interno**. Brasília, 2008. Disponível em: http://www.tre-df.gov.br/default/institucional/regimento.jsp

TRIBUNAL SUPERIOR ELEITORAL. **Manual do analista de jurisprudência**. Brasília, 2004

VOORHEES, Ellen M. **The Text Retrieval Conference**. In Proceedings. of the 16th Text REtrieval Conference, TREC 2007, at the National Institute of Standards and Technology (NIST) November 6–9, 2007

VOORHEES, Ellen M; HARMAN, Donna K. **The Text Retrieval Conference**. In.: Experiment and evaluation in information retrieval. Cambridge, Mass.: MIT Press, 2005. 472p. Cap. 1, p. 3-20.

WANGENHEIM, C. Gresse von, WANGENHEIM, A. von. ***Raciocínio Baseado Em Casos***. Barueri, SP: Manole, 2003.

WANGENHEIM, C. Gresse von, ALTHOFF, K.-D., BARCIA, R. M. **Intelligent Retrieval of Software Engineering Experienceware**. In Proc. of the 11th International Conference on Software Engineering and Knowledge Engineering (SEKE'99), June 16 to 19, 1999, 128-135.

WATSON, I. D. **Applying case-based reasoning**. São Francisco, Morgan Kaufmann. 1997.

WEBER-LEE, Rosina. **Pesquisa jurisprudencial inteligente**. Florianópolis, 1998. Tese (Doutorado em Engenharia de Produção) - Programa de Pós-Graduação em Engenharia de Produção, UFSC, 1998.

WEBER-LEE, R., Barcia, R. M., Costa, M. C., Filho, I. W., Hoeschl, H. C., Bueno, T. C., Martins, A., and Pacheco, R. C. **A Large Case-Based Reasoner for Legal Cases**. In Proceedings of the Second international Conference on Case-Based Reasoning Research and Development (July 25 - 27, 1997). D. B. Leake and E. Plaza, Eds. Lecture Notes In Computer Science, vol. 1266. Springer-Verlag, London, 190-199.1997.

WILLIS, J. (1910). **Common Law. In The Catholic Encyclopedia**. New York: Robert Appleton Company.Retrieved September 15, 2008. Disponível em: <http://www.newadvent.org/cathen/09068a.htm>. Acessado em novembro de 2008.

WITTGENSTEIN, Ludwig. **Philosophical Investigations**. Oxford, Blackwell. 1953.

APÊNDICE A – Resultado do Especialista 1

Tópico: 1

Objetivo: localizar todas as jurisprudências relacionadas com crime eleitoral por declaração falsa de domicílio prestada por terceiros visando apoiar confecção de habeas corpus.

Consulta utilizada: habeas corpus crime eleitoral falso domicílio

Documentos Recuperados (L): 3

Documentos Relevantes Recuperados (R): 2 (1º e 3º documentos)

Prec-R: 0,66

MPP : (1+0,66) = 0,83

Tópico: 2

Objetivo: localizar todas as jurisprudências relacionadas ao indeferimento de registro de candidatura ao cargo de deputado distrital por situação eleitoral irregular

Consulta utilizada: indeferimento de registro de candidatura

Documentos Recuperados (L): 25

Documentos Relevantes Recuperados (R): 8 (1º, 2º, 4º, 5º, 6º, 8º, 12º, 14º documentos)

Prec-R: 0,60

MPP: (1+1+0,75+0,8+0,83+0,75+0,58+0,57) = 0,78

Tópico: 3

Objetivo: localizar todas as jurisprudências relacionadas ao crime eleitoral por propaganda irregular com prisão em flagrante de cabo eleitoral

Consulta utilizada: prisão por propaganda eleitoral irregular

Documentos Recuperados (L): 9

Documentos Relevantes Recuperados (R): 5 (2º,3º,4º,6º,8º documentos)

Prec-R: 0,55

MPP: (0,5+0,66+0,75+0,66+0,62)=0,63

Tópico: 4

Objetivo: localizar todas as jurisprudências relacionadas ao cancelamento de filiação partidária devido à ocorrência de registro em dois partidos simultaneamente

Consulta utilizada: cancelamento de filiação partidária por duplicidade de registro

Documentos Recuperados (L): 8

Documentos Relevantes Recuperados (R): 5 (1º,2º,3º,5º,7º documentos)

