Recuperação Vertical de Informação

Um Estudo de Caso na Área Jurídica

Maria de Lourdes da Silveira

Maria de Lourdes da Silveira

Recuperação Vertical de Informação

Um Estudo de Caso na Área Jurídica

Tese de doutorado apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do grau de doutor em Ciência da Computação.

Orientador - Professor Berthier Ribeiro-Neto

Belo Horizonte, 10 de abril 2003

Ao

Professor José Chagas da Silveira, meu irmão primogênito,

por ser, para seus irmãos, o precursor do caminho do Campo à cidade.

À Carmélia Leonel de Oliveira, minha irmã,

por ser o esteio da família na cidade.

Agradecimentos

Agradeço ao professor Berthier Ribeiro-Neto por sua ajuda sem a qual seria impossível a conclusão desse trabalho.

Aos membros da banca, professores Imre Simon, Fábio Cozman, Nivio Ziviani e Wagner Meira por sua colaboração.

À Diretoria da Prodabel, nas pessoas dos Srs. Gustavo da Gama Torres e Eugênia Bossi, pela concessão da licença sem vencimentos que possibilitou a conclusão do curso.

Ao professor Nivio Ziviani por ceder gentilmente o Latin – Laboratório para Tratamento da Informação – para execução dos experimentos.

Agradeço aos professores Henrique Pacca, Alberto Laender e Carlos Camarão pelas "dicas" essenciais em momentos de fragilidade.

À Renata pela presteza e eficiência em seu trabalho na Secretaria do curso.

Aos amigos do Latin Pável, Claudine, Marco, Charles, Wesley, Tânia, Cristiane e Álvaro pela troca de experiências, pela colaboração e companhia nas muitas horas de convívio.

Aos amigos Ilmério, Altigran e Andrea pelas discussões teóricas sobre o meu trabalho.

Aos amigos Rodrigo Tôrres e Rodrigo Vale pela implementação do sistema utilizado nos experimentos e pelas discussões conceituais sobre a área jurídica.

Ao Juliano Veloso pelas discussões conceituais sobre a área jurídica, por sua colaboração na elaboração das consultas e na avaliação dos documentos relevantes da coleção Juris.

Aos amigos da qualificação Lucila, Silvio, Mark e Fátima pelo companheirismo naqueles momentos que imaginávamos serem os mais difíceis. Vã ilusão.

Aos amigos Karla, Juliana, Nazaré, Marcus Vinícius, Sônia, Patrícia e Murilo pelo carinho e atenção durante esse tempo.

Ao Jivan Aamin pelo apoio emocional e espiritual e pela alegria de viver e sonhar com uma vida melhor.

À minha família, especialmente à minha irmã Carmélia, pelo apoio constante e infinita compreensão pelas freqüentes ausências em inúmeros acontecimentos familiares ao longo de todos esses anos.

Finalmente, agradeço aos mestres do universo pela eterna fonte de sabedoria compartilhada em todos os momentos.

"... o real não está na saída nem na chegada: ele se dispõe para a gente é no meio da travessia."

João Guimarães Rosa

Grande Sertão: Veredas

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Os sistemas de Recuperação de Informação (RI) foram criados para facilitar o acesso à informação em bibliotecas digitais modernas. Esses sistemas permitem organizar, indexar e recuperar informação sobre os documentos de uma coleção. Nesse processo, a tarefa de indexação tem importância central. Para indexar os documentos é necessário identificar termos de indexação que reflitam o conteúdo semântico dos documentos e que possam ser lembrados pelos usuários no momento de efetuar as consultas.

Para normatizar o processo de seleção de termos de indexação, foram criados os tesauros. Tesauros são ferramentas desenvolvidas tanto para controlar e padronizar a linguagem de indexação, quanto para facilitar a construção e reformulação de consultas pelos usuários. Tesauros são, em geral, específicos para uma dada área do saber, tais como a médica e a jurídica, dentre outras.

Dado um tesauro para uma área específica, pergunta-se: é possível construir um sistema de RI especializado para aquela área do saber que permita a obtenção de resultados de qualidade superior? O objetivo deste trabalho é produzir uma resposta para essa questão.

Nossa proposta se baseia no projeto de um sistema de *Recuperação Vertical de Informação* que, utilizando-se do conhecimento codificado em um tesauro, facilite ao usuário encontrar informação de seu interesse. Para combinar informação evidencial proveniente da consulta do usuário (isto é, as palavras-chave especificadas pelo usuário) com informação evidencial proveniente do tesauro, adotamos o arcabouço teórico das redes Bayesianas. O resultado é um modelo de rede de crenças que leva a uma nova fórmula de ordenação dos documentos na resposta.

Para validar nosso sistema, realizamos um estudo de caso voltado para a área jurídica. Utilizando uma coleção de referência composta por jurisprudências dos tribunais federais brasileiros e o Tesauro da Justiça Federal, avaliamos nosso sistema de Recuperação Vertical de Informação. Nossos resultados indicam que a combinação de informação proveniente das consultas com informação proveniente do tesauro possibilita ganhos em precisão da ordem de 31%.

Abstract

Information Retrieval (IR) systems were created to facilitate the access to information in modern digital libraries. These systems organize and index the documents of the collection, so that information about them can be retrieved. In this process, the indexing task is of central importance. To index the documents, it is necessary to identify indexing terms that reflect the semantic content of the documents and that are likely to be remembered by the users at querying time.

To standardize the process of selecting the index terms, the thesauri were created. A thesaurus is a tool that allows standardizing the indexing language and facilitating the construction and reformulation of queries by the users. Thesauri are, in general, specific to a given knowledge area, such as the medical and juridical areas, among others.

Given a thesaurus to a specific area, one can ask: is it possible to build an IR system, specialized to that area of knowledge, that will generate higher quality results? The objective of this work is to produce a response to this question.

Our proposal is based on the design of a *Vertical Information Retrieval* system that, using the knowledge encoded in a thesaurus, facilitates the task of finding information of interest. In order to combine evidential information in the user query (i.e., the keywords specified by the user) with evidential information in the thesaurus, we adopt the theoretical framework of Bayesian networks. The result is a belief network model that yields a new equation for ranking the documents in the response set.

To validate our system, we carried out a case study in the juridical area. Using a reference collection composed of jurisprudences of Brazilian federal courts and the Federal Justice Thesaurus, we evaluated our Vertical Information Retrieval system. Our results indicate that the combination of information from the user queries with information from the thesaurus leads to gains in average precision of the order of 31%.

Capítulo 1

Introdução

O problema de acessar documentos de interesse em uma biblioteca existe desde a criação das primeiras bibliotecas, desde a edição dos primeiros livros. Para facilitar o acesso aos documentos[1], foram criadas, ao longo dos anos, estratégias de coleta e organização de informação sobre os mesmos. Para grandes acervos, é necessária a criação de ferramentas computacionais que auxiliem os bibliotecários na catalogação, indexação e organização dos documentos, como os sistemas de Recuperação de Informação (RI). Esses sistemas têm outra função importante que é auxiliar os usuários a encontrar documentos de seu interesse, particularmente em grandes bibliotecas.

As bibliotecas tradicionais são compostas por documentos em papel, aos quais estão associadas informações pré-definidas, tais como autor, editora, título, etc. Contam ainda com a presença de bibliotecários que, com seu conhecimento sobre a organização da biblioteca, operam como intermediários e intérpretes, ajudando os usuários a descobrir os documentos que tratam de assunto de seu interesse.

As bibliotecas digitais, por sua vez, são compostas por documentos em formato digital, que são, na maioria das vezes, sem estrutura ou semi-estruturados. Tais bibliotecas possibilitam o acesso ao conteúdo de seus documentos a partir de locais remotos, comumente via *Web*. Nesse caso, o usuário não conta com a ajuda dos bibliotecários e tem que, ele mesmo, traduzir sua necessidade de informação em uma consulta para o sistema, consulta essa que deve ser expressa na linguagem do sistema.

Tanto em bibliotecas tradicionais como em digitais, a indexação dos documentos, que pode ser feita de forma manual ou automática, é um processo essencial, pois o índice facilita o acesso ao conteúdo dos documentos. No processo de inde-

[1]O termo documento é aqui utilizado para designar indistintamente os vários elementos de uma coleção de ítens de texto, tais como livros, artigos, revistas, páginas *Web*, documentos jurídicos, etc.

xação manual, os bibliotecários fazem uma análise conceitual do conteúdo de cada documento e elegem *termos de indexação*, ou *palavras-chave*[2], que são associados ao documento. Os termos de indexação fornecem uma visão lógica do documento e constituem apontadores para o seu conteúdo. Quanto maior o número de termos de indexação associados a um documento, maior é a possibilidade de que seja recuperado, ou seja, maior a sua *revocação*[3]. Por outro lado, a *precisão* do resultado da pesquisa tende a ser melhor quanto menor o número de termos de indexação associados a um documento [29]. Esse é um ponto a se considerar, pois existe uma relação inversa entre precisão e revocação. Um ganho em precisão é, em geral, acompanhado por uma perda em revocação e vice-versa [1].

A dificuldade do indexador está em escolher termos mais prováveis de serem utilizados pela comunidade de usuários e que, ao mesmo tempo, bem representem o conteúdo do documento, especialmente daqueles mais extensos, que tratam de vários assuntos. Como cada indexador carrega consigo suas características pessoais e tem preferências distintas em relação ao uso da linguagem, diferentes termos de indexação podem ser utilizados para representar o mesmo documento. Além disso, tal representação pode ser muito detalhada e específica em determinados aspectos e muito superficial em outros.

Para minimizar tais problemas, pode-se recorrer à padronização e à normatização da linguagem de indexação. Para tanto, foram criadas ferramentas como os vocabulários controlados, dentre eles os tesauros. Além de permitirem a padronização dos termos de indexação, os tesauros constituem uma base de conhecimento sobre a terminologia utilizada em uma dada área do conhecimento. Eles podem ser utilizados tanto pelos indexadores, na construção do índice, como pelos usuários, na construção ou reformulação de suas consultas.

Mesmo com a utilização da linguagem normatizada de um tesauro, o processo de indexação, quando efetuado manualmente, pode levar a inconsistências, perda de precisão e classificação errônea. Para evitar tais problemas, é necessário normatizar também o processo de indexação, o que pode ser feito através de sua automatização.

No processo de indexação automática em uma biblioteca digital, a seleção das palavras-chave, que representarão os documentos no índice, é feita através da análise automática do texto dos documentos. A partir de operações pré-definidas, os programas de indexação selecionam automaticamente palavras-chave contidas no texto para compor o índice. O vocabulário obtido pode ainda ser enriquecido com o de um tesauro e, nesse caso, seus conceitos podem funcionar como um classificador, pois

[2] Termo, termo de indexação ou palavra-chave serão utilizados indistintamente nesse trabalho.
[3] Os conceitos de revocação e precisão são apresentados na Seção 2.4.

permitem associar os documentos às suas diferentes classes.

Numa biblioteca digital, na qual os documentos são representados também por conceitos de um tesauro, dizemos que uma forma de conhecimento foi agregada à coleção. Quando essa agregação de conhecimento é voltada para uma área específica, como a jurídica e a médica, temos um sistema de recuperação de informação específico para a área do saber em questão — um *Sistema de Recuperação Vertical de Informação*.

1.1 Descrição do Problema

Na *Web*, um usuário eventual (não especialista) tende a formular consultas compostas por duas ou três palavras somente [14, 55]. Como conseqüência, uma longa lista de documentos é retornada como resposta e o trabalho do usuário passa a ser, então, selecionar aqueles que lhe interessam. Para evitar esse trabalho tedioso, o usuário pode tentar reformular sua consulta a partir de palavras-chave mais apropriadas. No caso em que o conhecimento de um tesauro foi utilizado para indexar os documentos, o usuário pode acrescentar conceitos desse tesauro à sua consulta original.

Entretanto, reformular a consulta de modo que a mesma forneça uma descrição mais detalhada do tópico de interesse do usuário, que descreva o "caso ou tarefa" de interesse o mais detalhadamente possível, não é um problema trivial. Freqüentemente o próprio usuário não tem uma visão completa do problema a ser resolvido, não conseguindo definir claramente sua necessidade de informação. Ela é difusa e se altera com o passar do tempo, à medida que ele adquire novas informações.

O objetivo desse trabalho é propor e avaliar soluções para o problema de ordenação dos documentos retornados como resposta a uma consulta de um usuário. Para isso, introduzimos um novo modelo de recuperação de informação, baseado no arcabouço teórico das redes Bayesianas. Esse modelo é utilizado para gerar uma nova fórmula de ordenação dos documentos na resposta.

A idéia é utilizar o conhecimento organizado de uma área específica, codificado na forma de um tesauro, como fonte de evidência[4] suplementar a ser acrescida ao conhecimento advindo do conteúdo dos documentos e das consultas dos usuários. Nossa proposta é a utilização do tesauro tanto na construção automática do índice quanto na composição das consultas. A partir da consulta inicial formulada pelo usuário, propomos melhorar a especificação da mesma, de forma totalmente automática, sem qualquer ajuda adicional do usuário.

[4]Evidência é uma informação adicional sobre um fato.

Em nossa proposta, o vocabulário do tesauro é tratado como fonte de evidência semântica suplementar sobre uma dada área do saber. Tal fonte de evidência é combinada com o conteúdo dos documentos através do arcabouço probabilístico das redes Bayesianas. O resultado é um modelo Bayesiano para uma área específica de conhecimento, aqui denominado *Sistema de Recuperação Vertical de Informação*. Essa idéia se justifica porque os tesauros constituem uma base de conhecimento sobre a terminologia mais apropriada em uma dada área, segundo os especialistas.

A verificação da qualidade do modelo que estamos propondo é feita a partir de um estudo de caso voltado para uma área particular do conhecimento, a saber, a área jurídica. Uma coleção de referência composta por jurisprudências[5] produzidas pelos tribunais superiores e pelos tribunais federais brasileiros[6] é utilizada em nossa validação experimental. O tesauro considerado é um tesauro jurídico brasileiro produzido pelo Conselho de Justiça Federal (CJF) para a Justiça Federal.

Os resultados de nossa avaliação sugerem que a adoção de um sistema de recuperação de informação especializado para uma área específica oferece vantagens sobre a adoção de um sistema de recuperação de informação não especializado, como as máquinas de busca tradicionais[7].

1.2 Estudo de Caso: O Problema de Recuperação Vertical de Informação Jurídica

Quando um advogado está compondo o processo de defesa de um novo caso, ele deve preparar sua argumentação de forma tão completa quanto possível, para melhorar as chances de obter um parecer favorável para o seu cliente. Ele necessita desenvolver argumentos sólidos para dar suporte à tese defendida no processo. Para isso, ele normalmente pesquisa por decisões em jurisprudências similares ao novo processo jurídico que irá impetrar. Também é importante encontrar leis que foram aplicadas a casos concretos similares ao caso de interesse e que tenham interpretações favoráveis.

O conteúdo de uma jurisprudência pode ser utilizado pelo advogado como argumento para melhor evidenciar seus pontos de vista sobre o direito que está sendo

[5] Jurisprudência é o conjunto de todos os processos já julgados sobre o mesmo assunto ou matéria e que representam a interpretação ou posição de um tribunal sobre o assunto.

[6] Uma breve introdução sobre as famílias de sistemas de direito e sobre a estrutura do Poder Judiciário no Brasil pode ser encontrada no Apêndice A.

[7] A expressão *máquina de busca tradicional* é utilizada para referenciar as máquinas de busca que não utilizam informação evidencial adicional.

reclamado. Ele pode utilizar os mesmos argumentos e conceitos de casos passados. Por outro lado, para os juízes também é importante encontrar processos passados, porque eles representam interpretações possíveis da lei aplicada a casos concretos em situações similares.

De um modo geral, na área jurídica é importante para o pesquisador encontrar todos os casos relacionados a um dado caso de interesse. Toda jurisprudência pode ser citada porque ela não se torna obsoleta e pode representar uma interpretação possível pelo tribunal que irá julgar o caso atual. O uso da jurisprudência ressalta a importância de se construir sistemas de RI específicos para a área jurídica.

1.3 Trabalho Proposto e Contribuições

A proposta desse trabalho é o projeto, a implementação e a validação experimental de um modelo de Recuperação Vertical de Informação que auxilie o usuário a encontrar mais facilmente os documentos de seu interesse. O modelo é avaliado experimentalmente através de um estudo de caso sobre uma coleção de documentos jurídicos do direito brasileiro.

As seguintes hipóteses são investigadas nesse trabalho:

- a utilização dos conceitos de um tesauro como evidência semântica melhora o resultado da pesquisa;
- a utilização dos relacionamentos entre os conceitos de um tesauro como evidência semântica melhora o resultado da pesquisa.

Para verificar tais hipóteses, os seguintes procedimentos foram executados:

- definição de um modelo de rede Bayesiana que permita combinar, em um mesmo arcabouço teórico, o resultado obtido através da utilização de palavras-chave provenientes do conteúdo dos documentos com os resultados obtidos utilizando informações evidenciais provenientes do conteúdo de um tesauro;
- implementação do modelo de rede Bayesiana proposto;
- avaliação do modelo proposto, mediante comparação com um algoritmo convencional de recuperação de informação, utilizando uma coleção de documentos jurídicos brasileiros e o tesauro jurídico da Justiça Federal.

As principais contribuições desse trabalho são a definição de um modelo de recuperação vertical de informação, que utiliza um tesauro como evidência semântica adicional, e sua avaliação experimental através de um estudo de caso voltado para a área jurídica.

1.4 Organização da Tese

No Capítulo 2 introduzimos o problema de Recuperação de Informação (RI) e descrevemos os componentes de um sistema de RI. A seguir apresentamos o modelo vetorial clássico para recuperação de informação e alguns conceitos utilizados para avaliar sistemas de RI. Ao final apresentamos alguns conceitos básicos da teoria de probabilidades.

No Capítulo 3 introduzimos o conceito de tesauros, os requisitos para a construção manual de tesauros e apresentamos alguns princípios para a avaliação da qualidade de um tesauro. A seguir apresentamos o Tesauro Jurídico da Justiça Federal e uma avaliação da sua qualidade.

No Capítulo 4 introduzimos as redes Bayesianas, o modelo de rede de crenças e sua adaptação ao modelo vetorial. A seguir apresentamos a utilização das redes Bayesianas na área de RI.

No Capítulo 5 apresentamos nosso modelo de rede de crença Bayesiana que permite combinar os resultados obtidos por palavras-chave com os resultados obtidos utilizando informações contidas no tesauro. Apresentamos a idéia da utilização do tesauro como evidência semântica, mostramos nosso modelo de rede e a equação derivada do modelo para calcular a ordenação do novo conjunto resposta. Ao final apresentamos a análise do custo computacional.

No Capítulo 6 apresentamos a coleção de referência e analisamos os resultados experimentais executados para verificar a qualidade do modelo proposto.

No Capítulo 7 apresentamos os trabalhos relacionados. As conclusões são apresentadas no Capítulo 8.

No Apêndice A introduzimos brevemente as maiores famílias de direito vigentes atualmente com o objetivo de situar o sistema de direito brasileiro. A seguir introduzimos a organização do Poder Judiciário no Brasil. No Apêndice B apresentamos algumas siglas e definições utilizadas nesse trabalho. No Apêndice C apresentamos o texto das consultas utilizadas nos experimentos e no Apêndice D apresentamos exemplos de conceitos do tesauro e seus relacionamentos.

Capítulo 2

Conceitos Básicos

Neste Capítulo introduzimos o problema de Recuperação de Informação (RI), descrevemos os componentes de um sistema de RI, apresentamos o modelo vetorial clássico para recuperação de informação, introduzimos alguns conceitos utilizados para avaliar sistemas de RI e alguns conceitos básicos da teoria de probabilidades.

2.1 O Problema de Recuperação de Informação

O problema de recuperar informação em um sistema de Recuperação de Informação (RI) pode ser dividido em duas partes. A primeira trata da constituição da coleção de documentos e da geração de um índice para a mesma. A segunda trata da utilização do sistema por seus usuários.

Uma coleção de documentos trata de assuntos de interesse da sua comunidade de usuários. Dada a coleção, é necessário que a mesma seja organizada de forma a facilitar o acesso aos documentos. Para tanto, um instrumento central em uma biblioteca, tradicional ou digital, é o índice. O índice pode ser construído manualmente, através da análise do conteúdo dos documentos por bibliotecários, ou automaticamente, por algoritmos que avaliam o conteúdo dos documentos e selecionam termos de indexação para compor o índice.