Prec-R: 0,62

MPP: (1+1+1+0,8+0,71)=0,90

Tópico: 5

Objetivo: localizar todas as jurisprudências relacionadas à inegibilidade de candidato por abuso do poder econômico desrespeitando a Lei de responsabilidade fiscal

Consulta utilizada: prestação de contas rejeitada Lei de responsabilidade fiscal

Documentos Recuperados (L): 4

Documentos Relevantes Recuperados (R): 1 (1º documento)

Prec-R: 0,25

MPP: 1,0

APÊNDICE B – Resultado do Especialista 1 Tabelado

Tópico	DocRec	DocRel	Prec-R	MPP	Média do Tópico
Habeas Corpus	3	2	0,66	0,83	**0,75**
Registro de Candidaturas	25	8	0,60	0,78	**0,69**
Propaganda Eleitoral	9	5	0,55	0,63	**0,59**
Filiação Partidária	8	5	0,62	0,90	**0,76**
Prestação de Contas	3	1	0,25	1,00	**0,63**
Média Geral	**9,6**	**4,2**	**0,54**	**0,83**	**0,68**

Legenda:

DocRec – Total de documentos recuperados no tópico

DocRel – Total de documentos recuperados relevantes no tópico

Prec-R – Precisão R com nível de corte 10 (k=10)

RkgDocRel – Ranking dos documentos relevantes recuperados

MPP – Média principal de precisão

APÊNDICE C – Resultado do Especialista 2

Tópico: 6

Objetivo: localizar todas as jurisprudências relacionadas ao registro de candidatura para o cargo de governador após exercício consecutivo de mais de dois mandatos

Consulta utilizada: incompatibilidade de candidatura a governador por reeleição

Documentos Recuperados (L): 10

Documentos Relevantes Recuperados (R): 4 (3º,6º,7º,9º documentos)

Prec-R: 0,40

MPP: (0,33+0,33+0,42+0,44)=0,38

Tópico: 7

Objetivo: localizar todas as jurisprudências relacionadas a eleitor com direitos políticos suspensos por condenação criminal transitada em julgado

Consulta utilizada: direitos suspensos por condenação criminal

Documentos Recuperados (L): 15

Documentos Relevantes Recuperados (R): 9 (2º,3º,5º,6º,8º,9º,11º,12º,14º documentos)

Prec-R: 0,60

MPP: (0,5+0,66+0,6+0,66+0,62+0,66+0,63+0,66+0,64)=0,62

Tópico: 8

Objetivo: localizar todas as jurisprudências relacionadas à ação de impugnação de mandato eletivo por denuncia de fraude no processo de transferência de domicílio eleitoral

Consulta utilizada: impugnação de cargo eletivo por fraude

Documentos Recuperados (L): 8

Documentos Relevantes Recuperados (R): 3 (1º,2º,5º documentos)

Prec-R: 0,37

MPP: (1+1+0,6)=0,86

Tópico: 9

Objetivo: localizar todas as jurisprudências relacionadas à irregularidade na formação de coligação partidária para o cargo de governador em virtude de não respeitar a verticalização

Consulta utilizada: coligação partidária irregular por não verticalização

Documentos Recuperados (L): 6

Documentos Relevantes Recuperados (R): 4 (1º,3º4º,5º documentos)

Prec-R: 0,66

MPP: (1+,66+0,75+0,8)=0,80

Tópico: 10

Objetivo: localizar todas as jurisprudências relacionadas a questões relativas a eleições em órgãos partidários que sejam de competência da Justiça Eleitoral

Consulta utilizada: órgãos partidários matéria interna corporis

Documentos Recuperados (L): 11

Documentos Relevantes Recuperados (R): 5 (1º,2º,5º,6º,8 documentos)

Prec-R: 0,50

MPP: (1+1+0,6+0,66+0,62)=0,77

APÊNDICE D – Resultado do Especialista 2 Tabelado

Tópico	DocRec	DocRel	Prec-R	MPP	Média do Tópico
Reeleição	10	4	0,40	0,38	**0,39**
Suspensão dos Direitos	15	9	0,60	0,62	**0,61**
Fraude Eleitoral	8	3	0,37	0,86	**0,62**
Verticalização	6	4	0,66	0,80	**0,73**
Partidos Políticos	11	5	0,50	0,77	**0,64**
Média Geral	**10**	**5**	**0,51**	**0,69**	**0,60**

Legenda:

DocRec – Total de documentos recuperados no tópico

DocRel – Total de documentos recuperados relevantes no tópico

Prec-R – Precisão R com nível de corte 10 (k=10)

RkgDocRel – Ranking dos documentos relevantes recuperados

MPP – Média principal de precisão

APÊNDICE E – Resultado do Especialista 3

Tópico: 11

Objetivo: localizar todas as jurisprudências relacionadas ao cancelamento da inscrição eleitoral em virtude da falta do eleitor em três pleitos consecutivos

Consulta utilizada: inscrição cancelada por ausência a pleitos

Documentos Recuperados (L): 2

Documentos Relevantes Recuperados (R): 2 (1º, 2º documentos)

Prec-R: 1,0

MPP: (1+1)=1,0

Tópico: 12

Objetivo: localizar todas as jurisprudências relacionadas ao indeferimento de registro de candidatura para o cargo de senador por ausência de quitação eleitoral

Consulta utilizada: ausência de quitação eleitoral indefere candidatura

Documentos Recuperados (L): 16

Documentos Relevantes Recuperados (R): 4 (1º, 3º,4º,8º documento)

Prec-R: 0,25

MPP: (1+0,66+0,75+0,5)=0,72

Tópico: 13

Objetivo: localizar todas as jurisprudências relacionadas ao pagamento de multa por motivo de decisão transitada em virtude da ausência ao pleito eleitoral

Consulta utilizada: pagamento de multa por ausência ao pleito

Documentos Recuperados (L): 9

Documentos Relevantes Recuperados (R): 2 (3º,6º documentos)

Prec-R: 0,22

MPP: (0,33+0,33)=0,33

Tópico: 14

Objetivo: localizar todas as jurisprudências relacionadas à ocorrência de propaganda eleitoral irregular por tratar-se de pintura em mura de instituição pública

Consulta utilizada: propaganda eleitoral irregular ocorrida em área pública

Documentos Recuperados (L): 1

Documentos Relevantes Recuperados (R): 1 (1º documento)

Prec-R: 1,0

MPP: 1,0

Tópico: 15

Objetivo: localizar todas as jurisprudências relacionadas à perda de cargo eletivo por infidelidade partidária configurada pela troca de partido político

Consulta utilizada: troca de partido infidelidade partidária

Documentos Recuperados (L): 8

Documentos Relevantes Recuperados (R): 3 (2º,3º,5º documentos)

Prec-R: 0,37

MPP: (0,5+0,66+0,6)=0,58

APÊNDICE F – Resultado do Especialista 3 Tabulado

Tópico	DocRec	DocRel	Prec-R	MPP	Média do Tópico
Cancelamento de inscrição	2	2	1,00	1,00	**1,00**
Quitação Eleitoral	16	4	0,25	0,72	**0,49**
Ausência às Urnas	9	2	0,22	0,33	**0,28**
Propaganda Eleitoral	1	1	1,00	1,00	**1,00**
Partidos Políticos	8	3	0,37	0,58	**0,48**
Média Geral	**7,2**	**2,4**	**0,57**	**0,73**	**0,65**

Legenda:

DocRec – Total de documentos recuperados no tópico

DocRel – Total de documentos recuperados relevantes no tópico

Prec-R – Precisão R com nível de corte 10 (k=10)

RkgDocRel – Ranking dos documentos relevantes recuperados

MPP – Média principal de precisão

APÊNDICE G – Resultado do Especialista 4

Tópico: 16

Objetivo: localizar todas as jurisprudências relacionadas ao indeferimento de registro de candidatura para o cargo de deputado distrital por possuir o candidato menos de um ano de domicílio eleitoral na circunscrição em que concorrerá ao cargo público

Consulta utilizada: indeferimento de registro de candidatura transferência de domicílio

Documentos Recuperados (L): 7

Documentos Relevantes Recuperados (R): 2 (1º, 5º documentos)

Prec-R: 0,28

MPP: (1+0,4)=0,70

Tópico: 17

Objetivo: localizar todas as jurisprudências relacionadas a pedido de recurso contra expedição de diploma eleitoral para o cargo de governador por motivo de condenação em crime eleitoral de abuso do poder de autoridade

Consulta utilizada: recurso contra expedição de diploma crime eleitoral

Documentos Recuperados (L): 19

Documentos Relevantes Recuperados (R): 4 (1º, 2º, 3º, 7º documentos)