No processo de indexação manual, os bibliotecários analisam o conteúdo de cada documento e elegem termos de indexação para representar os documentos no índice. Devido a características e preferências pessoais que surgem no processo de análise do conteúdo dos documentos, diferentes termos de indexação podem ser selecionados para representar os mesmos documentos. Para evitar problemas advindos da variabilidade no uso da linguagem, ferramentas como os vocabulários controlados, dentre eles os tesauros, foram criadas ao longo dos anos. Durante a fase de indexação, essas ferramentas funcionam como uma espécie de vocabulário autorizado para orientar o

indexador na seleção dos termos de indexação. Na fase de pesquisa, essas ferramentas podem ser utilizadas para orientar o usuário na formulação e contextualização da consulta.

Nos sistemas de RI automatizados, o índice é gerado por algoritmos que analisam sintaticamente os documentos e extraem os termos de indexação para compor o índice, com base em operações previamente estabelecidas (ver Figura 2.1 e 2.2). Para cada palavra-chave, o sistema de RI gera uma lista de documentos nos quais a palavra-chave ocorre. Uma observação importante é que, em tal processo de indexação, a estrutura dos documentos é perdida, como também é perdida parte da semântica contida nos documentos.

O processo de acessar documentos em um sistema de RI começa quando o usuário se depara com uma necessidade de informação. Em contato com a interface, o usuário traduz sua necessidade de informação em uma consulta. A tarefa do sistema é, através do processamento da consulta, encontrar os documentos que possam lhe interessar.

Em uma biblioteca tradicional, os bibliotecários podem operar como intermediários, consultores ou intérpretes. Utilizando seus conhecimentos sobre a coleção e sobre o índice, eles traduzem a necessidade de informação, declarada pelo usuário, em uma consulta na linguagem específica do sistema. Nos sistemas automatizados com acesso via *Web*, entretanto, não existem tais intérpretes e o usuário deve fazer, ele mesmo, a tradução, utilizando seu conhecimento parcial sobre o assunto e sobre a coleção. No entanto, a transformação da necessidade de informação do usuário em uma consulta não é uma tarefa trivial, como passamos a discutir.

Segundo Gilchrist (Taylor [58] apud [19]), o processo de formulação de consultas se dá em quatro diferentes níveis ou etapas. No primeiro momento, a necessidade de informação pode não ter expressão em palavras, é uma necessidade visceral. No segundo, ela pode ser consciente, com expressão interna no cérebro, mas ainda sem palavras. No terceiro, ela pode ser formalizada em frases e expressões e, no quarto momento, pode ser transformada em palavras-chave. Ademais, a necessidade de informação se altera à medida que o usuário obtém as primeiras respostas e passa a conhecer melhor a coleção. Esse fenômeno se dá devido à característica dinâmica e evolutiva do conhecimento.

Uma dificuldade adicional é que a percepção da necessidade de informação pelo usuário se dá em linguagem natural, que permite a expressão de conteúdo semântico através do uso das estruturas da língua. Tal conteúdo é perdido na transformação da necessidade de informação em uma consulta, usualmente expressa como uma seqüência de palavras-chave, conectadas ou não por operadores booleanos. Além

disso, a necessidade de informação se dá em um contexto específico, que também é perdido na formulação da consulta. Outro aspecto é que, freqüentemente, o usuário conhece pouco ou nada sobre a coleção em que está executando a pesquisa, nem tão pouco conhece a constituição do vocabulário e as operações executadas quando o índice foi criado.

A possibilidade de um sistema de RI recuperar todos os documentos relevantes a uma necessidade de informação de um usuário é limitada, uma vez que o ato de encontrar documentos relevantes envolve incerteza. Por um lado, porque a necessidade de informação não é inteiramente especificada, pois ela se altera à medida que o usuário obtém as primeiras informações e, por outro, porque um documento pode atender apenas parcialmente a necessidade de informação representada na consulta. E, ainda, pela restrição da pesquisa ser feita por casamento de padrão, o que restringe o processamento ao uso das mesmas palavras na consulta e no índice.

Tais dificuldades são agravadas, pois os usuários tendem a construir suas consultas com duas ou três palavras-chave somente [14, 55], tornando-as pouco específicas. Como conseqüência, uma longa lista de documentos é apresentada ao usuário e sua tarefa passa a ser, então, selecionar aqueles que são relevantes à sua necessidade de informação. Nesse contexto, encontrar documentos de interesse pode exigir várias interações e reformulações da consulta inicial, tornando-se um processo trabalhoso. Uma alternativa para simplificar esse processo é expandir a consulta do usuário com termos adicionais que forneçam um melhor contexto para a consulta. Isso pode ser feito de duas formas principais: com uma abordagem interativa ou com uma abordagem automática. O que as diferencia é a participação ou não do usuário na seleção dos termos a serem utilizados na expansão da consulta original. Se efetuada de forma apropriada, a expansão da consulta pode levar a melhores respostas e reduzir o esforço despendido pelo usuário na seleção dos documentos de interesse.

As abordagens interativa e automática podem se utilizar de bases de conhecimento externos à coleção, como os tesauros. O objetivo de se utilizar um tesauro na reformulação de consultas é recuperar parte da semântica perdida na transformação da necessidade de informação em consulta. O tesauro possibilita ao usuário selecionar conceitos com algum relacionamento semântico ao texto da consulta, de acordo com sua estrutura. Pesquisas recentes com expansão de consultas utilizando os diferentes relacionamentos de um tesauro mostram resultados favoráveis [20, 21, 32, 39, 49], assim como em pesquisas anteriores [25, 27].

2.2 Componentes de um Sistema de Recuperação de Informação

Um sistema de RI lida com a organização, o armazenamento e o acesso aos itens de informação que compõem uma biblioteca digital e pode ser dividido em duas partes. A primeira trata da constituição da coleção de documentos, da sua organização e da geração de um índice para a mesma. A segunda trata da utilização do sistema por seus usuários. Uma visão geral dos componentes de um sistema de RI é apresentada na Figura 2.1.

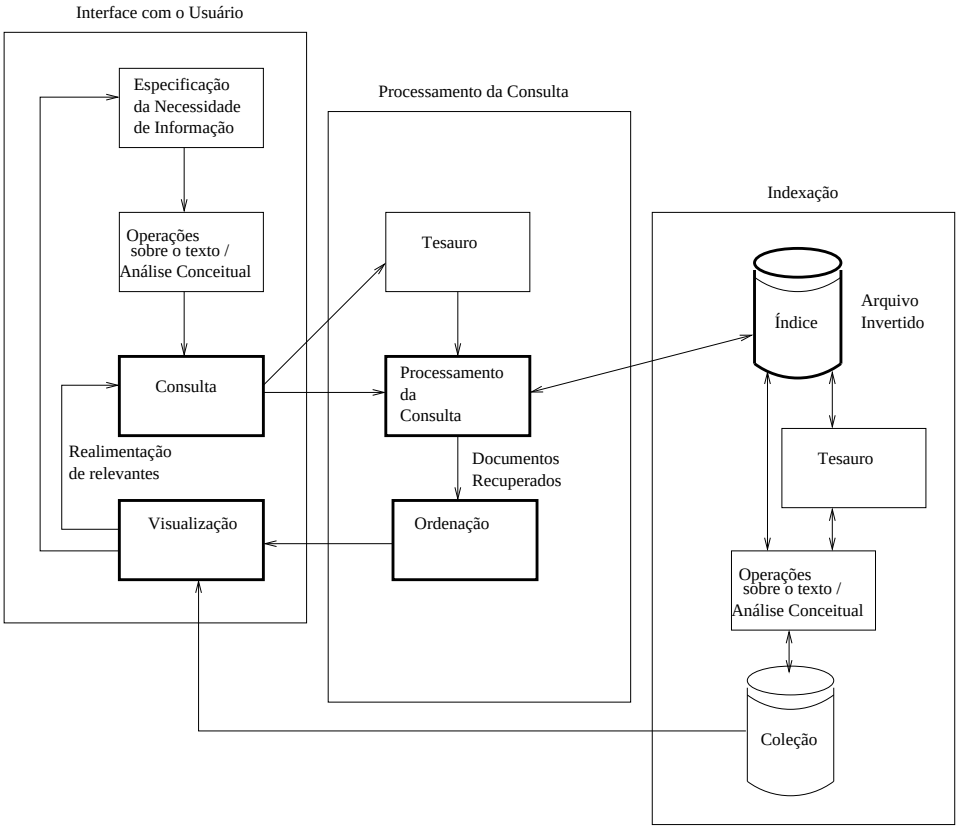

Figura 2.1: Componentes de um Sistema de RI

A constituição da coleção é o primeiro passo na criação de um sistema de RI. Em uma biblioteca digital, uma coleção é composta por um conjunto de documentos de interesse da comunidade de usuários, que pode ser sobre uma área específica do conhecimento humano, composta por documentos provenientes de uma fonte específica ou de várias fontes. A coleção pode também ser composta por documentos de diferentes áreas, como aquelas formadas a partir da Internet. Dada uma coleção, é necessário que a mesma seja organizada e armazenada de forma a facilitar o acesso

aos documentos. Para isso, um instrumento central em uma biblioteca é o índice. O processo de indexação é dividido em duas partes: a avaliação conceitual do conteúdo de cada documento e a tradução dessa análise no conjunto dos termos de indexação. Os termos de indexação selecionados representam o documento no índice, fornecem uma visão lógica dos mesmos e possibilitam o acesso aos documentos.

No processo de indexação manual, os bibliotecários analisam o conteúdo de cada documento e selecionam os termos de indexação para representá-los no índice. Os vocabulários controlados são ferramentas utilizadas na tradução do conteúdo dos documentos, com o objetivo de padronizar a linguagem de indexação e funcionam como uma espécie de vocabulário autorizado, para orientar os indexadores na seleção dos termos de indexação.

O projetista de um sistema de RI deve decidir se haverá controle ou não de vocabulário na indexação e/ou na pesquisa. Segundo Lancaster e Warner [30], existem quatro maneiras de se utilizar vocabulários controlados para representar documentos e necessidades de informação. A primeira é controlar o vocabulário na indexação (entrada) e na pesquisa (saída). A segunda é controlar o vocabulário na indexação e não controlar na pesquisa. A terceira é não controlar a indexação, mas sim a pesquisa, ou seja, utiliza-se um tesauro somente na pesquisa. A quarta é não exercer qualquer controle tanto na indexação quanto na pesquisa, o que caracteriza um sistema de RI de linguagem natural.

Nos sistemas de RI automatizados, o índice é gerado por algoritmos que analisam sintaticamente os documentos e extraem os termos de indexação para compor o índice. Esses algoritmos utilizam operações previamente estabelecidas, tais como operações para redução das palavras à sua raiz gramatical (*stemming*), exclusão de acentos, hífens, espaços em branco e exclusão de palavras sem valor de indexação (*stopwords*[1]). Em tais operações de transformação, a estrutura da língua é perdida e também parte da semântica contida nos documentos. Os vocabulários controlados permitem recuperar parte da semântica perdida, através da utilização de conceitos e seus relacionamentos. Essas operações são executadas para todos os documentos para obter as palavras-chave que formam o vocabulário da coleção. Para cada palavra-chave, o sistema de RI gera o conjunto de documentos onde a palavra-chave ocorre, na forma de uma lista invertida [5]. A figura 2.2 ilustra as possíveis operações a serem executadas no texto dos documentos e suas possíveis formas de representação no índice. Os sistemas de RI tradicionais não levam em

[1] *Stopwords* são termos de um idioma, considerados no âmbito de uma aplicação específica, que agregam pouca ou nenhuma semântica à sentença, tais como artigos, conjunções, preposições e conectivos.

conta a estrutura dos documentos.

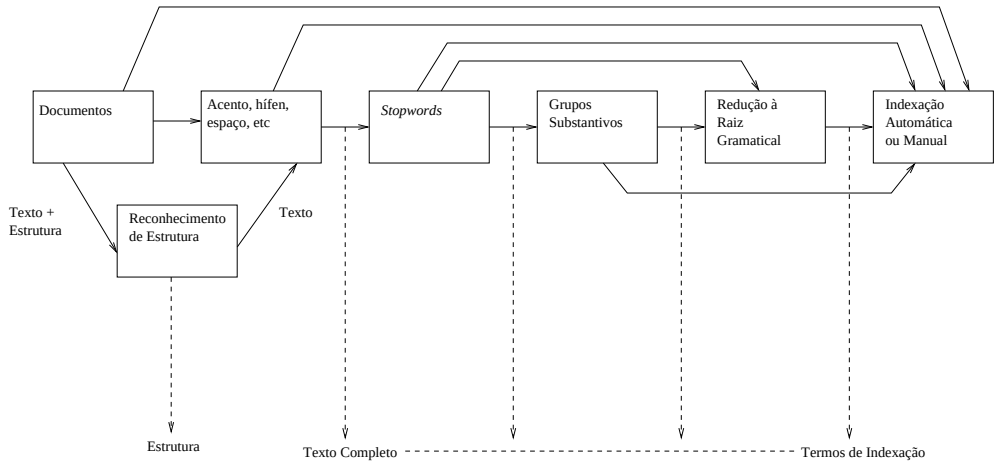

Figura 2.2: Operações para extração dos representantes de documentos (de [5])

O processo de acessar documentos em um sistema de RI se dá através da interface do sistema e começa quando o usuário se depara com uma necessidade de informação. Esse processo pode se dar de dois modos: o *browsing* ou o *searching*. O modo *browsing* ocorre quando o usuário tem uma idéia vaga sobre o assunto que deseja pesquisar. O usuário inicia sua pesquisa a partir de uma página conhecida e, seguindo as conexões (*links*) com outras páginas estabelecidas na página atual, prossegue sua pesquisa, ou seja, escolhendo as conexões que ele avalia estarem relacionados com sua necessidade de informação inicial. Esse procedimento pode ser demorado e pode levá-lo a se distanciar do seu objetivo inicial. Tal efeito é conhecido como *lost in hyperspace*.

O modo *searching* se dá quando o usuário, em contato com a interface, traduz sua necessidade de informação em uma consulta que é submetida ao sistema de RI. Como no processo de indexação, a necessidade de informação deve ser traduzida em uma consulta composta por palavras-chave a ser submetida ao sistema. Esse processo de tradução é conhecido como *estratégia de pesquisa*. Com base na consulta, a função do sistema de RI passa a ser, então, encontrar em seus índices os documentos mais prováveis de serem relacionados à necessidade de informação do usuário, expressa na consulta. Para tanto, o sistema deverá encontrar os documentos cujos termos de indexação coincidam com os da consulta, através de operações de casamento de padrões.

Uma tarefa central nos sistemas de RI é estimar a relevância de cada documento em relação à consulta de um usuário. Essa estimativa é feita associando-se um

número real a cada documento na resposta, usualmente referido como o *rank* do documento, o que permite ordená-los por similaridade com relação à consulta. A lista ordenada dos documentos da resposta é usualmente referida como *ranking*.

Diferentes algoritmos de ordenação são utilizados em sistemas de RI. Esses algoritmos podem ser baseados em álgebra Booleana, em modelagem probabilística e em representações vetoriais. O modelo vetorial é um modelo clássico e popular na área de RI e será utilizado como base de comparação nesse trabalho. A seguir apresentamos uma breve introdução desse modelo. Os demais modelos não serão abordados. Maiores informações podem ser encontradas em [5].

2.3 Modelo Vetorial

Os sistemas de RI tradicionais adotam palavras-chave para indexar os documentos e também para traduzir a necessidade de informação do usuário em uma consulta. A representação de documentos e consultas por palavras-chave é uma simplificação, pois parte da semântica dos documentos e da necessidade de informação é perdida quando ambos são transformados em um conjunto de palavras-chave. Outra simplificação que ocorre nos sistemas de RI atuais é que as palavras-chave são consideradas independentes, ou seja, conhecer alguma informação sobre um termo i nada implica sobre seu vizinho $i + 1$.

O modelo vetorial utiliza um espaço t-dimensional, onde t é o número de termos do vocabulário, para representar os documentos e as consultas como vetores. Operações algébricas são efetuadas sobre tais vetores para obter uma ordenação dos documentos por similaridade com relação à consulta [5, 45, 46].

No modelo vetorial, a consulta e os documentos são representados por vetores em um espaço t-dimensional, ou seja, $\vec{d_j} = (w_{1j}, w_{2j}, \ldots, w_{ij}, \ldots, w_{tj})$ e $\vec{q} = (w_{1q}, w_{2q}, \ldots, w_{iq}, \ldots, w_{tq})$, onde w_{ij} (w_{iq}) é um peso associado ao termo i no documento j (na consulta q) e é calculado como segue [47]:

$$w_{ij} = f_{ij} \times \log \frac{N}{n_i} \tag{2.1}$$

onde f_{ij} é a freqüência do termo i no documento d_j, N é o número de documentos da coleção e n_i é o número de documentos em que o termo i ocorre. A parcela $\log \frac{N}{n_i}$ é denominada *idf* - *inverse document frequency*. O valor f_{ij} é uma medida da importância do termo i no documento, ou seja, quanto mais freqüente é um termo, mais importante ele é para o documento, e *idf* é uma medida da importância do termo i na coleção, ou seja, quanto menos freqüente é um termo, mais importante ele é na coleção.

A similaridade entre a consulta e cada documento é calculada pela correlação entre os vetores $\vec{q}$ e $\vec{d_j}$. A correlação mais utilizada é o cosseno do ângulo formado entre os dois vetores, ou o produto interno entre eles, dado pela fórmula:

$$\begin{aligned} sim(\vec{d_j}, \vec{q}) &= \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} \\ &= \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \end{aligned} \tag{2.2}$$

onde $|\vec{d_j}|$ e $|\vec{q}|$ são as normas do documento e da consulta. A norma da consulta não influencia o resultado da ordenação e a norma do documento permite a normalização no espaço de documentos.

As principais vantagens do modelo vetorial são: implementação simples e largamente utilizada; utilização de pesos para os termos melhora o desempenho dos sistemas de RI; casamento parcial entre os termos da consulta e os termos dos documentos permitindo recuperar documentos com similaridade somente parcial com relação à consulta. A principal desvantagem é que o modelo adota, como premissa, a independência dos termos. Essa premissa é simplificadora, mas permite que o cálculo do peso e da similaridade sejam processados eficientemente.

2.4 Precisão e Revocação

Precisão e revocação relativas[2] são conceitos amplamente utilizados em RI para avaliar o desempenho de sistemas de RI e foram originalmente propostos por Kent et all. (Kent [26] apud [34]). São medidas utilizadas para avaliar a eficácia de um sistema de RI, ou seja, elas medem a habilidade do sistema de recuperar os documentos relevantes e, ao mesmo tempo, de evitar os não relevantes [66].

As seguintes definições são necessárias para se entender esses dois conceitos. Seja C o conjunto de documentos da coleção. Seja R o conjunto de documentos relevantes para uma dada consulta, identificado por um grupo de especialistas, e $|R|$ o número de documentos em R. Seja A o conjunto de documentos da resposta retornado pelo sistema de RI e $|A|$ o número de documentos em A. Seja Ra o conjunto de documentos relevantes do conjunto resposta A, resultado da interseção entre os conjuntos R e A, e seja $|Ra|$ o número de documentos no conjunto Ra. A Figura 2.3 ilustra esses conceitos e seus relacionamentos.

[2] Precisão e revocação *relativas* porque são valores calculados com base no número de documentos relevantes de uma dada consulta, número esse gerado a partir de uma amostra de documentos da coleção.

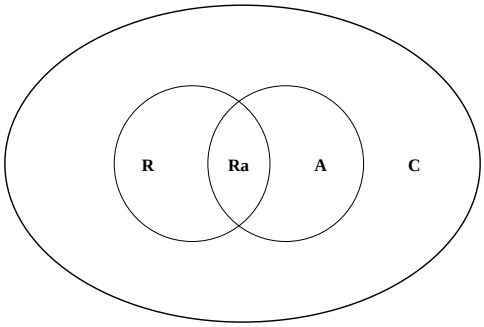

Figura 2.3: Conjuntos para definição de precisão e revocação

Precisão

Precisão é a fração de documentos recuperados que é relevante [5], ou seja, é uma medida da capacidade do sistema de recuperar somente documentos relevantes [69]. É calculada pela fórmula:

$$P = \frac{|Ra|}{|A|} \tag{2.3}$$

A precisão média para todas as consultas de um conjunto de consultas de teste é calculada pela fórmula:

$$PM = \left[\sum_{i=1}^{Nq} \frac{|Ra_i|}{|A_i|}\right] \times \frac{1}{Nq} \tag{2.4}$$

onde Nq é o número de consultas.