Prec-R: 0,21

MPP: (1+1+1+0,57)=0,89

Tópico: 18

Objetivo: localizar todas as jurisprudências relacionadas ao pedido de impugnação na apuração de votos apresentado por fiscal de partido à junta apuradora

Consulta utilizada: impugnação de voto na junta apuradora

Documentos Recuperados (L): 13

Documentos Relevantes Recuperados (R): 6 (3º,4º,7º, 8º, 11º, 13º documentos)

Prec-R: 0,40

MPP: (0,33+0,5+0,42+0,5+0,45+0,46)=0,44

Tópico: 19

Objetivo: localizar todas as jurisprudências relacionadas ao pedido de direito de resposta por irregularidade em propaganda eleitoral gratuita

Consulta utilizada: propaganda eleitoral gratuita direito de resposta

Documentos Recuperados (L): 5

Documentos Relevantes Recuperados (R): 5 (1º, 2º, 3º, 4º, 5º documentos)

Prec-R: 1,0

MPP: (1+1+1+1+1)=1,0

.

Tópico: 20

Objetivo: localizar todas as jurisprudências relacionadas à substituição de candidato a vice-governador por renuncia ao cargo antes do pleito eleitoral

Consulta utilizada: substituição de candidato a vice-governador por renuncia

Documentos Recuperados (L): 2

Documentos Relevantes Recuperados (R): 1 (2º documento)

Prec-R: 0,50

MPP: (0,5)=0,5

APÊNDICE H – Resultado do Especialista 4 Tabelado

Tópico	DocRec	DocRel	Prec-R	MPP	Média do Tópico
Registro de Candidatura	7	2	0,28	0,70	**0,49**
Expedição de Diploma	19	4	0,21	0,89	**0,55**
Impugnação de Voto	13	6	0,40	0,44	**0,42**
Propaganda Eleitoral	5	5	1,00	1,00	**1,00**
Substituição de Candidato	2	1	0,50	0,50	**0,50**
Média Geral	**9,2**	**3,6**	**0,48**	**0,71**	**0,59**

Legenda:

DocRec – Total de documentos recuperados no tópico

DocRel – Total de documentos recuperados relevantes no tópico

Prec-R – Precisão R com nível de corte 10 (k=10)

RkgDocRel – Ranking dos documentos relevantes recuperados

MPP – Média principal de precisão

APÊNDICE I – Resultado do Especialista 5

Tópico: 21

Objetivo: localizar todas as jurisprudências relacionadas ao impedimento de participação no pleito eleitoral aos eleitores que prestam serviço militar obrigatório mesmo que tenham inscrição eleitoral anterior ao ingresso no serviço militar

Consulta utilizada: impedimento de votação de conscrito

Documentos Recuperados (L): 1

Documentos Relevantes Recuperados (R): 1

Prec-R: 1,0

MPP: 1,0

Tópico: 22

Objetivo: localizar todas as jurisprudências relacionadas ao registro de pesquisa eleitoral de opinião pública relativa às eleições ou aos candidatos feita fora do prazo legal

Consulta utilizada: registro de pesquisa eleitoral

Documentos Recuperados (L): 7

Documentos Relevantes Recuperados (R): 2 (2º, 3º documentos)

Prec-R: 0,28

MPP: (0,5+0,66)=0,58

Tópico: 23

Objetivo: localizar todas as jurisprudências relacionadas à disposição sobre apuração e totalização dos votos, proclamação e diplomação dos eleitos nas eleições de 2006

Consulta utilizada: dispõe sobre apuração de votos

Documentos Recuperados (L): 12

Documentos Relevantes Recuperados (R): 5 (1º, 3º, 8º, 11º, 12º documentos)

Prec-R: 0,25

MPP: (1+0,66+0,37+0,36+0,41)=0,56

Tópico: 24

Objetivo: localizar todas as jurisprudências relacionadas com a ocorrência de urnas anuladas e apuradas em separado durante o pleito eleitoral por falsidade ideológica de eleitor

Consulta utilizada: urna anulada e apurada em separado

Documentos Recuperados (L): 3

Documentos Relevantes Recuperados (R): 3 (1º, 2º, 3º documentos)

Prec-R: 1,0

MPP: (1+1+1)=1,0

Tópico: 25

Objetivo: localizar todas as jurisprudências relacionadas à isenção de penalidades para eleitores analfabetos por ausência às urnas em pleito eleitoral

Consulta utilizada: ausência às urnas de eleitor analfabeto

Documentos Recuperados (L): 5

Documentos Relevantes Recuperados (R): 1 (2º documento)