Revocação

Revocação é a fração de documentos relevantes recuperados [5], ou seja, é uma medida da capacidade do sistema de recuperar todos os documentos relevantes [69]. É calculada pela fórmula:

$$R = \frac{|Ra|}{|R|} \tag{2.5}$$

A revocação média para todas as consultas de um conjunto de consultas de teste é calculada pela fórmula:

$$RM = \left[\sum_{i=1}^{Nq} \frac{|Ra_i|}{|R_i|}\right] \times \frac{1}{Nq} \tag{2.6}$$

onde Nq é o número de consultas.

Ganho

O Ganho (G) é um número utilizado para comparar dois algoritmos, digamos a e b. O ganho de b em relação a a é dado pela equação:

$$G(b,a) = \frac{PM(b) - PM(a)}{PM(a)} \tag{2.7}$$

onde PM é a precisão média.

Precisão Média para Diferentes Níveis de Revocação

Os valores de precisão e revocação dados pelas fórmulas 2.3 e 2.5 são calculados assumindo-se que todos os documentos do conjunto resposta A foram analisados pelo usuário. No entanto, o usuário analisa os documentos a partir do topo da lista ordenada de documentos do conjunto resposta, o que implica que os valores de precisão e revocação variam à medida que o usuário prossegue a sua análise da lista ordenada. Uma forma mais comum de apresentar os resultados é gerar o gráfico de precisão versus revocação para vários níveis de revocação, ou seja, a precisão é calculada quando 10% dos documentos relevantes são analisados, quando 20% dos documentos relevantes são analisados, assim por diante, até que 100% dos documentos relevantes são analisados. Assim, esses gráficos são obtidos através do cálculo da precisão média em pontos padrão de revocação, tais como 10% (ou 0.1), 20% (ou 0.2), ..., 100% (ou 1.0). A curva de precisão versus revocação é também uma forma comum de apresentar e comparar os resultados de diferentes algoritmos de RI.

Para ilustrar, consideremos a lista ordenada de 20 documentos da Tabela 2.1 gerada em resposta a uma consulta q.

Tabela 2.1: Lista de Documentos Gerados em Resposta à Consulta q

1 - D_{203}	6 - D_{7}	11 - D_{80}	16 - D_{167}
2 - D_{202}	7 - D_{183}	12 - D_{81}	17 - D_{173}
3 - D_{310}	8 - D_{195}	13 - D_{82}	18 - D_{178}
4 - D_{415}	9 - D_{110}	14 - D_{95}	19 - D_{181}
5 - D_{620}	10 - D_{53}	15 - D_{152}	20 - D_{420}

Consideremos ainda que o número de documentos relevantes para a consulta q, de acordo com uma análise feita por especialistas, é 10. Esses 10 documentos são os

seguintes:

$$R_q = \{D_{202}, D_{310}, D_{415}, D_7, D_{195}, D_{53}, D_{82}, D_{95}, D_{152}, D_{420}\}$$

Para calcular a precisão nos diferentes níveis de revocação, procedemos como segue. Examinamos o primeiro documento na lista ordenada de documentos na resposta (1 - D_{203}) e constatamos que ele não é relevante (não está no conjunto R_q). Examinamos então o segundo documento na resposta (2 - D_{202}) e constatamos que ele é relevante. Nesse ponto temos então que:

- o primeiro documento relevante, dentre os 10 documentos relevantes em R_q, foi observado, ou seja, o nível de revocação é 10%;
- para observar o primeiro documento relevante, foi preciso examinar dois documentos, ou seja, a precisão no ponto de revocação 10% é de 50%.

Assim, o primeiro ponto da curva de revocação versus precisão é (10%, 50%).

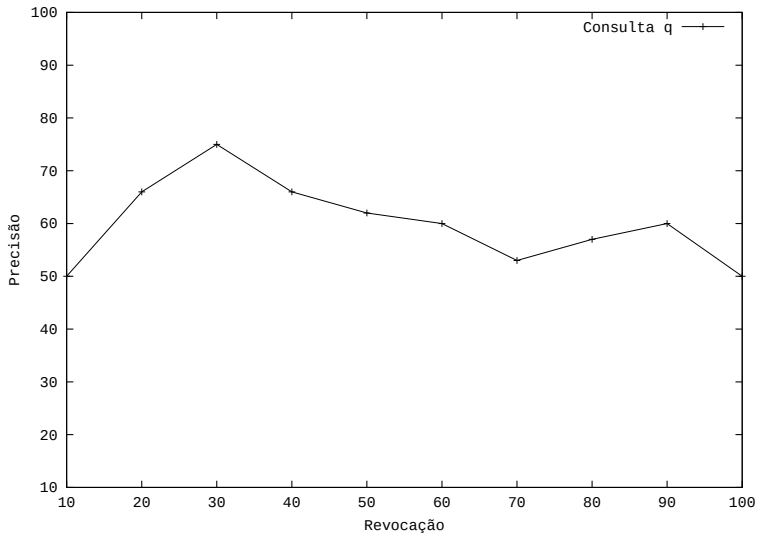

Figura 2.4: Exemplo da Curva de Precisão versus Revocação para a Consulta q

Continuando a análise, o terceiro documento da lista (3 - D_{310}) é um documento relevante. Para encontrar o segundo documento relevante, foi necessário analisar três documentos. Assim, o segundo ponto da curva de precisão versus de revocação é (20%, 66%). Se prosseguirmos examinando a lista ordenada de documentos na resposta, teremos 10 pontos de precisão-revocação que permitem traçar a curva da Figura 2.4.

A curva de precisão versus revocação pode também ser traçada para um número N_q de consultas de teste. Nesse caso, a precisão média $P(r)$ correspondendo ao nível de revocação r é calculada pela fórmula:

$$P(r) = \frac{\sum_{i=1}^{N_q} P_i(r)}{N_q}$$

onde $P(r)$ é a precisão média no ponto de revocação r, N_q é o número de consultas e $P_i(r)$ é o valor da precisão da consulta i no ponto de revocação r.

Interpolação de Precisão-Revocação

Os níveis de revocação para as várias consultas podem ser diferentes dos níveis de revocação padrão. Isso ocorre, por exemplo, quando o número de documentos relevantes no conjunto R_q é menor que 10. Nesse caso, utiliza-se a interpolação para calcular a precisão média nos níveis padrão de revocação. A precisão interpolada $P(r_i)$ no nível de revocação padrão r_i é definida pela equação: $P(r_i) = \max_{r_i \leq r \leq r_{i+1}} P(r)$, onde $r_i \in \{0.0, 0.1, 0.2, ..., 1.0\}$ é o ponto de revocação padrão, $P(r_i)$ é a precisão interpolada no nível r_i, $P(r)$ é o valor máximo da precisão conhecida em qualquer ponto de revocação entre os níveis r_i e r_{i+1}.

Exemplificamos o conceito de interpolação a seguir. Suponhamos que uma consulta q_2 obtenha como resultado a lista apresentada na Tabela 2.1 e que seu conjunto de documentos relevantes seja:

$$R_{q_2} = \{D_{152}, D_{195}, D_{202}\}$$

Nesse exemplo, o primeiro documento relevante da lista ordenada da Tabela 2.1 é 2 - D_{202}, que fornece um nível de revocação de 33,33% (um dentre três documentos relevantes foi analisado). A precisão nesse ponto é de 50% (um em dois documentos analisados é relevante). O segundo documento relevante é o 8 - D_{195}, que fornece um nível de revocação de 66,66% (dois dentre três documentos relevantes foram analisados). A precisão é de 25% (dois em oito documentos analisados são relevantes). O terceiro documento relevante é o 15 - D_{152}, que fornece um nível de revocação de 100% (três dentre três documentos relevantes foram analisados). A precisão é de 20% (três em quinze documentos analisados são relevantes).

Os valores de precisão para os pontos de revocação padrão são interpolados como segue. A precisão interpolada nos pontos 0%, 10%, 20% e 30% é de 50%, que é a precisão conhecida no ponto de revocação 33,33%. A precisão interpolada nos pontos 40%, 50% e 60% é de 25%, que é a precisão conhecida no ponto de revocação 66,66%. A precisão interpolada nos pontos 70%, 80%, 90% e 100% é de 20%, que é a precisão conhecida no ponto de revocação 100%. Esse exemplo gera o gráfico da Figura 2.5.

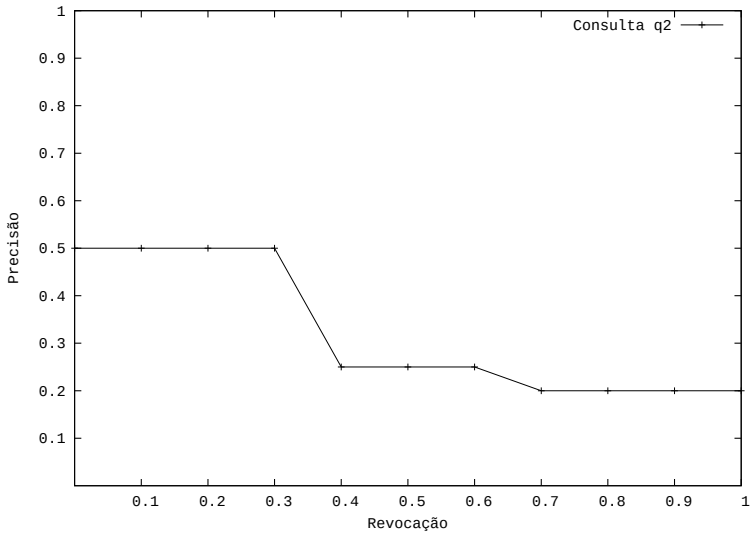

Figura 2.5: Exemplo da Curva de Precisão versus Revocação para a Consulta q_2 com interpolação

2.5 Teoria das Probabilidades

No mundo em que vivemos lidamos constantemente com a incerteza, ou por não possuirmos informações completas sobre os fatos que nos cercam, ou mesmo por desconhecermos alguns deles. Consciente ou inconscientemente, tomamos decisões com graus de crença baseados em fatos passados ou em regras gerais. Quando dizemos, por exemplo, que a probabilidade de acontecer tal fato é de 80%, nós estamos exprimindo um grau de crença ou expectativa em que tal fato irá acontecer.

As duas correntes mais importantes na área de Probabilidades são as correntes freqüentista e epistemológica. A corrente freqüentista defende a posição de que números que representam as probabilidades são provenientes de experimentos, os quais estão relacionados às leis de chance (*Laws of chance*). A corrente epistemológica interpreta os números como graus de crença que podem ser obtidos sem experimentação. O modelo de redes de Bayesianas, descrito na Seção 4 e utilizado nesse trabalho, advém da corrente epistemológica, pois utiliza graus de crença no cálculo das probabilidades.

Antes de coletarmos evidências sobre um fato, nós falamos em probabilidade incondicional ou anterior. Depois de obtermos alguma evidência, falamos de probabilidade condicional ou posterior. A probabilidade incondicional, denotada por P(A), é a probabilidade da proposição A ser verdadeira, se houver ausência de qualquer outra informação. A probabilidade condicional, denotada P(A|B), é a probabili-

dade de A ser verdadeira dada a ocorrência de B. A probabilidade condicional pode ser expressa em termos das probabilidades incondicionais, $P(A|B) = \frac{P(A \wedge B)}{P(B)}$, para $P(B) > 0$, ou ainda, $P(A \wedge B) = P(A|B) \times P(B)$. Para $A \wedge B$ ser verdadeiro, B deve ser verdadeiro e então A ser verdadeiro dado B. Escrito de outra forma, $P(B|A) = \frac{P(A \wedge B)}{P(A)}$, para $P(A) > 0$, ou ainda, $P(A \wedge B) = P(B|A) \times P(A)$. Para $A \wedge B$ ser verdadeiro, A deve ser verdadeiro e então B ser verdadeiro dado A. Essas proposições são as duas formas da regra do produto.

Axiomas da probabilidade

1. Todas as probabilidades estão entre 0 e 1, ou seja, $0 \leq P(A) \leq 1$;

2. Proposições necessariamente verdadeiras têm probabilidade 1 e proposições necessariamente falsas têm probabilidade 0, ou seja, $P(verdadeiro) = 1$; $P(falso) = 0$;

3. A probabilidade da disjunção é dada por: $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$.

Desses três axiomas podemos derivar todas as propriedades de probabilidade [44]. Se fizermos, por exemplo, $P(B) = P(\neg A)$ no axioma 3, temos:

$$\begin{aligned} P(A \vee \neg A) &= P(A) + P(\neg A) - P(A \wedge \neg A) \\ P(verdadeiro) &= P(A) + P(\neg A) - P(falso) \\ 1 &= P(A) + P(\neg A) - 0 \\ P(A) &= 1 - P(\neg A) \end{aligned}$$

que é a probabilidade da negação de uma proposição em termos da probabilidade da própria proposição.

Distribuição Conjunta de Probabilidade

Um modelo probabilístico de um domínio consiste de um conjunto de variáveis aleatórias que podem ter valores particulares com certas probabilidades. A distribuição conjunta de probabilidade (*Joint Probability Distribution*) especifica completamente todas as proposições do domínio. Um evento atômico é uma especificação completa do estado do domínio, ou seja, uma atribuição de valores particulares para todas as variáveis. Sejam as variáveis aleatórias $X_1, X_2 \cdots X_n$. A distribuição conjunta de probabilidade $P(X_1, \cdots X_n)$ atribui probabilidades para todos os possíveis eventos atômicos. A distribuição conjunta de probabilidade é uma tabela

n-dimensional na qual cada célula fornece a probabilidade de que tal estado específico ocorra. $P(X_i)$ é um vetor uni-dimensional de probabilidades para todos os possíveis valores da variável X_i.

Teorema de Bayes

Pelas duas formas da regra do produto temos:

$$P(A \wedge B) = P(A|B)P(B),$$

$$P(A \wedge B) = P(B|A)P(A).$$

E podemos escrever:

$$P(A|B)P(B) = P(B|A)P(A)$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Dada uma evidência E, podemos reescrevê-la:

$$P(B|A, E) = \frac{P(A|B, E)P(B|E)}{P(A|E)}$$

Capítulo 3

Tesauros

Os tesauros são ferramentas de vocabulário específicas para um domínio do conhecimento, construídas por especialistas para representar conceitos de tal domínio e especificar seus relacionamentos. Os relacionamentos mais comumente representados são os de equivalência, de hierarquia e de associatividade. Nos relacionamentos de equivalência, os conceitos são sinônimos ou sinônimos parciais. Nos de hierarquia, os conceitos possuem relação de subordinação, como termos genéricos e termos específicos. Nos de associatividade, os conceitos possuem um relacionamento horizontal, diferente dos anteriores, conforme determinado pelos especialistas da área em questão [29]. Exemplos de conceitos e relacionamentos podem ser encontrados no Apêndice D.

Um tesauro é uma ferramenta que pode ser utilizada tanto no processo de indexação dos documentos quanto no de pesquisa por informação. Para tal, deve ser uma ferramenta que facilite, ao indexador, representar o conteúdo de um documento no índice e, ao usuário, representar a sua necessidade de informação. Na indexação dos documentos da coleção, o tesauro pode ser utilizado como uma lista de palavras autorizadas para normatizar a linguagem de indexação. Durante a fase de pesquisa, o tesauro pode ser utilizado para mostrar ao usuário associações entre conceitos que podem conduzi-lo a outro conceito útil na formulação da consulta [19]. Outro uso do tesauro é como ferramenta para ajudar o usuário a situar melhor o assunto ou o contexto de sua consulta [49].

Os vários elementos de um tesauro são ferramentas apropriadas, ou para melhorar a revocação, garantindo que o maior número de documentos relevantes sejam recuperados, ou para melhorar a precisão, garantindo que os documentos não desejados sejam excluídos. Os elementos do tesauro apropriados para melhorar a revocação são [1, 19]:

- Conceitos (CC): devem representar o mais amplamente possível a área de

conhecimento em questão;

- Sinônimos (SY): indicam os termos preferenciais ou descritores (USE) e os não-preferenciais ou não-descritores (UP);
- Termos específicos (TE) e termos genéricos (TG): podem introduzir termos próximos relacionados hierarquicamente;
- Termos relacionados (TR): termos próximos conceitualmente, mas sem relação de hierarquia, que podem levar à lembrança de outros termos associados.

Os elementos apropriados para melhorar a precisão são:

- Especificidade da linguagem do tesauro: quanto mais detalhado for o tesauro, mais precisamente os assuntos poderão ser descritos;
- Nível de coordenação: nível de combinação dos conceitos do tesauro, que podem ser utilizados tanto na indexação quanto na pesquisa;
- Indicadores de relação: associações feitas no momento de indexação manual, no contexto de um documento específico, que evitam combinações indevidas em tempo de pesquisa;
- Peso: diferenciação entre conceitos de maior ou menor relevância em documentos particulares.

3.1 Formato das Entradas de um Tesauro

Um tesauro é um vocabulário controlado, organizado em uma estrutura definida de tal forma que os relacionamentos entre os conceitos devem estar claramente identificados. Apresentamos na Figura 3.1 um esquema da estrutura que deve ser seguida na apresentação dos conceitos de um tesauro.

Os campos dessa estrutura representam:

- **(id)**: define a identificação de um descritor (conceito) no tesauro;
- **<nome-do-descritor-A>**: nome do conceito identificado por (id);
- **UP**: utilizado para definir, no relacionamento de equivalência, o descritor não-preferido de <nome-do-descritor-A>, ou seja, define um sinônimo que deve ser evitado;
- **<nome-do-descritor-B>**: nome do conceito não-preferido;

- **TG1**: utilizado para definir, no relacionamento de hierarquia, um conceito genérico ao conceito <nome-do-descritor-A>;
- **<nome-do-descritor-genérico>**: nome do conceito genérico;
- **TGm**: utilizado para definir conceitos genéricos em níveis crescentes de generalização;
- **TE1**: utilizado para definir, no relacionamento de hierarquia, um conceito específico ao conceito <nome-do-descritor-A>;
- **<nome-do-descritor-específico>**: nome do conceito específico;
- **TEn**: utilizado para definir conceitos específicos em níveis crescentes de especificidade;
- **TR**: utilizado para definir, no relacionamento de associatividade, um conceito relacionado ao conceito <nome-do-descritor-A>;
- **<nome-do-descritor-relacionado>**: nome do conceito relacionado.
- **USE**: utilizado para definir, no relacionamento de equivalência, o descritor preferido de <nome-do-descritor-B>, ou seja, define que o sinônimo <nome-do-descritor-A> deve ser utilizado em vez de <nome-do-descritor-B>;

3.2 Histórico e Requisitos para a Construção Manual de Tesauros

Ao longo dos anos, com a experiência na organização de acervos e na criação de seus índices, tem-se observado uma grande variabilidade na seleção dos conceitos e termos representantes dos documentos no índice. Essa variação surge das diferenças e preferências pessoais dos indexadores em relação ao vocabulário a ser utilizado para representar o conteúdo dos documentos. Essa constatação sugere a necessidade de se exercer controle sobre o vocabulário a ser utilizado pelos indexadores para garantir consistência no serviço de indexação.