Prec-R: 0,20

MPP: 0,50

APÊNDICE J – Resultado do Especialista 5 Tabulado

Tópico	DocRec	DocRel	Prec-R	MPP	Média do Tópico
Votação de Conscrito	1	1	1,00	1,00	**1,00**
Pesquisa Eleitoral	7	2	0,28	0,58	**0,43**
Apuração de Votos	12	5	0,25	0,56	**0,41**
Urna Anulada	3	3	1,00	1,00	**1,00**
Eleitor Analfabeto	5	1	0,20	0,50	**0,35**
Média Geral	**5,6**	**2,4**	**0,55**	**0,73**	**0,64**

Legenda:

DocRec – Total de documentos recuperados no tópico

DocRel – Total de documentos recuperados relevantes no tópico

Prec-R – Precisão R com nível de corte 10 (k=10)

RkgDocRel – Ranking dos documentos relevantes recuperados

MPP – Média principal de precisão

APÊNDICE K – Resultado Consolidado Tabulado

Na Tabela de Dados Estatísticos Consolidados (TDEC) são apresentados os resultados obtidos com a avaliação do sistema de recuperação de informação baseado em casos.

Tópico	DocRec	DocRel	Prec-R	MPP	Média do Tópico
Habeas Corpus	3	2	0,66	0,83	**0,75**
Registro de Candidaturas	25	8	0,60	0,78	**0,69**
Propaganda Eleitoral	9	5	0,55	0,63	**0,59**
Filiação Partidária	8	5	0,62	0,90	**0,76**
Prestação de Contas	3	1	0,25	1,00	**0,63**
Reeleição	10	4	0,40	0,38	**0,39**
Suspensão dos Direitos	15	9	0,60	0,62	**0,61**
Fraude Eleitoral	8	3	0,37	0,86	**0,62**
Verticalização	6	4	0,66	0,80	**0,73**
Partidos Políticos	11	5	0,50	0,77	**0,64**
Quitação Eleitoral	16	4	0,25	0,72	**0,49**
Ausência às Urnas	9	2	0,22	0,33	**0,28**
Partidos Políticos	8	3	0,37	0,58	**0,48**
Registro de Candidatura	7	2	0,28	0,70	**0,49**
Expedição de Diploma	19	4	0,21	0,89	**0,55**
Impugnação de Voto	13	6	0,40	0,44	**0,42**
Substituição de Candidato	2	1	0,50	0,50	**0,50**
Pesquisa Eleitoral	7	2	0,28	0,58	**0,43**
Apuração de Votos	12	5	0,25	0,56	**0,41**
Eleitor Analfabeto	5	1	0,20	0,50	**0,35**
Média Global	**9,8**	**3,9**	**0,41**	**0,67**	**0,54**

Legenda:

DocRec – Total de documentos recuperados no tópico

DocRel – Total de documentos recuperados relevantes no tópico

Prec-R – Precisão R com nível de corte 10 (k=10)

RkgDocRel – Ranking dos documentos relevantes recuperados

MPP – Média principal de precisão

ANEXO A – Formulário de indexação: Modelo TSE

1) Campo de identificação do documento

Nº do acórdão/resolução:_____ data: __/__/__ tipo/nº do proc:____UF:____

2) Campo indexação para base de dados

_____

3) Campo catálogo

_____

4) Campo indexação para revista de jurisprudência do TSE

_____

5) Campos precedentes

Tipo de decisão(Ac./Res./Dec.):___nº___data:__/__/__classe/tipo/nº do proc ___

6) Campo vide

Alteração/revogação/vide:__tipo de decisão(Ac./Res./Dec.):____nº__data:__/__/__

UF:__Rel./Red. Desig.:_____

Observações:_____

7) Campo referência legislativa

_____

8) Campo doutrina

_____

9) Indexado por

_____

ANEXO B – Exemplo de Acórdão

ACÓRDÃO N° 2686
Habeas Corpus nº 172 – Classe 7

Relator: Desembargador Estevam Maia
Paciente: Altamiro Rajão

EMENTA
HABEAS CORPUS - TRANCAMENTO DE AÇÃO PENAL - SENTENÇA PROLATADA - PERDA SUPERVENIENTE DO OBJETO - PROCESSO EXTINTO SEM JULGAMENTO DO MÉRITO.
Ao impetrar o *habeas corpus,* o paciente pretendia o trancamento da ação penal. Ocorreu, todavia, a perda superveniente do objeto, eis que a sentença já foi prolatada, fato que deu origem ao Recurso Criminal n°460, em tramitação neste Tribunal.
Ante o exposto, extingue-se o processo sem o enfrentamento do mérito.
Acordam os juizes do **TRIBUNAL REGIONAL ELEITORAL, ESTEVAM MAIA - relator, JOSE LUIZ DA CUNHA FILHO, CARLOS FERNANDO MATHIAS DE SOUZA, ROBERVAL CASEMIRO BELINATI, SILVÂNIO BARBOSA DOS SANTOS e FREDERICO BERNARDES VASCONCELOS** - vogais, em julgar prejudicado o pedido. Decisão **UNANIME**, de acordo com a ata do julgamento e as notas taquigráficas.
Publicada no Diário da Justiça, Seção 3, de 11de 12 de 2007, fls. 93
RELATÓRIO
Cuida-se de *habeas corpus,* com pedido de concessão de liminar, impetrado por Manoel Ninaut filho, em face de ato do ilustre Juiz da 1ª Zona Eleitoral do Distrito Federal, consubstanciado no recebimento de denúncia em face da prática, *in thesi,* do tipo penal previsto no art. 347 do Código Eleitoral.
Defende a inexistência de justa causa para o prosseguimento da ação penal, uma vez que a conduta descrita na peça acusatória, segundo alegado, não configura crime eleitoral.
Em apertada síntese, assinala que o paciente encontrava-se em local próximo ao Centro de Formação de Praças do CBMDF para acompanhar seu genitor, o sr. José Rajão Filho, então candidato ao cargo de deputado distrital, o qual fora convidado para participar da cerimônia de encerramento do Curso de Cabos do Corpo de Bombeiros Militar do Distrito Federal.
Assevera que o paciente foi abordado por pessoas sem identificação, que o indagaram sobre, o que ele portava dentro de sua vestimenta. Acresce que; 'ante a recusa do paciente de apresentar seus pertences, a pessoa não identificada solicitou auxílio de autoridades militares no local e deu "voz de prisão" ao paciente, sem que houvesse qualquer comportamento ilegal ou repreensivel do Sr. Altamiro, pois não distribuía 'santinhos' ou fazia qualquer outro tipo de propaganda politica.
Aduz que o 'paciente foi denunciado por infração ao artigo 347 (desobediência) e não por algum tipo de propaganda eleitoral irregular. Tece considerações sobre a inexistência da configuração dos elementos do tipo previsto no dispositivo legal citado.
Ao final defende a necessidade do deferimento de liminar a fim de suspender o curso da ação penal em trâmite perante a 1ª Zona Eleitoral, uma vez que no dia 13 de setembro de 2007, às 15h, o paciente deverá comparecer a audiência de instrução e julgamento para prestar depoimento.
O pedido de liminar foi indeferido, consoante decisão proferida às fis. 24/26.
Dispensadas as informações, os autos foram encaminhados ao Procurador Regional Eleitoral para parecer.
Por sua vez, o Ministério Público, às fis. 30/34, manifestou-se **contrariamente à concessão da ordem** de habeas corpus, por não vislumbrar ilegalidade ou abuso na ação penal intentada.
Consoante decisão de fi. 36, determinei que fosse oficiado ao juiz de primeiro grau para que informasse o andamento da ação penal.
À fI. 38, informou a Chefe da Seção de Processamento que, a Ação Penal A-0010/2007 já foi sentenciada, tendo sido inclusive objeto de recurso (Recurso Criminal n° 460) distribuído ao ilustre Juiz Frederico Bernar Vasconcelos.

É o breve relato.