Os Vocabulários Controlados surgiram para atender originalmente a dois tipos de requisições: classificação bibliográfica, de origem inglesa, e indexação alfabética de assuntos, de origem americana [28]. O primeiro tesauro foi construído pelo Departamento de Defesa Americano, no centro de informação ASTIA - *Armed Services Technical Information Agency*, em 1960. As primeiras regras para a construção

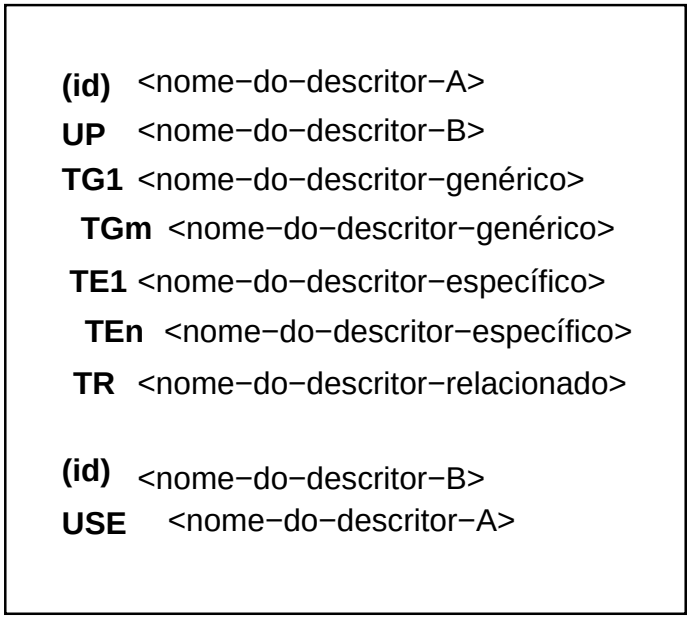

Figura 3.1: Estrutura de Apresentação das Entradas de um Tesauro

de tesauros foram publicadas em 1962. Em 1970, a UNESCO publicou a primeira versão do *Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri*, que foi a base para o primeiro padrão internacional de construção de tesauros, a ISO 2788-1974 (*International Organization for Standardization*), revisado em 1986 [24]. Tais padrões definem princípios gerais que se aplicam a todos os domínios de conhecimento e devem ser seguidos na construção de qualquer tesauro.

Os Vocabulários Controlados podem ser uma lista de cabeçalhos de assunto, um esquema de classificação, um tesauro ou, simplesmente, uma lista de frases e palavras autorizadas [29], que sejam reconhecidos de forma genérica e independente de documentos específicos. Na sua forma mais simples, os tesauros podem ser uma lista de termos com uma relação semântica entre eles como, por exemplo, os sinônimos, que podem ter uma forma padronizada de uso e os homógrafos, que podem ser diferenciados por uma descrição entre parênteses. Em sua forma mais complexa, os tesauros devem representar os conceitos e os relacionamentos existentes entre eles.

Os conceitos, representados por descritores, podem ser de vários tipos, tais como coisas e suas partes, materiais, atividades ou processos, eventos ou ocorrências, propriedades ou estados de coisas, pessoas, materiais ou ações, disciplinas ou assuntos, ou, ainda, nomes próprios ou de entidades [3]. A seleção de termos e descritores deve levar em conta sua relevância e seu uso tanto pela literatura da área quanto

por seus usuários. Termos e descritores devem ser selecionados, inicialmente, a partir da literatura e de publicações que são referência na área (*garantia literária*). Adicionalmente, a linguagem utilizada pelos usuários do tesauro deve ser também levada em conta (*garantia do usuário*).

Outro aspecto a ser considerado é com relação à forma gramatical. Os descritores devem ser substantivos ou frases substantivadas que representem um só conceito. Verbos e advérbios não devem ser considerados como descritores. Adjetivos podem ser utilizados em casos especiais, com o objetivo de reduzir o número de descritores compostos ou quando fazem parte do conceito. Além disso, a forma singular e plural deverá ser também observada. Objetos ou conceitos contáveis devem estar na forma plural, a não ser que sejam utilizados pela literatura ou pelos usuários na forma singular. Conceitos relativos a materiais e substâncias incontáveis, conceitos abstratos e nomes de entidades devem estar na forma singular. As abreviaturas e sua forma escrita por extenso devem ser consideradas como sinônimos. A forma escrita por extenso deve ser preferida, criando-se uma referência cruzada entre as duas formas de representar o mesmo conceito [3].

Após a seleção dos conceitos, os relacionamentos existentes entre eles devem ser claramente especificados através de indicadores de relacionamento. Idealmente, cada conceito deve estar relacionado a, pelo menos, um outro conceito do tesauro. Cada relacionamento possui a propriedade de reciprocidade, ou seja, todo relacionamento de A para B deve implicar em um relacionamento correspondente de B para A. Os relacionamentos mais comumente representados nos tesauros são os de equivalência, de hierarquia e de associatividade.

Nos relacionamentos de equivalência, os conceitos são sinônimos ou sinônimos parciais. Esse relacionamento indica que um conceito pode ser expresso por dois ou mais termos. Nesse caso, o mais utilizado pela comunidade de usuários deve ser definido como preferencial e os outros como não-preferenciais, quando utilizados pelos indexadores. Os indicadores de relacionamento utilizados são USE e UP. USE precede o termo preferencial (descritor) e UP, o termo não-preferencial (não-descritor).

Nos relacionamentos de hierarquia, os conceitos possuem relação de superordenação ou de subordinação, como os termos genéricos, com conotação mais ampla, e os termos específicos, com conotação mais específica. Os relacionamentos hierárquicos cobrem três tipos de situação: os relacionamentos genéricos que identificam classes e seus membros ou espécies; os relacionamentos que indicam todo-parte em situação em que um conceito é parte do outro; os relacionamentos de instância que indicam associações entre categorias de coisas ou eventos. Os indicadores de relacionamento hierárquico são TG, para termos genéricos, e TE, para termos específicos.

Nos relacionamentos de associatividade, indicador de relacionamento TR, os conceitos possuem um relacionamento simétrico, diferente dos anteriores. São termos tais que um pode ser utilizado para explicar o outro, de tal forma que um pode ajudar uma pessoa a se lembrar do outro. Os termos relacionados são termos semântica ou conceitualmente associados, mas não equivalentes, utilizado no mesmo contexto, conforme determinado pelos especialistas. Enquadram-se nesse tipo de relacionamento, conceitos que definem disciplinas ou campos de estudo e objetos ou fenômenos estudados; operações ou processos e seus agentes e instrumentos; ações e seus produtos ou objetos; objetos ou substâncias e suas propriedades; conceitos ligados por dependência causal; conceitos e suas unidades ou mecanismos de medida [29, 3]. Exemplos de conceitos e relacionamentos podem ser encontrados no Apêndice D.

Os padrões internacionais têm como objetivo definir princípios e regras a serem seguidos na construção manual de um tesauro, por especialistas de uma área de conhecimento. A qualidade de um tesauro depende, em parte, da observação e prática de tais princípios. No entanto, é a sua utilização por indexadores, no processo de indexação, e por usuários, no processo de pesquisa, que determinará a verdadeira qualidade de um tesauro.

3.3 Avaliação da Qualidade de um Tesauro

A qualidade de um tesauro pode ser verificada pela observância de padrões e normas internacionais vigentes durante seu processo de construção, conforme descrito anteriormente. Além desses padrões, algumas medidas estatísticas foram sugeridas por Lancaster [28] para se verificar a qualidade de um tesauro, medidas que descrevemos a seguir.

Medidas Estatísticas

Connectedness Ratio (Cr): é a razão de referência cruzada entre conceitos (i.e., conceitos associados a pelo menos um outro conceito) em relação ao total de conceitos do tesauro. Um valor próximo de 1 é recomendado. Esse valor pode ser calculado pela equação: $Cr = \frac{tot-at-least-one}{nc}$, onde *tot-at-least-one* é o total de conceitos que fazem pelo menos uma referência a outro conceito e nc é o número de conceitos do tesauro.

Accessibility (Acc): é o número médio de referências recebidas por conceitos do tesauro. Um valor entre 2 e 5 é recomendado. Esse valor pode ser calculado pela equação: $Acc = \frac{nr}{nc}$, onde nr é o número total de referências recebidas por conceitos

e *nc* é o número de conceitos do tesauro.

Equivalence ratio (Er): é a razão entre conceitos descritores e não-descritores no vocabulário do tesauro. Tal valor deve exceder 1 (mais termos que conceitos) e pode ser calculado pela equação: $Er = \frac{nc}{(nc-nondesc)}$, onde *nc* é o número de conceitos do tesauro e *nondesc* é o número de conceitos que são não-descritores.

Reciprocity ratio (Rr): é a extensão em que os conceitos classificados pelos relacionamentos (TR, TE, TG, USE, UP) são recíprocos, ou seja, quando existem as relações recíprocas TR-TR, TE-TG, TG-TE, USE-UP, UP-USE definidas no tesauro. Não há recomendação de valor para essa medida. Esse valor pode ser calculado pela equação: $Rr(Rel1\text{-}Rel2) = \frac{R1}{R2R1}$, onde $(Rel1\text{-}Rel2) \in \{$TR-TR, TE-TG, TG-TE, USE-UP, UP-USE$\}$, *R1* é o número de vezes que um conceito é classificado como *Rel1*, *R2R1* é o número de vezes que um conceito classificado como *Rel2* referencia outro classificado como *Rel1* e $(Rel1,\ Rel2) \in \{TR, TE, TG, USE, UP\}$.

Flexibility (F): é a proporção de palavras em conceitos multi-palavras (conceitos compostos por mais de uma palavra) que são conceitos do tesauro. É recomendado que esse valor seja 0,6 ou mais e pode ser calculado pela equação: $F = \frac{totpehcc}{nc}$, onde *totpehcc* é o número total de palavras em conceitos multi-palavras que é conceito e *nc* é o número de conceitos do tesauro.

Level of Pre-Coordination(PC): é a média do número de palavras por conceito. Tal valor deve estar entre 1,5 e 2,0 para inglês e francês. Esse valor pode ser calculado pela equação: $PC = \frac{TotP}{nc}$, onde *TotP* é o número total de palavras e *nc* é o número de conceitos do tesauro.

Apesar dessas medidas serem indicativas da qualidade de um tesauro, Lancaster [28] enfatiza que a sua qualidade só pode ser verificada pela utilização por seus usuários, ou seja, por indexadores e pesquisadores.

3.4 O Tesauro Jurídico da Justiça Federal

O Conselho de Justiça Federal — CJF [11] — e a Justiça Federal têm trabalhado na especificação, construção e manutenção de um tesauro para o Direito Brasileiro, o Tesauro Jurídico da Justiça Federal [57]. O objetivo do Tesauro da Justiça Federal é definir conceitos das áreas do direito brasileiro e seus respectivos relacionamentos. A motivação foi a construção de uma ferramenta para padronizar a linguagem utilizada pelos indexadores na construção da Seção de Indexação das jurisprudências, visando melhorar a recuperação de informação contida nas jurisprudências. O Tesauro da Justiça Federal foi elaborado para atender prioritariamente à Justiça Federal de segunda instância, ou seja, aos TRF's [60, 61, 62, 63, 64]. Ver Figura A.1

no Apêndice A.

O processo de construção, que ocorreu entre 1993 e 1997, contou com a coordenação do CJF e com a participação de representantes dos cinco Tribunais Regionais Federais que são componentes da Justiça Federal[1]. Os ramos do direito abordados pelo Tesauro da Justiça Federal são: Direito Penal, Processo Penal, Direito Civil, Processo Civil, Direito Comercial, Direito Previdenciário, Direito Tributário, Direito Administrativo, Direito Constitucional, Direito do Consumidor, Direito Eleitoral, Direito Internacional Público, Direito Internacional Privado, Direito Econômico-Financeiro e Direito Autoral.

O Tesauro da Justiça Federal contém 8.357 conceitos jurídicos apresentados em ordem alfabética. Desses, 6.103 são classificados como termos específicos, 891 como termos genéricos, 7.301 como termos relacionados e 1.702 são classificados como sinônimos, dos quais 674 são termos preferenciais (descritores) e 1.028 são termos não-preferenciais (não-descritores). Observamos que um conceito A, classificado como termo específico de um conceito B, pode ser classificado como termo genérico de outro conceito C e como termo relacionado de um quarto conceito D. Os operadores de relação definidos no Tesauro Jurídico da Justiça Federal são: UP, USE, TE, TG, TR. Os operadores UP e USE são utilizados para definir relacionamentos de equivalência, TE e TG para definir relacionamentos de hierarquia, TR para definir relacionamentos de associatividade. Ver exemplos de conceitos e relacionamentos no Apêndice D.

O Tesauro Jurídico da Justiça Federal foi construído por profissionais da Ciência da Informação e de Direito pertencentes aos quadros funcionais do CJF e dos TRF's. A estrutura do tesauro foi definida por profissionais da Ciência da Informação e a definição dos conceitos e seus relacionamentos foi de responsabilidade dos profissionais de Direito dos vários órgãos da Justiça Federal.

O processo de construção do Tesauro da Justiça Federal teve uma abordagem *bottom up* [28], que consistiu no levantamento dos conceitos considerados significantes e úteis para a indexação e para a pesquisa. Esse processo de levantamento é denominado *garantia literária* (*literary warrant* [28]). A outra forma de obtenção de conceitos, denominada *garantia do usuário*, que consiste em uma análise do vocabulário utilizado pelos usuários potenciais do sistema de informação, não foi utilizada na construção do Tesauro da Justiça Federal.

Os ramos do Direito a serem abordados pelo tesauro foram atribuídos aos gru-

[1]Informações obtidas em entrevista informal concedida por Neide de Sordi, coordenadora do processo de construção do tesauro, e Mônica Lacerda de Medeiros, ambas da Secretaria de Pesquisa e Informação Jurídicas do CJF, participantes do processo de construção e manutenção do Tesauro Jurídico da Justiça Federal.

pos participantes. No primeiro momento, eles leram os Códigos[2] e/ou as Doutrinas[3] existentes em cada ramo do direito. A partir dessa leitura, os participantes selecionaram as categorias e subcategorias a serem abordadas pelo tesauro. A seguir, os conceitos foram extraídos e seus relacionamentos foram propostos. Dessa forma, cada grupo, que trabalhou em separado, formou mini-tesauros para os ramos do Direito atribuídos a ele.

No segundo momento, os grupos se reuniram para discutir a união dos vários mini-tesauros em um único tesauro, fazendo a consolidação dos conceitos e relacionamentos que compõem o Tesauro Jurídico da Justiça Federal.

Como um tesauro é uma ferramenta que deve representar de forma atualizada os conceitos principais de uma área do conhecimento, ele deve ser mantido e revisado periodicamente. O processo de manutenção do Tesauro da Justiça Federal se dá através de proposições originadas nos grupos responsáveis nos órgãos da Justiça Federal, proposições essas que são discutidas entre os participantes. A alteração de conceitos existentes ou a inclusão de novos conceitos se dá por consenso.

[2]Código é uma coleção de Leis destinadas a reger a matéria que faz parte ou que é objeto de um ramo do Direito [50].

[3]Doutrina é um conjunto de princípios que firmam teorias ou interpretações sobre a ciência jurídica. Pode ainda representar a opinião particular de um ou de vários jurisconsultos a respeito de um ponto de direito controvertido [50].

3.5 Avaliação da Qualidade do Tesauro da Justiça Federal

Conforme descrito na Seção 3.3, a qualidade de um tesauro pode ser avaliada pela satisfação dos usuários mediante seu uso. No entanto, tal qualidade pode também ser determinada de duas formas alternativas. Primeiro, pela observância de normas internacionais que regulamentam o processo de construção de tesauros. Segundo, pelo uso de algumas medidas estatísticas adotadas como referência na avaliação da qualidade de um tesauro. Para nos certificarmos que o Tesauro da Justiça Federal apresentava boa qualidade, realizamos as avaliações apresentadas abaixo.

Com relação à observação de padrões internacionais [3, 24], o Tesauro da Justiça Federal apresenta boa qualidade. Os conceitos foram obtidos a partir da literatura referencial da área (garantia literária), obedecem às regras da forma gramatical, ou seja, os conceitos são substantivos ou frases substantivadas e aparecem na forma singular ou plural conforme seu uso. Verbos e adjetivos não são utilizados como conceitos.

Outro aspecto diz respeito à clareza na especificação dos relacionamentos, através dos indicadores de relacionamento USE, UP, TG, TE e TR. Cada conceito é precedido por um número de controle interno. A seguir, os relacionamentos são definidos, com cada conceito relacionado sendo precedido pelo indicador de relacionamento equivalente, sempre na mesma ordem, observando-se que nem todos os conceitos possuem todos os relacionamentos. Ver exemplos no Apêndice D.

Com relação à avaliação de qualidade baseada em medidas estatísticas, definidas na Seção 3.3, o Tesauro da Justiça Federal também apresenta boa qualidade, pois os resultados se enquadram dentro dos valores recomendados pela literatura. A medida *Connectedness Ratio (Cr)* é igual a 1, quando o valor recomendado é que seja aproximadamente igual a 1. A medida *Accessibility (Acc)* é igual 5,27, quando o valor recomendado é que deve estar entre 2 e 5. A medida *Equivalence ratio (Er)* é igual a 1,14, quando o valor recomendado é que deve exceder 1. A medida *Reciprocity ratio (Rr)*, calculada para todos os relacionamentos recíprocos, ou seja, TE/TG e TG/TE, USE/UP e UP/USE e TR/TR, é igual a 1. A medida *Flexibility (F)* é igual a 0,9, quando o valor recomendado é que seja maior que 0,6. A medida *Level of Pre-coordination (PC)* é igual a 2,63, quando o valor recomendado é entre 1,5 e 2 para inglês e francês. Valor referência para o Português não foi encontrado.

Com base nesse levantamento estatístico e na observância de padrões internacionais na sua construção, podemos inferir que o Tesauro da Justiça Federal é um tesauro bem construído e, portanto, confiável.

Capítulo 4

As Redes Bayesianas e sua Utilização em RI

Neste Capítulo introduzimos os conceitos sobre redes Bayesianas, sobre o modelo de rede de crenças utilizado em RI. Apresentamos também sua utilização na área de RI.

4.1 Redes Bayesianas

As redes Bayesianas fornecem um formalismo gráfico para representar as variáveis de uma distribuição conjunta de probabilidades e para representar as dependências, e, conseqüentemente as independências, conhecidas entre as variáveis. Elas fornecem uma descrição completa do domínio e podem ser utilizadas para responder a qualquer pergunta sobre o domínio. A distribuição conjunta de probabilidade é modelada por um grafo direcionado acíclico, no qual os nodos representam as variáveis aleatórias da distribuição e os arcos direcionados representam os relacionamentos causais existentes entre elas.

A topologia da rede é uma representação abstrata do conhecimento do domínio, ou seja, a estrutura causal entre os processos do domínio. Uma vez que a topologia da rede está definida, é necessário especificar as probabilidades condicionais para os nodos que participam diretamente das relações de dependência. Cada nodo possui uma tabela de probabilidade condicional que quantifica a influência que os nodos pais têm sobre cada nodo filho.

O princípio fundamental é que as dependências conhecidas entre as variáveis aleatórias do domínio são declaradas explicitamente na rede e que a distribuição conjunta de probabilidade pode ser inferida a partir dessas dependências. Os relacionamentos entre os nodos, indicados pelos arcos direcionados do grafo, representam

dependências causais ou as influências diretas entre as variáveis do domínio. A intensidade dessas influências ou dependências é expressa por probabilidades condicionais associadas aos arcos do grafo. As dependências declaradas são utilizadas para inferir as crenças (probabilidades) associadas a todas as variáveis da rede.

Como ilustração, vejamos o exemplo da Figura 4.1. Nessa figura, $X_1, X_2, \cdots, X_i, \cdots, X_n$ e Y são variáveis aleatórias representadas como nodos da rede Bayesiana. Se o valor de uma variável X_1 tem influência direta no valor de outra variável Y, um arco direcionado de X_1 para Y é inserido no grafo (X_1 é dito pai de Y). Sejam (x_1, y) dois possíveis valores para as variáveis X_1 e Y, respectivamente. A influência de X_1 sobre Y é expressa pela probabilidade condicional $P(y|x_1)$.