VOTOS
O Senhor Desembargador ESTEVAM MAIA - relator: Cuida-se de *habeas corpus,* com pedido de concessão de liminar, impetrado por Manoel Ninaut filho, em face de ato do ilustre Juiz da 1ª Zona Eleitoral do Distrito Federal, **consubstanciado no recebimento de denúncia** em face da prática, *in thesi,* do tipo penal previsto no art. 347 do Código Eleitoral.
Alega o paciente, em suma, que a ação penal ofertada não teria justa causa, uma vez que a conduta descrita na peça acusatória não configura crime eleitoral.
Em que pesem os esforços empreendidos pelo impetrante, bem como o bem lançado parecer do representante do Ministério Público Eleitoral, o mérito do presente feito não pode ser enfrentado.
É que, conforme informação constante de fI. 38, a Ação Penal A-0010/2007 já foi sentenciada, dando origem, inclusive a Recurso Criminal n° 460, dirigido a este Tribunal Eleitoral, tendo como relator o Juiz Frederico Bernardes Vasconcelos.
Diante disso, verifica-se a perda do objeto do presente pedido de *habeas corpus,* em face da ausência de interesse processual.
Se o escopo pretendido pela impetração é o trancamento da ação penal, inexoravelmente ocorreu a perda superveniente do objeto com a prolação da sentença, posto que esgotada a jurisdição do primeiro grau. Portanto, inexiste ato passível de controle por meio da presente via.
Nesse sentido tem se posicionado o colendo Tribunal de Justiça do Distrito Federal, consoante ementa ora transcrita:
"HABEAS CORPUS. CRIME DE FALSA IDENTIDADE (ART. 307 DO CÓDIGO PENAL). DENÚNCIA, AUSÉNCIA DE JUSTA CAUSA. TRANCAMENTO. AÇÃO PENAL. SENTENÇA CONDENATÓRIO. PERDA SUPERVENIENTE DO INTERESSE PROCESSUAL. PEDIDO PREJUDICADO.
A prolação de sentença na ação penal que se pretende trancar por falta de justa causa em sede de habeas corpus, faz desaparecer o objeto da proteção jurídica vindicada pelo remédio heróico e, por conseguinte, a perda do interesse processual.
HC extinto sem exame do mérito, em face da perda superveniente do objeto. (HC, acórdão n° 280352; relator Sandoval Oliveira; publicação em *18/09/2007,* pág. 144)
Ante o exposto, **julgo extinto o processo sem enfrentamento do mérito**, em face da perda superveniente do objeto.
O Senhor Juiz JOSÉ LUIZ DA CUNHA FILHO - vogal: Acompanho o relator.
O Senhor Desembargador Federal CARLOS FERNANDO MATHIAS DE SOUZA vogal:. Acompanho o relator.
O Senhor Juiz ROBERVAL CASEMIRO BELINATI - vogal: Acompanho o relator.
O Senhor Juiz SILVÂNIO BARBOSA DOS SANTOS - vogal: Acompanho o relator.
O Senhor Juiz FREDERICO BERNARDES VASCONCELOS - vogal: Acompanho o relator.
DECISÃO
prejudicado o pedido. Decisão unânime. Em 4 de dezembro de 2007.

ANEXO C – Formato padrão de um tópico

Número: ___ (01..25)

Título: _____ (resume o tópico)

Descrição: _____

_____

_____ (descreve a necessidade de informação)

Narrativa: _____

_____

_____

_____ (apresenta critérios que caracterizam a relevância do documento)

Identificação dos casos relevantes: _____, _____, _____, _____, _____ (número do acórdão)

_____

Nome e assinatura

ANEXO D – Instrumento de coleta de dados

Objetivo: Avaliação do nível de precisão no resultado da busca em sistema de recuperação de casos jurídicos

1. Especialista: _____

2. Caso de entrada utilizado:

Ponto de acesso	**Descritores**
Fato	
Matéria	
Entendimento	
Argumento	

3. Identificação do tópico utilizado: _____
4. Qtd de casos recuperados: _____
5. Qtd de casos relevantes: _____

6. Identificação e julgamento dos casos recuperados:

Seqüência	Identificação do caso similar	Julgamento de relevância

_____

Nome e assinatura

ANEXO E – Formulário de indexação: Modelo de RBC

1) Descrição física do acórdão

Nº do acórdão/resolução:_____ data: __/__/__ tipo/nº do proc:____UF:____

2) Descrição de conteúdo do acórdão

Pontos de acesso	Termos índices
Fato	
Matéria	
Entendimento	
Argumento	

3) Campo referência legislativa

_____

4) Campo doutrina

_____

_____

Nome e assinatura

ANEXO F – Tabela de dados estatísticos consolidados (TDEC)

Tópico	DocRec	DocRel	Prec-R	MPP	Média
Média					

Legenda:

DocRec – Total de documentos recuperados no tópico

DocRel – Total de documentos recuperados relevantes no tópico

Prec-R – Precisão R com nível de corte 10 (k=10)

RkgDocRel – Ranking dos documentos relevantes recuperados

MPP – Média principal de precisão