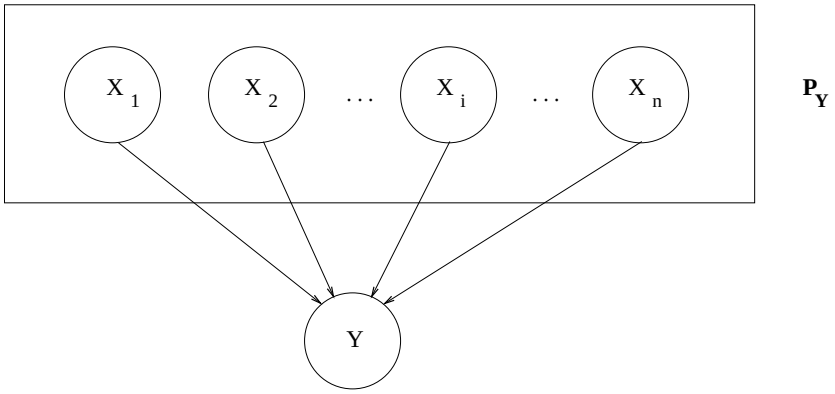

Figura 4.1: Exemplo de Rede Bayesiana

Seja $\mathbf{P_Y} = \{X_1, X_2, \cdots, X_i, \cdots, X_n\}$ o conjunto de variáveis pais de Y e seja $p_y = \{x_1, x_2, \cdots, x_i, \cdots, x_n\}$ um conjunto de valores possíveis de $\mathbf{P_Y}$. A influência de $\mathbf{P_Y}$ sobre Y pode ser especificada por qualquer função F, que satisfaça:

$$\sum_{\forall y \in D_Y} F(y, p_y) = 1 \quad e \quad 0 \leq F(y, p_y) \leq 1,$$

onde D_Y é o domínio da variável Y. A função $F(y, p_y)$ fornece a quantificação da probabilidade condicional $P(y|p_y)$. Essa especificação representa a distribuição conjunta de probabilidade para os nodos da rede da Figura 4.1 (mais informações em Pearl [37]).

Para exemplificar, consideremos a Figura 4.2. Essa figura representa a distribuição conjunta de probabilidade $P(x_1, x_2, x_3, x_4, x_5)$ para as variáveis aleatórias $\{X_1, X_2, X_3, X_4, X_5\}$, onde $\{x_1, x_2, x_3, x_4, x_5\}$ são seus respectivos valores. O nodo X_1 é denominado raiz da rede (nodo sem pai) e é o pai de X_2 e X_3. A probabilidade

$P(x_1)$, associada com o valor x_1 do nodo raiz X_1, é denominada probabilidade *a priori* e é utilizada para representar o conhecimento prévio sobre o domínio modelado. Dado o valor da variável X_1, as variáveis X_2 e X_3 são independentes. Dados os valores das variáveis X_2 e X_3, as variáveis X_4 e X_5 são independentes. Devido às dependências declaradas na rede, a distribuição conjunta de probabilidade pode ser calculada por:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\ P(x_2|x_1)\ P(x_3|x_1)\ P(x_4|x_2, x_3)\ P(x_5|x_3).$$

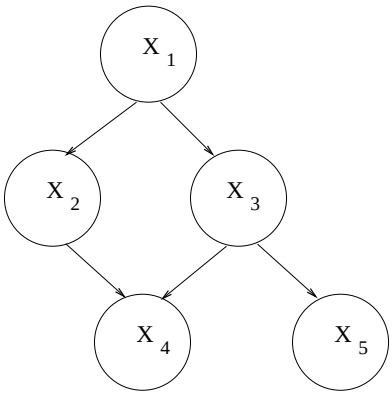

Figura 4.2: Rede Bayesiana com independências declaradas

A principal vantagem das redes Bayesianas é o poder de síntese de representação dos relacionamentos probabilísticos. Em vez de considerar a distribuição conjunta de probabilidades completa, elas permitem considerar apenas as independências entre as variáveis declaradas na rede, independências essas que são utilizadas para inferir as crenças de todas as variáveis. Outra grande vantagem das redes Bayesianas é que elas podem ser naturalmente estendidas por evidências geradas a partir de fontes independentes de conhecimento, tais como os tesauros, conforme será discutido no Capítulo 5.

4.2 Modelo de Rede de Crenças

O modelo de rede de crenças para RI, proposto por Ribeiro-Neto e Muntz [40], segue a linha epistemológica, na qual as probabilidades são interpretadas como graus de crença. Daí o seu nome. A rede Bayesiana proposta é um grafo direcionado acíclico, no qual as probabilidades são interpretadas como graus de crença, sem

necessidade de experimentação. Os nodos da rede representam variáveis aleatórias e os arcos representam o relacionamento entre as variáveis, cuja intensidade é expressa pela probabilidade condicional da relação.

Uma simplificação comumente aceita nos modelos clássicos de RI, também assumida por Ribeiro-Neto e Muntz, é que documentos e consultas são representados por palavras-chave e que essas são independentes entre si. Como resultado, consultas e documentos da coleção são modelados como subconjuntos de palavras-chave.

Uma consulta Q é modelada como um nodo da rede e é representada por uma variável binária aleatória Q[1]. Essa variável pode estar ativa (indicado por q) ou inativa (indicado por $\bar{q}$)[2]. Cada documento da coleção é modelado como um nodo da rede e é representado por uma variável binária aleatória D_j. Essa variável pode estar ativa (indicado por d_j) ou inativa (indicado por $\bar{d_j}$). Similarmente, cada termo de indexação é modelado como um nodo da rede. A cada termo K_i está associada a variável binária aleatória K_i. Essa variável pode estar ativa (indicado por k_i) ou inativa (indicado por $\bar{k_i}$). O fato de uma variável binária estar ativa representa que o evento associado, consulta, documento ou termo, é observado.

O espaço conceitual no qual as variáveis aleatórias se inserem é denominado universo de discurso. Nesse trabalho, assumimos que o universo de discurso U é o vocabulário da coleção que é composto pelos termos de indexação advindos dos documentos da coleção e pelos conceitos advindos de um tesauro relacionado, ou seja, $U = (K_1, K_2, \cdots, K_t)$, onde t é o número de termos únicos do vocabulário (enriquecido com os conceitos do tesauro).

Nesse espaço, palavras-chave ou sub-conjuntos de palavras-chave são interpretadas como conceitos em U. Consultas e documentos são modelados como conceitos em U, ou seja, $D = (K_1, K_2, \cdots, K_j, \cdots, K_t)$ e $Q = (K_1, K_2, \cdots, K_i, \cdots, K_t)$, onde $D \subseteq U$ e $Q \subseteq U$. Como resultado, consultas e documentos são tratados de modo análogo, o que leva à rede Bayesiana da Figura 4.3.

Na rede da Figura 4.3, o nodo Q modela a consulta Q, os nodos K_i modelam as palavras-chave do vocabulário da coleção e os nodos D_j, os documentos da coleção. A cada nodo está associada uma variável binária aleatória para representar se o nodo está ativo ou inativo. Seja $\mathbf{K}$ o conjunto de todas as variáveis K_i. O nodo Q é apontado pelos nodos K_i que representam as palavras-chave presentes em Q.

[1] Nossa notação adota o mesmo rótulo para o nome do nodo e para o nome da variável aleatória. Isso não gera ambigüidade, pois estará sempre claro no texto quando nos referimos ao nodo ou à variável aleatória associada.

[2] Em nossa notação, letras maiúsculas designam variáveis aleatórias (ou nodos da rede), letras minúsculas designam o estado dessas variáveis e letras maiúsculas em negrito designam conjuntos de variáveis.

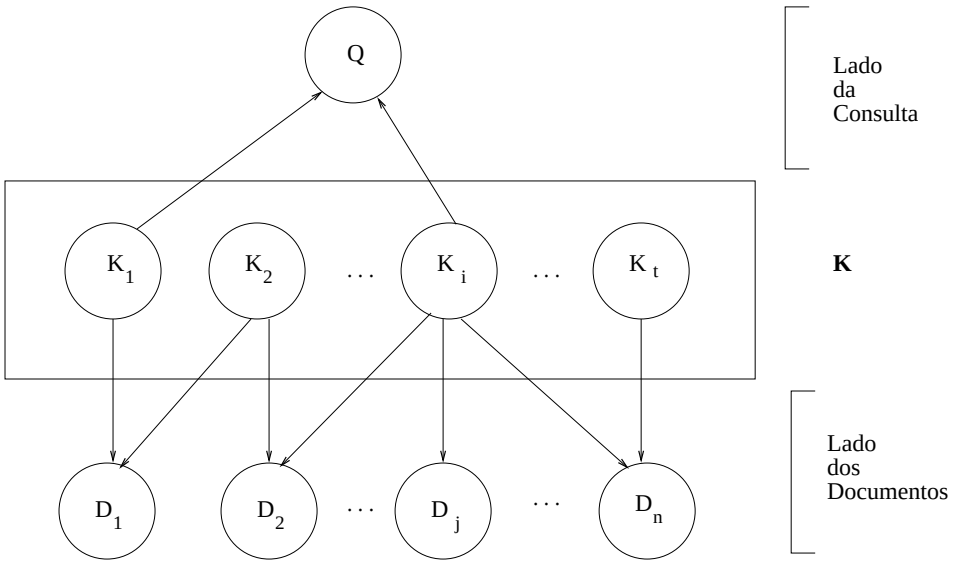

Figura 4.3: Rede Bayesiana para Q com termos K_1 e K_i (de [51])

Os nodos D_j são apontados pelas palavras-chave K_i que estão presentes em D_j. Por exemplo, na Figura 4.3, $Q = (K_1 = 1, K_2 = 0, \cdots, K_i = 0, \cdots, K_t = 0)$ e $D_1 = (K_1 = 1, K_2 = 1, \cdots, K_j = 0, \cdots, K_t = 0)$.

Na rede da Figura 4.3, a ordenação dos documentos é baseada na quantificação da similaridade entre documento e consulta dada pela probabilidade $P(d_j|q)$. Pelo Teorema de Bayes, a probabilidade $P(d_j|q)$ pode ser calculada como segue:

$$P(d_j|q) = \frac{P(d_j \wedge q)}{P(q)} \tag{4.1}$$

Como $P(q)$ não influencia a ordenação, podemos substituir $\frac{1}{P(q)}$ por um fator de normalização, ou seja, $\frac{1}{P(q)} = \eta$. Como $P(d_j \wedge q)$ é dependente do estado das variáveis em $\mathbf{K}$, podemos reescrever a Equação (4.1), como segue:

$$P(d_j|q) = \eta \sum_{\forall k} P(d_j \wedge q \mid k)\ P(k) \tag{4.2}$$

onde k é um estado das variáveis em $\mathbf{K}$. Pela rede da Figura 4.3, D_j e Q são independentes dado k. Logo:

$$P(d_j \wedge q \mid k) = P(d_j \mid k)\ P(q \mid k) \tag{4.3}$$

Substituindo (4.3) em (4.2), temos:

$$P(d_j \mid q) = \eta \sum_{\forall k} P(d_j \mid k)\ P(q \mid k)\ P(k) \tag{4.4}$$

A Equação (4.4) é a equação geral para estimar a probabilidade que um documento D_j seja relevante para a consulta Q. Observamos que essa equação pode ser adaptada para representar outros modelos clássicos de RI, como o booleano e o probabilístico, bastando para isso fazer definições apropriadas para as probabilidades que a compõem. A seguir mostramos como adaptar essa equação para representar o modelo vetorial.

4.2.1 Rede de Crenças para o Modelo Vetorial

O modelo vetorial, que é muito popular na área de Recuperação de Informação (RI), foi introduzido na Seção 2.3. A Equação (2.2) define a similaridade $sim(\vec{d_j}, \vec{q})$ entre um documento D_j e uma consulta Q, segundo o modelo vetorial. No modelo de rede de crenças, a similaridade entre a consulta Q e um documento D_j é dada pela probabilidade $P(d_j|q)$, conforme Equação (4.4). A probabilidade $P(d_j|q)$ pode ser feita equivalente a $sim(\vec{d_j}, \vec{q})$, dada pelo modelo vetorial (veja Equação 2.2), através das seguintes especificações:

$$P(k) \quad = \quad (1/2)^t \tag{4.5}$$

onde t é o número de termos do vocabulário da coleção. Essa formulação implica que todas as combinações de termos são, *a priori*, igualmente prováveis, tanto na constituição de consultas como na de documentos.

Para $P(q|k)$ definimos:

$$P(q|k) = \begin{cases} 1 & \text{se } \forall\ i,\ g_i(q) = g_i(k) \\ 0 & \text{caso contrário} \end{cases} \tag{4.6}$$

onde $g_i(q)$ retorna o estado da variável k_i segundo q e $g_i(k)$ retorna o estado da variável k_i segundo k. A equação (4.6) estabelece que apenas as palavras-chave presentes em Q são de interesse. Ou seja, o nodo Q é utilizado somente para selecionar os termos a serem ativados. E para $P(d_j|k)$ definimos:

$$P(d_j|k) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{ik}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{ik}^2}} \tag{4.7}$$

onde w_{ij} e w_{ik} são os pesos $td - idf$ definidos para o modelo vetorial. Substituindo (4.5), (4.6) e (4.7) em (4.4) obtemos $P(d_j|q)$ que é coincidente com a similaridade fornecida pelo modelo vetorial.

4.3 Uso de Redes Bayesianas em RI

Os primeiros autores a utilizar redes Bayesianas em sistemas de RI foram Turtle e Croft [65], quando propuseram o modelo de rede de inferência para representar termos de indexação, consultas e documentos como nodos da rede. A rede é utilizada para representar como a necessidade de informação é atendida por documentos de uma coleção, quando mediados pelas palavras-chave que os representam.

A rede de inferência é composta pela rede de documentos e pela rede da consulta. A rede dos documentos é construída uma única vez para a coleção, enquanto a rede da consulta é construída quando uma consulta é observada. Numa rede de inferência, a observação de um documento é causa para aumentar a crença nas variáveis randômicas associadas às palavras-chave do documento. Esse relacionamento causal é modelado por um arco direcionado do documento para as palavras-chave. A crença na consulta é uma função das crenças associadas aos nodos das palavras-chave que a compõem, o que implica em inserir arcos direcionados das palavras-chave para a consulta. Dada uma consulta, cada documento d_j é observado e a crença no mesmo em relação à consulta pode ser calculada. Esse procedimento é repetido para todos os documentos da coleção, gerando uma probabilidade de que cada documento da coleção satisfaça a necessidade de informação expressa na consulta do usuário. A ordenação do documento d_j em relação à consulta q é dada pela probabilidade $P(q \mid d_j)$.

Ribeiro-Neto e Muntz [40] propõem que a consulta seja uma evidência da necessidade de informação e que a ordenação dos documentos na resposta seja calculada pela probabilidade de um documento dada uma consulta, ou seja, $P(d_j \mid q)$. Esses autores utilizaram redes Bayesianas para combinar evidências de consultas passadas com evidências obtidas do conteúdo dos documentos da coleção para melhorar a qualidade das respostas. Esses autores mostraram que o formalismo das redes Bayesianas fornece uma poderosa base conceitual para representar, quantificar e combinar múltiplas fontes de evidência em um sistema de RI, com o objetivo de melhorar a qualidade da ordenação.

Em Ribeiro-Neto et al. [41] novos experimentos com evidências de consultas passadas são executados. Utilizando quatro coleções de referência distintas, esses autores reportam melhoria em precisão, da ordem de 59% para a coleção CFC, 86% para a coleção CISI, 93% para a coleção CACM e de 123% para a coleção TREC-8.

Silva et al. [52] utilizaram rede de crenças para combinar evidências obtidas da estrutura de *links* de páginas da *Web* com evidências advindas de palavras-chave. A informação contida na estrutura de *links* é composta por evidência *hub*, que são páginas que apontam para muitas páginas, e evidência *authority*, que são páginas

apontadas por muitas outras. Esses autores reportam experimentos combinando evidências de palavras-chave, *hub* e *authority*, com ganhos em precisão de até 74%.

Calado et al. [7] utilizaram rede de crenças para combinar evidências advindas da estrutura de *links* de páginas *Web*. Assim como em Silva et al. [52], esses autores utilizaram evidências *hub* e *authority* quando calculadas em relação à coleção local e, adicionalmente, em relação à coleção global. A informação local é obtida através da análise de *links* dos documentos do conjunto resposta retornado à consulta, enquanto a informação global utiliza análise de *links* dos documentos da coleção completa. Esses autores relataram ganhos de 74% em precisão média quando combinaram evidências de palavras-chave, *hub* e *authority* para a coleção local e de 35% em precisão média para a coleção global. O resultado da coleção global supera o da coleção local para os pontos mais baixos de revocação.

Haines e Croft [22] utilizaram rede de inferência para fazer realimentação de relevantes. A partir da informação de relevância recebida do usuário, *links* para os novos termos são inseridos na rede da consulta, o que implica na alteração das dependências entre os nodos da rede. Esse fato implica que novos valores de probabilidades são calculados, gerando novas crenças em documentos e, conseqüentemente, um novo conjunto resposta. Esses autores relataram ganhos em precisão de 90% para uma coleção de documentos com títulos e resumos e de 30% para uma coleção de documentos com texto completo.

No tocante à utilização de tesauros jurídicos, resultados preliminares apresentados por Ribeiro-Neto e Assumpção [4, 39] utilizando o Tesauro Jurídico da Justiça Federal e redes Bayesianas, sugeriram que expansão de consultas baseada em relacionamentos do tesauro melhora o resultado da busca. Esses autores concluíram que a utilização de sinônimos e termos específicos levam a melhores resultados.

Capítulo 5

O Modelo de Rede de Crenças utilizando Tesauros

Neste capítulo apresentamos o modelo proposto com o objetivo de melhorar a precisão dos resultados gerados para uma consulta de um usuário utilizando o arcabouço probabilístico das redes Bayesianas. As redes Bayesianas são utilizadas como arcabouço teórico para combinar os resultados obtidos por palavras-chave com os resultados obtidos utilizando informações contidas em um tesauro.

Uma justificativa para a utilização do conceitual teórico das redes Bayesianas é que elas permitem modelar a incerteza que, nesse caso, está na formulação da consulta do usuário. Através do tesauro, geramos novas consultas relacionadas às consultas originais, isto é, inferimos o que o usuário poderia estar querendo dizer. Além disso, elas fornecem uma base formal flexível, permitindo incluir novas fontes de evidência sem perder a consistência.

5.1 Utilização de um Tesauro como Evidência Semântica

A idéia é que os conceitos de um tesauro e seus relacionamentos sejam utilizados como evidência semântica para melhor contextualizar a consulta e, conseqüentemente, para melhorar a qualidade dos resultados. Para isso, adicionamos novos nodos, arestas e probabilidades à rede original mostrada na Figura 4.3 da Seção 4.2. Esses novos nodos e arestas são utilizados para representar conceitos de um tesauro relacionados com a consulta original do usuário. A Figura 5.2 ilustra a rede resultante.

Os termos da consulta original são representados pelo conjunto $\mathbf{KY}$, como ante-

riormente. Do texto da consulta obtemos automaticamente os conceitos do tesauro jurídico relacionados a ela, os quais são representados pelo conjunto **CC**. A partir dos conceitos do tesauro, os diferentes relacionamentos, sinônimos, termos relacionados, termos específicos e termos genéricos, são obtidos também de forma automática. Tais mapeamentos são representados na rede pelos conjuntos de nodos **SY**, **TR**, **TE** e **TG**, respectivamente. Quando os documentos são recuperados utilizando somente informação proveniente do conjunto **KY**, dizemos que a pesquisa é baseada em palavras-chave. Quando os documentos são recuperados utilizando informação proveniente dos conjuntos **CC**, **SY**, **TR**, **TE** e **TG**, dizemos que a pesquisa é baseada em conceitos.

Os nodos $D_{Kj}, D_{Cj}, D_{Sj}, D_{Rj}, D_{Ej}$ e D_{Gj} representam o documento D_j em diferentes contextos. O nodo D_{Kj} é utilizado para representar o documento D_j quando ele aparece como resposta a uma pesquisa baseada em palavras-chave, ou seja, a consulta é considerada como composta por palavras-chave e é processada utilizando a fórmula do cosseno do modelo vetorial aplicada a palavras-chave. O nodo D_{Cj} é utilizado para representar o documento D_j quando ele aparece como resposta a uma pesquisa baseada em conceitos, ou seja, o documento deve conter o conceito do tesauro referenciado na consulta e é processada utilizando a fórmula do cosseno aplicada a conceitos. D_{Kj} e D_{Cj} são modelados separadamente na rede para avaliar o impacto dos conceitos do tesauro na qualidade dos resultados. Os nodos D_{Sj}, D_{Rj}, D_{Ej} e D_{Gj} são utilizados para representar o documento D_j quando ele aparece como resposta a uma pesquisa baseada em conceitos que são sinônimos, termos relacionados, termos específicos ou termos genéricos, respectivamente, dos conceitos presentes em **CC**.

Os procedimentos descritos acima são ilustrados com a consulta 25 do Apêndice C. Os procedimentos de geração dos conjuntos **KY**, **CC**, **SY**, **TR**, **TE** e **TG** são descritos a seguir e os conjuntos gerados podem ser vistos na Figura 5.1.

O texto da consulta 25 é: "Tutela antecipada em ação rescisória". Os mesmos procedimentos utilizados para extração dos representantes dos documentos no índice, descritos na Seção 6.2.1 e ilustrados na Figura 2.2, foram aplicados à consulta 25, ou seja, eliminação de *stopwords* e de acentos e tranformação em maiúsculas. Esses procedimentos geraram o conjunto **KY**. A seguir, o texto da consulta foi mapeado diretamente no Tesauro da Justiça Federal gerando o conjunto **CC**, ou seja, **CC** contém os conceitos do tesauro presentes no texto da consulta 25. Os componentes do conjunto **CC**, por sua vez, geraram os conjuntos **SY**, **TR**, **TE** e **TG**, que são os conceitos definidos nos relacionamentos dos conceitos em **CC**, conforme descrito no tesauro.

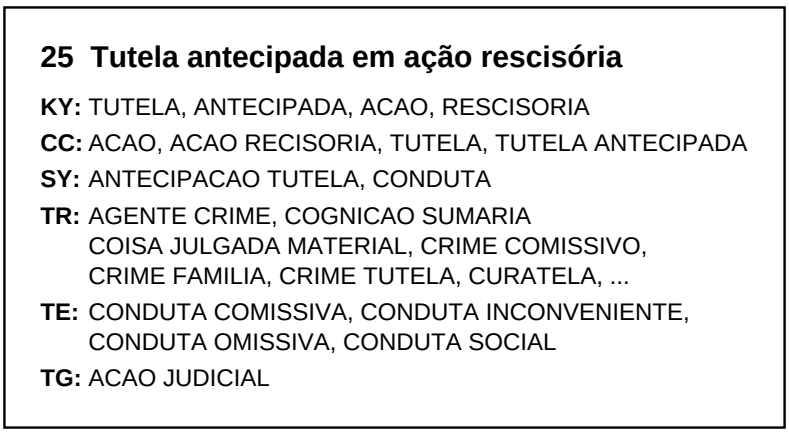

Figura 5.1: Exemplo de geração de consultas utilizando o tesauro

As consultas representadas pelos conjuntos **KY**, **CC**, **SY**, **TR**, **TE** e **TG** foram submetidas ao sistema gerando os conjuntos-resposta D_K, D_C, D_S, D_R, D_E e D_G, respectivamente. Esses conjuntos-resposta foram combinados de acordo com a Equação (5.6), deduzida a partir da rede ilustrada na Figura 5.2, gerando o conjunto D.

5.2 Equação Geral para o Cálculo da Similaridade entre um Documento na Resposta e a Consulta do Usuário

Nesse trabalho consideramos que a ordenação dos documentos do conjunto resposta é estabelecida pela probabilidade $P(d_j|q)$, que é calculada pela Equação (4.4), transcrita abaixo:

$$P(d_j \mid q) = \eta \sum_{\forall k} P(d_j \mid k)\ P(q \mid k)\ P(k) \tag{5.1}$$

Assumindo que os únicos termos de interesse são os termos mencionados na consulta, fazemos a seguinte redefinição:

$$P(q|k) = \begin{cases} 1 & \text{se } \mathbf{KY} = k_q \\ 0 & \text{caso contrário} \end{cases} \tag{5.2}$$

onde k_q é um estado das variáveis em **KY**, no qual os únicos nodos ativos são aqueles que correspondem a termos presentes na consulta Q. Isso permite que a

Equação (5.1) seja reescrita como segue:

$$P(d_j \mid q) = \eta \; P(d_j \mid k_q) \; P(q \mid k_q) \; P(k_q) \tag{5.3}$$

No entanto, $P(d_j|k_q)$ depende das evidências advindas dos diferentes conceitos e seus relacionamentos. Essas evidências são utilizadas para enriquecer a rede com distintas representações da consulta original. Para cada uma de tais representações são geradas as respectivas listas ordenadas de documentos, $D_{Cj}, D_{Sj}, D_{Rj}, D_{Ej}$ e D_{Gj}. Essas listas ordenadas são vistas como fontes distintas de evidência no cálculo final da relevância dos documentos para a consulta original. Para combinar essas fontes de evidência, nós utilizamos o operador disjuntivo (vide Figura 5.2) porque ele proporciona melhores resultados [40, 41, 52]. Assim, o nodo do documento D_j acumula os valores de todas as evidências através da disjunção das crenças associadas aos nodos $D_{Kj}, D_{Cj}, D_{Sj}, D_{Rj}, D_{Ej}$ e D_{Gj}. Isso permite reescrever a Equação (5.3) como segue:

$$\begin{aligned} P(d_j \mid q) = \eta \; [1 - P(\bar{dk}_j \mid k_q) \; \times P(\bar{dc}_j \mid k_q) \times P(\bar{de}_j \mid k_q) \times P(\bar{dg}_j \mid k_q) \times \\ P(\bar{ds}_j \mid k_q) \times P(\bar{dr}_j \mid k_q)] \times P(q \mid k_q) \times P(k_q) \end{aligned} \tag{5.4}$$

Avaliando cada uma das parcelas em separado, como, por exemplo, $P(\bar{dc}_j \mid k_q)$, obtemos:

$$\begin{aligned} P(\bar{dc}_j \mid k_q) &= \sum_{\forall cc} P(\bar{dc}_j \wedge cc \mid k_q) \\ &= \sum_{\forall cc} P(\bar{dc}_j \mid cc \wedge k_q) \times P(cc \mid k_q) \end{aligned}$$

Como cc torna $\bar{dc}_j$ independente de k_q, podemos escrever:

$$P(\bar{dc}_j \mid k_q) = \sum_{\forall cc} P(\bar{dc}_j \mid cc) \times P(cc \mid k_q) \tag{5.5}$$

onde cc é um estado das variáveis em $\mathbf{CC}$.

Assumindo que os únicos termos de interesse são aqueles com conceitos mapeados no tesauro, definimos:

$$P(cc|k_q) = \begin{cases} 1 & \text{se os únicos nodos ativos correspondem aos conceitos em Q} \\ 0 & \text{caso contrário} \end{cases}$$

Assim, se $P(cc|k_q) = 1$ representa o estado em que apenas os conceitos presentes na consulta estão ativos, chegamos à fórmula fechada:

$$\begin{aligned} P(\bar{dc}_j \mid k_q) &= P(\bar{dc}_j \mid cc) \\ &= (\; 1 - P(dc_j \mid cc) \;) \end{aligned}$$

onde cc é um estado das variáveis em **CC** no qual os únicos nodos ativos são aqueles correspondentes aos conceitos do tesauro mapeados na consulta.

Esse mesmo raciocínio vale para as contribuições das outras parcelas da Equação (5.4), que pode então ser reescrita como segue:

$$P(d_j \mid q) = \eta\,[1 - (1 - P(dk_j \mid k_q)) \times (1 - P(dc_j \mid cc)) \times (1 - P(de_j \mid ee)) \times (1 - P(dg_j \mid gg)) \times (1 - P(ds_j \mid ss)) \times (1 - P(dr_j \mid rr))] \times P(q \mid k_q) \times P(k_q) \quad (5.6)$$

onde k_q é um estado das variáveis em **KY** no qual os únicos nodos ativos são aqueles correspondentes aos termos da consulta. Similarmente, cc, ee, gg, ss, rr representam os estados das variáveis aleatórias que representam os nodos ativos em **CC, TE, TG, SY** e **TR**, respectivamente, nos quais os únicos nodos ativos são aqueles associados com os conceitos mapeados pelas palavras-chaves e seus respectivos relacionamentos.

A Equação (5.6) é a fórmula para ordenação dos documentos em nosso modelo Bayesiano para uma biblioteca digital que conta com um tesauro associado. Nossa fórmula de ordenação permite combinar evidências associadas à consulta do usuário de várias maneiras. Por exemplo, para considerar somente o efeito dos termos da consulta original, dado pelo modelo vetorial, basta definir:

$$P(dc_j \mid cc) = 0;\ P(de_j \mid ee) = 0;\ P(dg_j \mid gg) = 0;\ P(ds_j \mid ss) = 0;\ P(dr_j \mid rr) = 0.$$

Como resultado obtemos:

$$P(d_j|q) = \eta\ P(dk_j \mid k_q) \times P(q \mid k_q) \times P(k_q)$$

que ordena as respostas segundo o modelo vetorial.

Se quisermos considerar somente a evidência gerada por sinônimos dos conceitos associados aos termos da consulta, podemos definir:

$$P(dk_j \mid k_q) = 0;\ P(dc_j \mid cc) = 0;\ P(de_j \mid ee) = 0;\ P(dg_j \mid gg) = 0;\ P(dr_j \mid rr) = 0.$$

Adicionalmente, combinações de evidências duas a duas, três a três, e assim por diante, podem ser avaliadas definindo apropriadamente as probabilidades condicionais relacionadas. Notamos ainda que a Equação (5.6) permite calcular 63 diferentes formas de ordenação. Uma observação adicional é que, ao cancelarmos uma evidência fazendo sua probabilidade igual a zero, estamos eliminando parte da estrutura da rede, ou seja, estamos considerando uma rede diferente para cada combinação possível.

Custo Computacional

Apesar da distribuição de probabilidade conjunta ser exponencial em relação ao número de variáveis da distribuição, as independências declaradas na rede Bayesiana permitem um processamento mais simples.

No lado esquerdo da rede ilustrada na Figura 5.2, referente à rede original, somente um estado dos nodos raíz é considerado no cálculo de ordenação dos documentos, que é o estado em que apenas as palavras-chave da consulta são consideradas. Esse cálculo é feito de acordo com a fórmula do cosseno do modelo vetorial e é linear em relação ao número de documentos da coleção.

Com relação ao lado direito da rede ilustrada na Figura 5.2, os estados dos nodos considerados são aqueles que correspondem aos conceitos do tesauro associados à consulta. O custo adicionado por cada nova fonte de evidência considerada é equivalente ao custo do modelo vetorial. Como no modelo proposto temos ev evidências ($ev = 6$), o custo da equação para o cálculo da ordenação do conjunto resposta é $O(ev \times N)$[1], onde N é o número de documentos da coleção. Como $ev << N$, a complexidade é $O(N)$.

[1] A notação O significa o limite superior do custo de processamento de uma dada função [13].

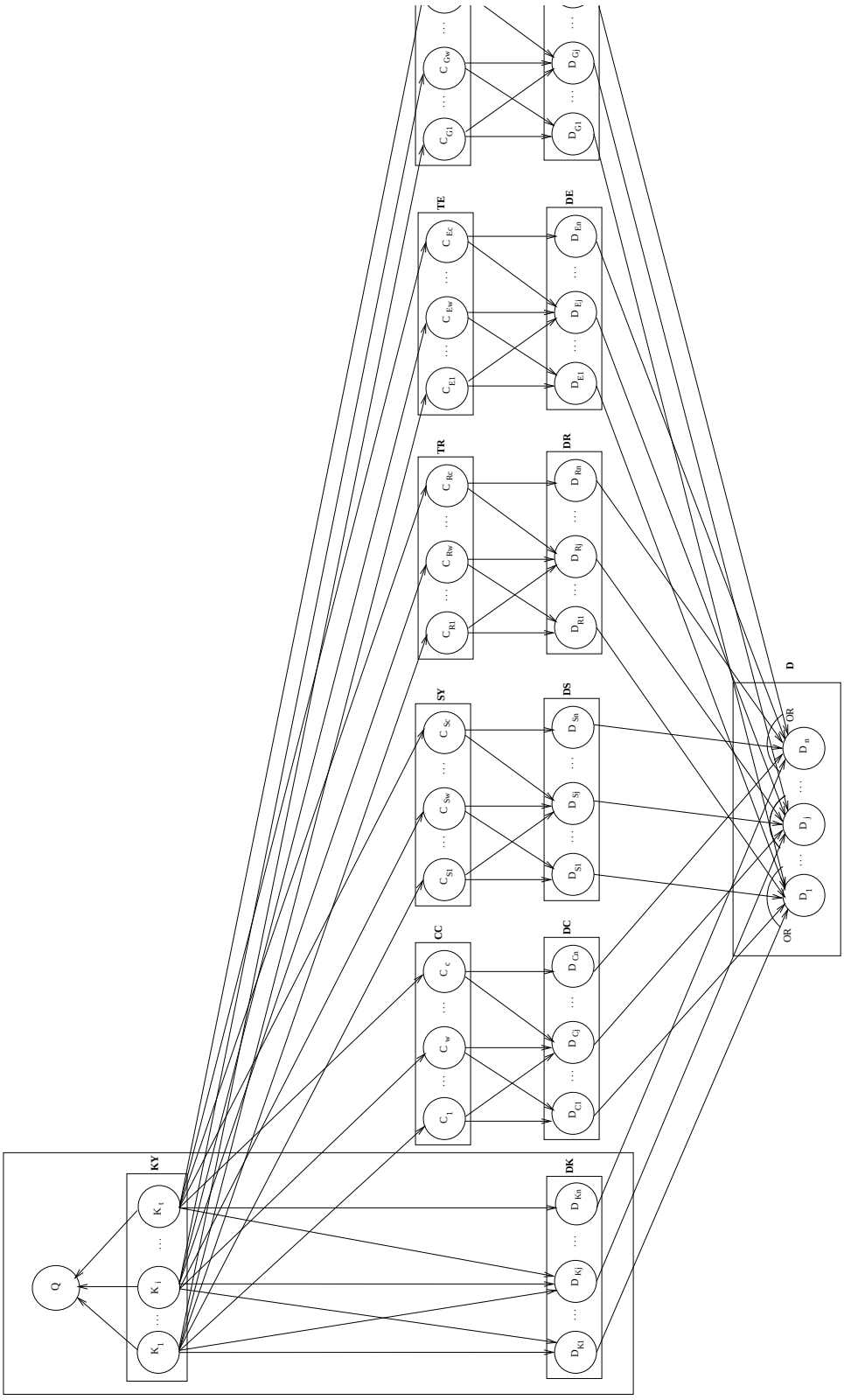

Figura 5.2: Rede de Crenças utilizando tesauro

Capítulo 6

Experimentos: Um Estudo de Caso na Área Jurídica

Neste Capítulo apresentamos os resultados experimentais utilizados para validar o modelo proposto. Nossa validação é baseada em um estudo de caso na área jurídica. Para fazer a validação do modelo proposto, uma coleção de referência da área jurídica brasileira foi composta, um tesauro jurídico brasileiro foi utilizado e um sistema de RI foi implementado. Descrevemos a seguir a constituição da coleção de referência, as características do sistema implementado e os resultados obtidos. O tesauro utilizado foi descrito nas Seções 3.4 e 3.5.

6.1 A Coleção de Referência

Uma coleção de referência utilizada para avaliar um sistema de RI possui três componentes: a coleção de documentos, um conjunto de consultas de teste e um conjunto de documentos relevantes para cada consulta de teste, determinado por especialistas no domínio do saber em questão. Descrevemos a seguir a constituição da coleção de referência, denominada coleção Juris, utilizada nesse trabalho.

A Coleção Juris

A coleção Juris é composta por documentos de três sub-coleções: a sub-coleção STF [56], composta por jurisprudências do Supremo Tribunal Federal, contendo 131.663 documentos; a sub-coleção STJ [57], composta por jurisprudências do Supremo Tribunal de Justiça, contendo 136.503 documentos; e a sub-coleção TRF's [60, 61, 62, 63, 64], composta por jurisprudências dos Tribunais Regionais Federais, contendo 284.407 documentos. No total, a coleção Juris é composta por 552.573 do-

cumentos (652 Mb). O tamanho médio dos documentos é de 164,68 termos por documento.

As Consultas de Teste

O conjunto de consultas de teste utilizado no processo de avaliação do modelo proposto foi constituído através dos procedimentos descritos a seguir.

Em abril/2002, obtivemos do CJF um conjunto de 744 consultas formuladas por usuários, público em geral e juízes, do sítio Internet do CJF [11]. Essas consultas foram analisadas e organizadas da seguinte forma: consultas com descrição clara de uma necessidade de informação (279), consultas a um documento específico (230), consultas a documentos com conteúdo similar ao de um documento citado (3), consultas com assunto muito genérico (30) e consultas com assuntos não associados às jurisprudências (202).

Do total de 744 consultas recebidas, 279 foram inicialmente selecionadas por possuírem a descrição clara de uma necessidade de informação. Essas 279 consultas passaram então pela avaliação de um especialista[1] que contribuiu com as seguintes informações adicionais: a qual Tribunal a consulta se referia e, para aquelas muito detalhadas ou redundantes, ele forneceu uma descrição resumida da intenção da consulta.

Um programa foi desenvolvido de tal forma que, para cada uma das 279 consultas avaliadas, o programa identificou as palavras-chave e os conceitos do tesauro presentes no texto da consulta. A partir dos conceitos identificados, o programa gerou quatro novos conjuntos de consultas com os conceitos associados àqueles mapeados e classificados de acordo com os relacionamentos do tesauro, (veja exemplo na Seção 5.1, Figura 5.1), ou seja, esse programa gerou seis conjuntos de consultas:

- **KY**: consultas compostos por palavras-chave;
- **CC**: consultas compostos por conceitos mapeados diretamente a partir do tesauro;
- **SY**: consultas compostos por sinônimos dos conceitos do conjunto **CC**;
- **TE**: consultas compostos por termos específicos dos conceitos do conjunto **CC**;
- **TG**: consultas compostos por termos genéricos dos conceitos do conjunto **CC**;

[1] Juliano Veloso Leite e Silva, estudante do 8º período do curso de Direito da Pontifícia Universidade Católica de Minas Gerais (PUC-MG).

- **TR**: consultas compostos por termos relacionados aos conceitos do conjunto **CC**.

Após essa análise, foram selecionadas 50 consultas para compor nossa coleção de referência (ou coleção de teste). As consultas selecionadas foram aquelas que possuíam todos os relacionamentos do Tesauro e que eram requisições direcionadas aos tribunais que contribuíram com os documentos que formam a coleção, ou seja, STF, STJ e TRF's. Tais restrições foram devidas ao fato de termos como objetivo avaliar a contribuição de conceitos e relacionamentos do tesauro como evidência semântica.

Avaliação de Relevantes

Na avaliação de novos algoritmos na área de RI, usa-se o conjunto de documentos relevantes para cada consulta de teste. Dada uma consulta de teste, o procedimento ideal é avaliar todos os documentos da coleção, classificando-os como relevantes ou irrelevantes em relação àquela consulta. Tal método é impraticável para grandes coleções.

Nesse trabalho foi utilizado o método *pooling* [69] para a constituição do conjunto de documentos relevantes para cada consulta de teste. Esse método, que é largamente utilizado pela comunidade de RI, é composto pelas seguintes etapas: para cada consulta de teste, os vários algoritmos a serem avaliados são executados, gerando conjuntos de documentos resposta ordenados por relevância. O conjunto de documentos candidatos para cada consulta é constituído pela união dos primeiros documentos dos conjuntos respostas dos vários algoritmos em análise.

Nesse trabalho, os 50 primeiros documentos resposta gerados pelas várias combinações Bayesianas produzidas por nosso modelo de rede de crenças foram utilizados para compor o *pool* de documentos candidatos para cada consulta de teste. Após eliminação de duplicatas, o *pool* analisado resultou em uma média de 319 documentos por consulta de teste. O conjunto de documentos candidatos foi então analisado pelo especialista da área de direito e cada documento foi classificado como relevante ou irrelevante para uma dada consulta[2]. Esse método garante a imparcialidade da avaliação, pois o especialista não sabe qual algoritmo gerou o documento que está sendo correntemente analisado.

[2]Os documentos classificados como parcialmente relevantes foram considerados irrelevantes, diferentemente do usual, pois as consultas são compostas por mais de um conceito. O objetivo dessa diretriz foi para que apenas documentos que atendessem a todos os conceitos fossem considerados relevantes.

Desse processo é gerado o conjunto de documentos relevantes para cada consulta de teste utilizado para comparar novos algoritmos.

Consultas de Teste Selecionadas

Finalizada a etapa de avaliação de relevantes, constatamos que 10 consultas, dentre as 50 originalmente selecionadas, não possuíam nenhum documento relevante e outras 10 possuíam poucos documentos relevantes. Com a ajuda do especialista, constatamos que tais consultas ou eram muito específicas ou eram compostas por mais de uma necessidade de informação, ou possuíam conceitos errados ou contraditórios (por exemplo, "Devedor Mutuante" quando deveria ser "Devedor Mutuário"). Como resultado, foram selecionadas 30 consultas para compor a coleção de referência utilizada em nossos experimentos. O número médio de documentos relevantes por consulta é de 33 documentos por consulta.

6.2 As Características do Sistema de RI Implementado

O sistema de RI implementado para a execução dos experimentos é constituído por dois módulos: o indexador e o processador de consultas, detalhados a seguir.

6.2.1 O Indexador

O módulo indexador foi responsável pela criação do índice, cujas entradas foram formadas pelas palavras-chave extraídas do texto dos documentos e pelos conceitos ou descritores do tesauro. As palavras-chave foram obtidas através de operações de transformação em maiúsculas e de eliminação de *stopwords* e de acentos. O índice foi gerado na forma de listas invertidas [5], ou seja, para cada item (palavra-chave ou conceito) foi criada uma lista de documentos nos quais o item ocorria, juntamente com a freqüência de ocorrência.

Para cada entrada do índice foi calculado o seu *idf* e para cada documento foi calculada sua norma. Tais medidas são utilizadas para calcular a similaridade entre uma consulta e os documentos da coleção e estão definidas na Seção 2.3. A principal novidade dessa implementação foi o enriquecimento do índice com os conceitos do tesauro, que julgamos ser um dos fatores que contribuiu para o bom desempenho dos resultados apresentados na Seção 6.3.

6.2.2 O Processador de Consultas

O módulo processador de consultas foi o responsável pela geração da lista ordenada de documentos retornados como resposta a uma consulta. De posse de uma consulta, esse módulo recupera do índice as listas de documentos equivalentes aos itens da consulta (palavras-chave ou conceitos). Com base nessa lista, o programa calculava a similaridade da consulta em relação a cada documento, gerando uma lista ordenada através das Equações 2.1 e 2.2.

Observamos que um módulo de interface com o usuário, mostrado na Figura 2.1, não foi implementado, uma vez que a parte experimental desse trabalho não contou com a participação de usuários.

6.3 Avaliação dos Resultados

A análise dos resultados experimentais é apresentada em três etapas. Na primeira, mostramos os resultados gerados pela utilização de palavras-chave e ordenação vetorial. Esses resultados, que servirão de base de comparação para os demais resultados, são referidos como KY (*keywords*). A seguir exploramos as potencialidades dos conceitos do Tesauro da Justiça Federal. Para isso, avaliamos os resultados obtidos com as combinações Bayesianas geradas pelo nosso modelo de rede de crenças. Por fim, comparamos ainda os resultados produzidos pelas combinações Bayesianas com os resultados produzidos pelo procedimento de expansão automática de consultas, procedimento muito comum na área de RI. Todos os resultados são mostrados como curvas de precisão-revocação e precisão média.

6.3.1 Resultados com Uso de Palavras-chave, Conceitos e Relacionamentos

Nesta seção avaliamos os resultados produzidos com o uso das evidências palavras-chave, conceitos e relacionamentos, separadamente. Aqui, o objetivo é verificar as contribuições possíveis de cada um em separado. Como podemos ver na Figura 6.1 e na Tabela 6.1, o resultado produzido pela evidência palavras-chave, curva KY, já é em si um bom resultado, com precisão média de 0,48. Além disso, todos os documentos relevantes são retornados no conjunto resposta. Esse fato se deve à composição das consultas, que têm um número médio de 9,0 palavras-chave por consulta. Esse número é três vezes maior que o tamanho médio das consultas submetidas por usuários eventuais na Internet (2 a 3 termos por consulta [14, 55]). Uma justificativa para esse fato está na origem das nossas consultas de teste. Conforme descrito na

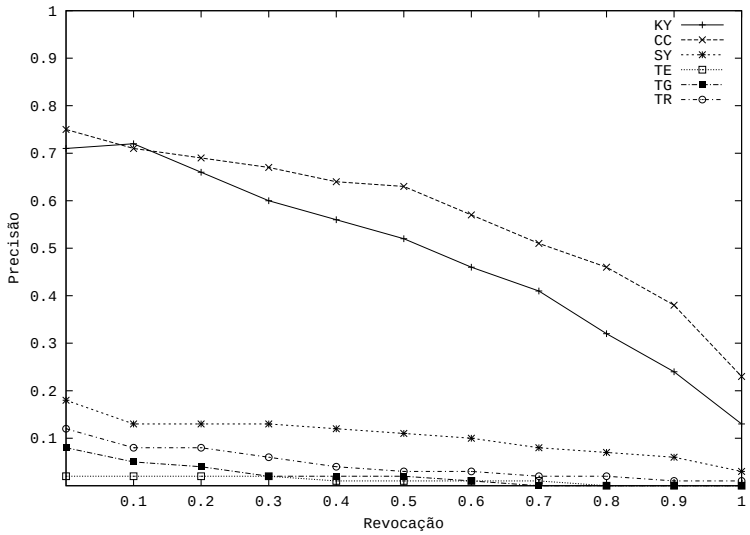

Figura 6.1: Resultados com palavras-chave, conceitos e relacionamentos

Seção 6.1, nossas consultas de teste têm origem em requisições feitas ao sítio Internet do Conselho de Justiça Federal [11], que é específico para a área jurídica. Esse é um indicativo de que os usuários do sítio do CJF possuem afinidades e interesses na área jurídica, ou seja, não se comportam como usuários eventuais.

Com relação aos conceitos associados às consultas, curva CC, observamos que eles são uma boa fonte de informação, com precisão média de 0,57. Esse resultado ultrapassa em precisão aqueles obtidos com o uso de palavras-chave, propiciando um ganho médio em precisão de 18,75%. Veja Tabela 6.1.

Com relação aos relacionamentos, curvas SY, TE, TG e TR, eles não são uma boa fonte de informação em termos de precisão dos resultados quando considerados separadamente. Observamos que somente sinônimos (SY) e termos relacionados (TR) retornam todos os documentos relevantes, ou seja, preservam os níveis de revocação relatados pela literatura. Ressaltamos, no entanto, que o procedimento aqui adotado considera os relacionamentos em separado, diferentemente dos outros pesquisadores que os consideram em conjunto com as palavras-chave, utilizando-os para expandir as consultas.

6.3.2 Resultados com Combinação Bayesiana

Nesta seção avaliamos, inicialmente, os resultados obtidos pelo modelo de rede de crenças proposto, com combinações dois a dois entre palavras-chave, conceitos e relacionamentos. A seguir avaliamos as combinações três a três e quatro a quatro,

entre palavras-chave, conceitos e relacionamentos. Combinações cinco a cinco e seis a seis não levam a melhores resultados e não acrescentam novas informações.

Combinações Dois a Dois

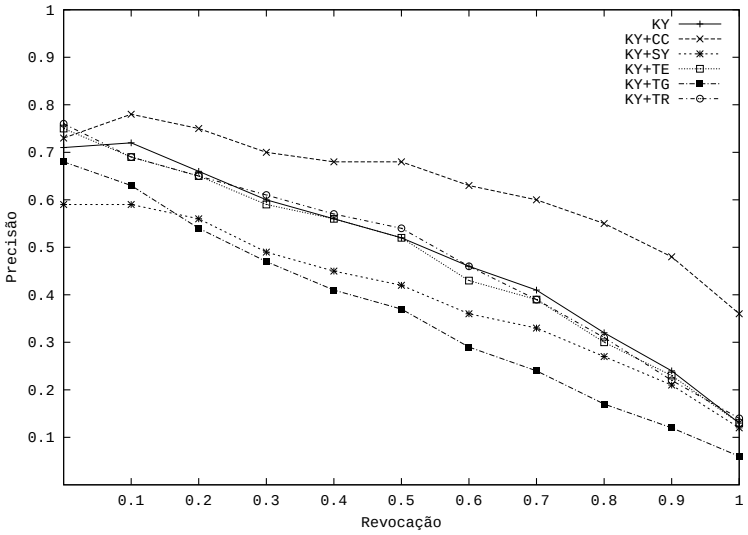

Figura 6.2: Resultados com as combinações Bayesianas dois a dois entre palavras-chave, conceitos e relacionamentos

A Figura 6.2 ilustra os resultados das combinações dois a dois que incluem palavras-chave. As demais combinações dois a dois, aquelas que não incluem palavras-chave, apresentam resultados similares ou inferiores. Conforme podemos observar, a combinação entre palavras-chave e conceitos, curva KY+CC, leva a resultados superiores, com precisão média de 0,63 (veja Tabela 6.1). Esse bom resultado é devido ao acúmulo de evidências produzido pela combinação do conhecimento do usuário, especificado na consulta (KY), com o conhecimento do especialista da área jurídica, especificado nos conceitos do tesauro (CC). Nosso modelo de rede de crenças possibilita acumular informação sobre essas duas fontes de evidência para melhorar os resultados. A combinação de evidências KY+CC propicia um ganho médio em precisão de 31,25%, relativo ao uso de palavras-chave somente (KY). Veja Tabela 6.1.

Com relação às combinações entre as palavras-chave e os relacionamentos, nossas observações são como se segue. As combinações com termos específicos (KY+TE) e termos relacionados (KY+TR) preservam os mesmos níveis de precisão do resultado com palavras-chave (KY), com precisão média de 0,48 e 0,49, respectivamente. Os sinônimos (KY+SY) e termos genéricos (KY+TG) degradam os resultados, com

precisão média de 0,40 e 0,36, respectivamente. Veja Tabela 6.1.

Com base nesses resultados, concluímos que os melhores resultados são gerados pela combinação de palavras-chave e conceitos (KY+CC). Assim sendo, consideramos a seguir somente combinações três a três e quatro a quatro, que envolvem palavras-chave (KY) e conceitos (CC). As demais combinações geram resultados inferiores.

Combinações Três a Três

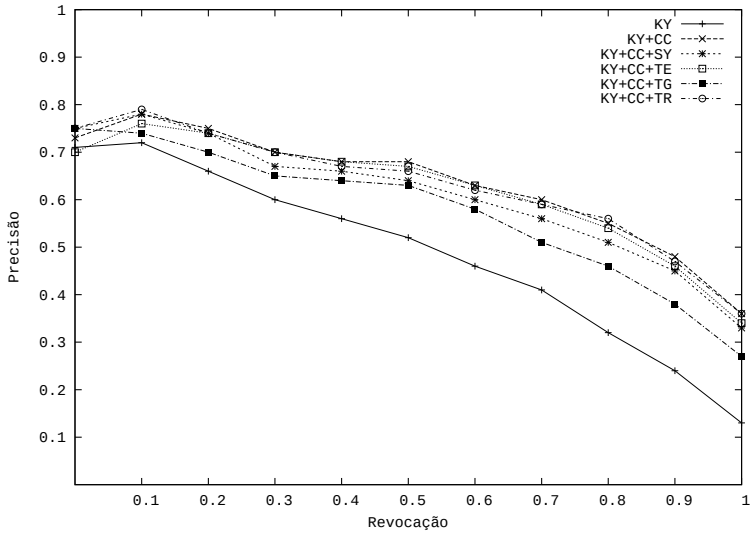

Figura 6.3: Resultados com as combinações Bayesianas três a três, envolvendo palavras-chave (KY), conceitos (CC) e um outro relacionamento (SY, TE, TG ou TR)

Avaliamos a seguir os resultados produzidos por combinações de evidências três a três, envolvendo palavras-chave (KY), conceitos (CC) e um relacionamento (SY, TE, TG ou TR). Na Figura 6.3 e na Tabela 6.2 observamos que há ganho em precisão para todas as combinações quando comparadas em relação à base, KY. Observamos ainda que as combinações com os relacionamentos termos específicos (KY+CC+TE) e termos relacionados (KY+CC+TR) proporcionam resultados equivalentes a KY+CC, com precisão média de 0,62 e 0,63, respectivamente. Os resultados com sinônimos (KY+CC+SY) e termos genéricos (KY+CC+TG) proporcionam resultados ligeiramente inferiores, com precisão média de 0,61 e 0,57, respectivamente. Ou seja, nenhuma combinação de evidência três a três ultrapassa os resultados obtidos com a combinação KY+CC, no caso da nossa coleção de referência.

Tabela 6.1: Ganhos e perdas com combinações dois a dois — R: Revocação, P: Precisão, G: Ganho relativo aos resultados obtidos com o uso de palavras-chave somente

R	KY	CC		KY+CC		KY+SY		KY+TE		KY+TG		KY+TR	
	P	P	G	P	G	P	G	P	G	P	G	P	G
0.0	0.71	0.75	5.63%	0.73	2.82%	0.59	-16.90%	0.75	5.63%	0.68	-4.23%	0.76	7.04%
0.1	0.72	0.71	-1.39%	0.78	8.33%	0.59	-18.06%	0.69	-4.17%	0.63	-12.50%	0.69	-4.17%
0.2	0.66	0.69	4.55%	0.75	13.64%	0.56	-15.15%	0.65	-1.52%	0.54	-18.18%	0.65	-1.52%
0.3	0.60	0.67	11.67%	0.70	16.67%	0.49	-18.33%	0.59	-1.67%	0.47	-21.67%	0.61	1.67%
0.4	0.56	0.64	14.29%	0.68	21.43%	0.45	-19.64%	0.56	0.00%	0.41	-26.79%	0.57	1.79%
0.5	0.52	0.63	21.15%	0.68	30.77%	0.42	-19.23%	0.52	0.00%	0.37	-28.85%	0.54	3.85%
0.6	0.46	0.57	23.91%	0.63	36.96%	0.36	-21.74%	0.43	-6.52%	0.29	-36.96%	0.46	0.00%
0.7	0.41	0.51	24.39%	0.60	46.34%	0.33	-19.51%	0.39	-4.88%	0.24	-41.46%	0.39	-4.88%
0.8	0.32	0.46	43.75%	0.55	71.88%	0.27	-15.62%	0.30	-6.25%	0.17	-46.88%	0.31	-3.13%
0.9	0.24	0.38	58.33%	0.48	100.00%	0.21	-12.50%	0.23	-4.17%	0.12	-50.00%	0.22	-8.33%
1.0	0.13	0.23	76.92%	0.36	176.92%	0.12	-7.69%	0.13	0.00%	0.06	-53.85%	0.14	7.69%
Média	0.48	0.57	18.75%	0.63	31.25%	0.40	-16.67%	0.48	0.00%	0.36	-25.00%	0.49	2.08%

Tabela 6.2: Ganhos com combinações três a três - R: Revocação, P: Precisão, G: Ganho relativo aos resultados obtidos com o uso de palavras-chave, conceitos e relacionamentos

R	KY	KY+CC		KY+CC+SY		KY+CC+TE		KY+CC+TG		KY+CC+TR	
	P	P	G	P	G	P	G	P	G	P	G
0.0	0.71	0.73	2.82%	0.75	5.63%	0.70	-1.41%	0.75	5.63%	0.75	5.63%
0.1	0.72	0.78	8.33%	0.78	8.33%	0.76	5.56%	0.74	2.78%	0.79	9.72%
0.2	0.66	0.75	13.64%	0.74	12.12%	0.74	12.12%	0.70	6.06%	0.74	12.12%
0.3	0.60	0.70	16.67%	0.67	11.67%	0.70	16.67%	0.65	8.33%	0.70	16.67%
0.4	0.56	0.68	21.43%	0.66	17.86%	0.68	21.43%	0.64	14.29%	0.67	19.64%
0.5	0.52	0.68	30.77%	0.64	23.08%	0.67	28.85%	0.63	21.15%	0.66	26.92%
0.6	0.46	0.63	36.96%	0.60	30.43%	0.63	36.96%	0.58	26.09%	0.62	34.78%
0.7	0.41	0.60	46.34%	0.56	36.59%	0.59	43.90%	0.51	24.39%	0.59	43.90%
0.8	0.32	0.55	71.88%	0.51	59.38%	0.54	68.75%	0.46	43.75%	0.56	75.00%
0.9	0.24	0.48	100.00%	0.45	87.50%	0.46	91.67%	0.38	58.33%	0.47	95.83%
1.0	0.13	0.36	176.92%	0.33	153.85%	0.34	161.54%	0.27	107.69%	0.36	176.92%
Média	0.48	0.63	31.25%	0.61	27.08%	0.62	29.17%	0.57	18.75%	0.63	31.25%

Combinações Quatro a Quatro

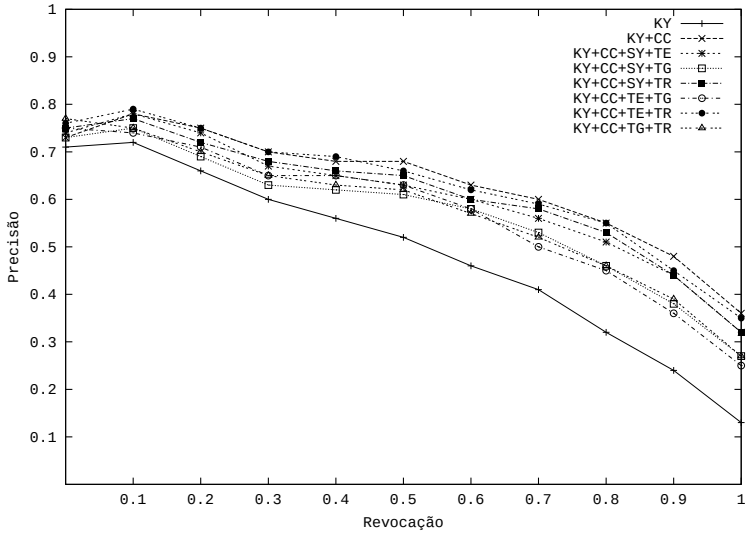

Figura 6.4: Resultados com as combinações Bayesianas quatro a quatro, envolvendo palavras-chave (KY), conceitos (CC) e dois outros relacionamentos

A Figura 6.4 ilustra os resultados com combinação de evidência quatro a quatro (somente os melhores resultados são mostrados). Observamos a mesma tendência da Figura 6.3. Todas as combinações de KY+CC com dois outros relacionamentos proporcionam ganho em precisão superiores a KY, mas nenhuma combinação ultrapassa os resultados obtidos com a combinação KY+CC. Observamos também que todas as combinações envolvendo termos genéricos (KY+CC+SY+TG, KY+CC+TE+TG, KY+CC+TG+TR) levam aos piores resultados.

As demais combinações, cinco a cinco e seis a seis, não levam a melhores resultados. Como já observado, os melhores resultados são obtidos com a combinação KY+CC.

6.3.3 Comparação entre Combinação Bayesiana e Expansão Automática de Consultas

Para ilustrar a diferença entre combinação Bayesiana e o procedimento de expansão automática de consultas utilizando conceitos e relacionamentos do tesauro, avaliamos a seguir os resultados obtidos com a utilização dessas duas técnicas.

No procedimento de expansão automática de consultas, conceitos (ou termos) relacionados às palavras-chave da consulta original são adicionados à consulta, ex-

pandindo-a. A nova consulta expandida é então processada. Os resultados obtidos são ordenados utilizando, por exemplo, a fórmula do cosseno do modelo vetorial.

Na literatura técnica, diferentes estratégias de modificação de pesos vetoriais associados aos termos originais da consulta e aos novos termos para expansão têm sido discutidas (veja Seção 7.1). Pode-se, por exemplo, atribuir pesos menores aos novos termos de modo a manter o "foco" da consulta original. Essas estratégias de modificação de pesos vetoriais fogem ao escopo de nosso trabalho e não são aqui discutidas.

A Figura 6.5 ilustra os resultados produzidos pela combinação Bayesiana (KY+-CC+SY+TE-TG+TR) e por expansão automática de consultas (KY.CC.SY.TE.-TG.TR) quando todas as evidências são consideradas. Observa-se que a combinação Bayesiana leva a ganhos médios de precisão com relação a nossa curva de referência (KY) da ordem de 18% e que expansão automática de consultas leva à redução de precisão da ordem de -4%. Veja Tabela 6.3.

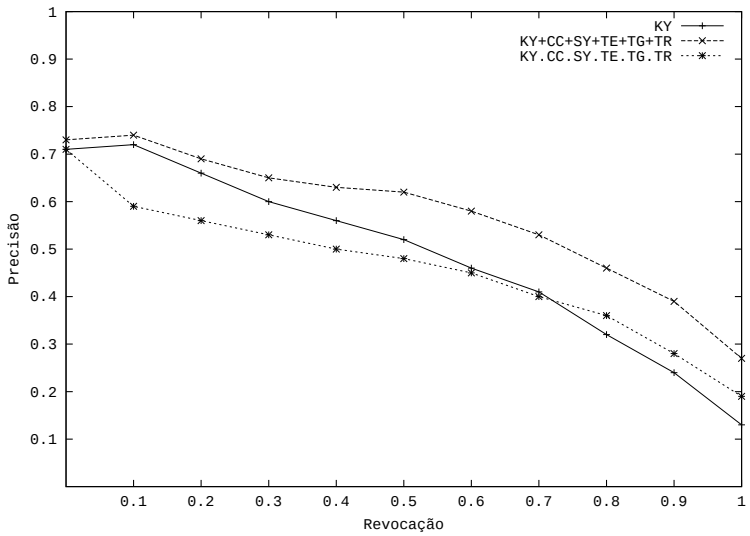

Figura 6.5: Comparação entre combinação Bayesiana e expansão automática de consultas utilizando todas as evidências

A razão é que evidências como termos relacionados recuperam novos documentos que, no caso de expansão automática de consultas, recebem valores elevados de similaridade para com a consulta expandida, afetando negativamente os resultados de precisão média. Em nosso modelo de rede de crenças, esses efeitos negativos são "controlados" em dois pontos: (a) somente resultados finais de similaridade entre documentos são combinados (ao contrário da técnica de expansão automática de consultas que agrega contribuições parciais, os novos termos e seus pesos, para

calcular um novo valor de similaridade) e (b) a combinação final de evidências é feita através de um operador disjuntivo que, naturalmente, varia de forma menos acentuada.

Tabela 6.3: Ganho com a combinação Bayesiana (KY+CC+SY+TE+TG+TR) e perda com a expansão automática de consultas utilizando todas as evidências (KY.CC.SY.TE.TG.TR)

	KY	KY+CC+SY+TE+TG+TR		KY.CC.SY.TE.TG.TR	
R	P	P	G	P	G
0.0	0.71	0.73	2.82%	0.71	0.00%
0.1	0.72	0.74	2.78%	0.59	-18.06%
0.2	0.66	0.69	4.55%	0.56	-15.15%
0.3	0.60	0.65	8.33%	0.53	-11.67%
0.4	0.56	0.63	12.50%	0.50	-10.71%
0.5	0.52	0.62	19.23%	0.48	-7.69%
0.6	0.46	0.58	26.09%	0.45	-2.17%
0.7	0.41	0.53	29.27%	0.40	-2.44%
0.8	0.32	0.46	43.75%	0.36	12.50%
0.9	0.24	0.39	62.50%	0.28	16.67%
1.0	0.13	0.27	107.69%	0.19	46.15%
Média	0.48	0.57	18.75%	0.46	-4.17%

Os resultados acima indicam a superioridade do modelo proposto em relação à abordagem de expansão automática de consultas.

Capítulo 7

Trabalhos Relacionados

Neste Capítulo discutimos os trabalhos relacionados à expansão de consultas, ao uso de tesauros em RI e aos sistemas de RI para a área jurídica.

7.1 Expansão de Consultas

O processo de selecionar documentos relevantes a uma necessidade de informação pode exigir várias interações e reformulações da consulta inicial, tornando-se um processo trabalhoso e demorado. Uma estratégia para simplificar esse processo é expandir a consulta inicial com termos relacionados, numa tentativa de melhorar o contexto da consulta e de minimizar o problema de diferenças no uso das palavras da língua por usuários e autores de documentos. Esse problema é referenciado como *word mismatch* e está relacionado ao uso de diferentes palavras para referenciar um mesmo conceito e de palavras com diferentes significados.

O processo de expansão de consultas deve sempre levar em conta: como selecionar os termos a serem utilizados na expansão; quais e quantos termos devem ser utilizados e onde obtê-los.

A seleção de termos pode ser feita com duas abordagens distintas, a interativa ou a automática. Na interativa, o usuário interage com o sistema para indicar quais termos ou documentos devem ser utilizados na expansão da consulta, técnica denominada realimentação de relevantes (*relevance feedback*). Na abordagem automática, o sistema se encarrega de determinar, sem auxílio do usuário, os termos a serem adicionados à consulta inicial. A definição de qual abordagem o sistema de RI vai adotar depende, principalmente, da comunidade de usuários. Segundo Greenberg [21], usuários especialistas têm disponibilidade de fornecer informações em um processo interativo, enquanto usuários eventuais e pouco conhecedores de uma área preferirão a abordagem automática. Por essa razão muitos autores têm adotado a

abordagem automática.

Com relação a quais termos utilizar para expandir a consulta, o método proposto por Rocchio [42], descrito a seguir, aborda esse problema através do recálculo dos pesos dos termos da consulta e a inserção de novos termos. Com relação à quantidade de termos utilizados na expansão, vários autores determinam esse valor experimentalmente.

As fontes mais utilizadas para a obtenção de termos para a expansão da consulta original são os documentos da coleção e bases de conhecimento externas, como os tesauros. A utilização de tesauros em expansão de consultas será apresentada da Seção 7.2. Com relação à expansão de consultas a partir da coleção de documentos, existem duas técnicas distintas: a análise global e a análise local. Na global, os relacionamentos existentes entre os termos são obtidos utilizando-se todos os documentos da coleção. Na análise local, apenas os documentos de topo, aqueles que aparecem melhor colocados na ordenação das respostas, são utilizados para a obtenção de novos termos. Ou seja, assume-se que os documentos do topo são mais importantes. Essa abordagem é denominada pseudo-realimentação de relevantes.

Uma das técnicas de análise global é o agrupamento de termos (*term clustering*), através da qual os grupos de termos são encontrados com base na co-ocorrência de termos em todos os documentos da coleção. Tal procedimento é computacionalmente complexo e não produz resultados efetivos [8]. Além disso, segundo Baeza-Yates e Ribeiro-Neto [5], estruturas globais não se adaptam bem ao contexto local de uma consulta, pois correlações existentes na coleção inteira podem não valer no contexto específico de uma consulta, especialmente para coleções genéricas.

O método mais tradicional de expansão de consultas é denominado realimentação de relevantes (*relevance feedback*), que envolve análise local e participação do usuário. Esse método visa obter mais informação do usuário com o objetivo de melhorar a especificação da consulta e ocorre como se segue. O conjunto de documentos gerado como resposta à consulta original é mostrado ao usuário que assinala aqueles que considera relevantes e aqueles que considera não relevantes. A partir da informação do usuário, o sistema deve reformular a consulta. O objetivo é que a nova consulta seja capaz de separar os documentos relevantes dos não relevantes. Rocchio [42] propôs um método para combinar expansão da consulta com recálculo de peso dos termos a serem utilizados e inserção de novos termos, conforme a fórmula a seguir, denominada Rocchio padrão:

$$QE = \alpha\ Q + \frac{\beta}{|R|} \sum_{i=1}^{|R|} R_i - \frac{\gamma}{|nR|} \sum_{i=1}^{|nR|} nR_i \tag{7.1}$$

onde QE é o vetor da consulta expandida, Q é o vetor da consulta original, $|R|$ é o número de documentos relevantes para a consulta Q, conforme indicado pelo usuário, R_i é o vetor do documento relevante i, $|nR|$ é o número de documentos não relevantes, conforme indicado pelo usuário, e nR_i é o vetor do documento não relevante i. As constantes α, β e γ podem ser variadas para determinar a importância dos termos da consulta original, dos termos provenientes dos documentos relevantes e dos termos provenientes dos documentos não relevantes, respectivamente, na consulta expandida. A fórmula original de Rocchio considera $\alpha = \beta = \gamma = 1$.

Variações dessa equação foram propostas ao longo dos anos. O principal questionamento diz repeito ao efeito negativo dos documentos não relevantes. Ide [23] propôs que apenas o documento não relevante mais alto na lista ordenada de respostas seja utilizado, ou seja, ela propõe minimizar os efeitos da realimentação negativa.

Salton e Buckley [48] realizaram uma série de experimentos, com diferentes coleções e diferentes parâmetros, e verificaram que a modificação proposta por Ide gera melhores resultados.

Losada e Barreiro [31] argumentaram que informações de documentos não relevantes devem ser consideradas, pois podem ajudar a determinar termos da consulta que só tenham retornado documentos não relevantes. Esses termos poderiam então ser removidos da consulta.

A análise local, denominada pseudo-realimentação de relevantes, é similar à realimentação de relevantes. Entretanto, a alteração ocorre sem a ajuda do usuário. Os documentos do topo são considerados relevantes e a fórmula de Rocchio é utilizada desprezando-se a parte relativa aos documentos não relevantes. Usualmente os termos mais freqüentes são selecionados. Pesquisas recentes mostram algum sucesso utilizando análise local para grandes coleções [70, 72, 73], tanto em relação a revocação quanto em relação a precisão, diferentemente de pesquisas anteriores [18] que relatam perda em precisão.

Voorhees e Harman [70] relataram que os experimentos de grupos da TREC, utilizando análise local para expansão de consultas, produziram melhoria nos resultados, especialmente quando há participação do usuário.

Xu e Croft [72, 73] propuseram uma técnica denominada "análise de contexto local" (*local context analysis*) que combina as técnicas de análise global e local. A co-ocorrência é determinada apenas para os termos da consulta na coleção local, ou seja, nos documentos do topo. Além disso, para evitar problemas de documentos longos e que tratam de vários assuntos, esses autores utilizaram técnicas de "passagem" para encontrar termos que co-ocorrem com os termos da consulta. Os termos escolhidos são aqueles que co-ocorrem com o maior número de termos da consulta.

Esses autores concluíram que expansão de consulta é uma ferramenta para melhorar tanto a revocação quanto a precisão e sugeriram que, para consultas bem formadas, deve-se considerar pesos menores para os termos da expansão. Tais resultados são consistentes com [48].

Utilizando análise local, Carpineto et al. [8] propuseram excluir a parte da fórmula referente aos documentos não relevantes e que os termos fossem selecionados observando-se a diferença de distribuição dos mesmos no conjunto pseudo-relevante e na coleção. Eles argumentaram que esse valor é um indicador da diferença semântica dos termos dos documentos da coleção local em relação à consulta. Os termos de maior peso são selecionados. Esses autores sinalizaram que seu método é mais apropriado para consultas pequenas, fato que requer mais experimentação.

Mitra et al. [35] propuseram um método automático para reordenar os documentos do conjunto resposta, para então utilizar os documentos do topo para expandir a consulta. Seu método utiliza co-ocorrência e proximidade de palavras da consulta nos documentos do topo para reordenar os documentos da resposta. A proximidade é determinada em "passagens" de tamanho variável. Esses autores reportaram resultados experimentais para TREC (3 a 6) com diferentes variações para o tamanho da coleção local e das "passagens". O seu método apresenta ganho, no entanto com grande variabilidade para as diferentes coleções e para os diferentes parâmetros.

Chang e Hsu [10] utilizaram agrupamento de documentos, *document clustering*, com o objetivo de agrupar os documentos da resposta para simplificar a expansão de consultas. Os grupos formados são mostrados ao usuário que seleciona o mais apropriado. Os termos mais relevantes do grupo escolhido são utilizados para expandir a consulta. Seus resultados indicaram que expansão da consulta com os termos do grupo escolhido causam perda de precisão, enquanto a utilização de termos de documentos selecionados dentro do melhor grupo possibilita pequeno ganho em precisão.

Allan [2] argumentou que o uso de documentos longos (acima de 500 palavras) na expansão de consultas degrada o desempenho. Esse autor sugeriu que o uso de "passagens" de documentos grandes possibilita resultados superiores em relação ao descarte de tais documentos. Seus resultados indicaram ganho em precisão média de 32% quando considera "passagem" de 100 ou 200 palavras em documentos de até 300 palavras.

O método de realimentação de relevantes utilizando os documentos do topo está restrito ao resultado da consulta inicial, pois, se a consulta não retorna nenhum documento relevante, esse método não é útil. Um outro problema da análise local é que no conjunto pseudo-relevante pode existir uma grande fração de documentos

não relevantes, documentos esses que podem introduzir termos não relacionados à consulta original, tirando o foco da mesma.

7.2 Uso de Tesauros em RI

Muitos autores têm utilizado tesauros para fazer expansão automática de consultas [20, 21, 27, 32, 33, 38, 68]. Greenberg [20, 21] investigou o uso de um tesauro da área comercial para expansão automática de consultas. Os experimentos foram executados sobre uma coleção de documentos também da área comercial e as consultas foram obtidas através de questões reais formuladas por estudantes da área comercial. Dentre as questões, a autora selecionou somente as que possuíam conceitos do tesauro em uso. Essa autora utilizou o tesauro somente como ferramenta de pesquisa (*search aid thesaurus*). Com base em resultados experimentais, sinônimos e termos específicos são bons candidatos se expansão automática de consultas for utilizada. Por outro lado, termos relacionados e termos genéricos são bons candidatos se expansão interativa de consultas for utilizada.

Kristensen [27] estudou o uso de sinônimos, termos específicos e termos relacionados para expansão de consultas em uma coleção de texto de artigos de jornais. Essa autora também utilizou o tesauro somente como ferramenta de pesquisa. Ela encontrou melhoria em revocação e redução em precisão e concluiu que tesauros são "claramente ferramentas para melhorar revocação". Sua pesquisa mostrou especificamente que sinônimos e termos relacionados são equivalentes em revocação e que sinônimos degradam mais a precisão que termos relacionados.

Mandala et al. [32, 33] utilizaram três tipos de tesauros para selecionar termos para expansão automática de consultas. Um tesauro construído manualmente e dois construídos automaticamente, sendo um utilizando co-ocorrência e o outro utilizando relações lingüísticas através de lógica de predicados. Esses autores desenvolveram algoritmos para estimar a probabilidade de que os termos de cada tesauro eram bons candidatos para a expansão da consulta. Eles utilizaram a coleção TREC7 com título e descrição e mostraram experimentalmente que a combinação dos três tesauros possibilitaram melhores resultados que cada um em separado. Seus resultados indicaram ganhos de 2% a 39% em precisão.

Qiu e Frei [38] propuseram um algoritmo para determinar pesos para selecionar os termos do tesauro a serem utilizados na expansão automática da consulta. O tesauro foi construído automaticamente com base em uma matriz de similaridade termo-termo. Esses autores consideraram a consulta como conceito, isto é, consideraram-na como um termo virtual representado pelo centróide dos termos da

consulta. Os termos candidatos utilizados para a expansão foram os termos da coleção mais similares ao termo virtual. Eles argumentaram que considerar a consulta como conceito é uma abordagem superior e reportaram ganhos de 20 a 30% em precisão média.

Segundo Voorhees [68], o uso de tesauros genéricos em coleções genéricas apresenta pouca contribuição, especialmente para consultas bem construídas. Consultas pequenas podem ser melhoradas com utilização de tesauros genéricos, no entanto, a quantificação da melhoria não é apresentada pela autora. Seus experimentos foram realizados com parte das coleções TREC-1 e TREC-2 e com os sinônimos extraídos do WordNet.

Clark et al. [12] utilizaram um tesauro para encontrar especialistas em certos assuntos em uma grande companhia. No sistema proposto, o especialista usa o tesauro para descrever o seu perfil e a descrição é armazenada como um documento. A interface de pesquisa foi utilizada pelos funcionários da empresa para encontrar especialistas com um determinado perfil. O tesauro foi utilizado para auxiliar na descrição desse perfil. O usuários reportaram satisfação no uso do sistema.

7.3 Sistemas de RI para a área Jurídica

A importância da pesquisa na área jurídica vem crescendo devido ao aumento da disponibilidade de documentos jurídicos em forma digital. Dempsey et al. [16] utilizaram um tipo de diretório no qual o usuário escolhia o tipo de documento (jurisprudência, lei, etc) e o tribunal. Além disso, o usuário podia especificar a consulta em linguagem natural. O sistema oferecia também a possibilidade de criação de uma coleção virtual. A consulta e os parâmetros eram submetidos a um conjunto de *sites* especializados na área jurídica. As páginas retornadas eram indexadas formando a coleção virtual e os usuários podiam fazer pesquisa nessa coleção. Os autores argumentaram que essa estratégia aumenta a cobertura e que isso é importante devido à característica dinâmica de geração de novos casos julgados na área jurídica.

Alguns autores têm utilizado técnicas da área de Inteligência Artificial (IA) em suas propostas para sistemas da área jurídica [6, 17, 43, 59, 71]. Dick [17] propôs um modelo para recuperar jurisprudências com ênfase no desenvolvimento de representação do conhecimento que é baseada em gramática. Seu foco foi em descobrir argumentos ou parte de argumentos que podiam ser úteis ao usuário. Essa autora utilizou um subconjunto da área jurídica de leis de contrato. Ela argumenta que essa opção é devida à clareza dos conceitos e precisão da linguagem utilizada nesses casos. Ela não apresentou resultados experimentais.

Rose and Belew [43] utilizaram uma combinação das técnicas simbólica e conexionista de IA. O paradigma simbólico "... é baseado na idéia que o núcleo da inteligência está na manipulação explícita, largamente seqüencial, de símbolos". O paradigma conexionista acredita que "... a inteligência é uma propriedade emergente da interconexão de um grande número de elementos de processamento muito simples que operam em paralelo e que comunicam somente informações sub-simbólicas". Esses autores afirmaram que uma jurisprudência possui informações simbólicas e sub-simbólicas e utilizaram os dois tipos de informação para construir seu sistema. Eles não apresentaraam resultados experimentais.

Bueno et al. [6] and Weber [71] utilizaram técnica de raciocínio baseado em caso (*case-base reasoning*) para encontrar jurisprudências similares em um banco de dados jurídico. Essa técnica utiliza o princípio de analogia no qual assume-se que problemas similares possuem soluções similares. Esse raciocínio foi aplicado a casos passados e a novos casos da área jurídica. Com essa técnica, Weber [71] argumentou que é possível modelar aspectos do raciocínio humano e que é similar ao julgamento feito por especialistas humanos. Esse trabalho é similar a uma aplicação de banco de dados com tuplas atributo-valor sendo extraídas do texto dos documentos. Essa autora utilizou uma pequena coleção jurídica brasileira (com 3.446 documentos), mas não apresentou resultados experimentais.

Bueno et al. [6] utilizaram dois tipos de vocabulários controlados. O primeiro, denominado Vocabulário Controlado, possui uma lista de palavras-chave normativas extraídas da literatura (Códigos e Normas). O segundo, denominado Tesauro Jurídico, foi construído por especialistas da área jurídica. Esse trabalho é também similar a uma aplicação de banco de dados com tuplas atributo-valor sendo extraídas do texto dos documentos. Eles utilizaram uma coleção jurídica brasileira, mas, infelizmente, não apresentaram resultados experimentais que possam ser comparados com o nosso trabalho.

Tong et al. [59] propuseram que o conhecimento sobre a requisição a ser pesquisada seja codificada como uma coleção de regras, uma representação do conhecimento baseada em regras (*rule-based knowledge representation*). O usuário podia especificar os relacionamentos entre conceitos e também a força desses relacionamentos. As regras eram do tipo "*antecedente regra conseqüente(força)*", na qual *antecedente* pode ser composto por um número de conceitos ou texto conectado por operadores, *regra* é um conjunto de regras pré-definidas e *força* é o grau em que o antecedente interfe no conseqüente. Esses autores não apresentaram resultados experimentais.

Resultados preliminares apresentados em Silveira et al. [53], baseados na utilização de conceitos e termos específicos de um tesauro jurídico brasileiro, reportaram

aumento de precisão quando as fontes de evidência são combinadas em uma rede Bayesiana. Os resultados mostraram ainda que a combinação Bayesiana é uma abordagem superior à forma tradicional de expansão de consultas.

O foco do nosso trabalho está na investigação de como os conceitos e relacionamentos de um tesauro podem ser utilizados para melhorar a ordenação dos documentos do conjunto resposta em um sistema de Recuperação Vertical de Informação para a área jurídica. Contrário ao trabalho apresentado em [20, 21], aqui é utilizado o arcabouço das redes Bayesianas como arcabouço teórico. Esse modelo propicia modularidade e extensibilidade no sentido de que novas evidências podem ser naturalmente incorporadas ao modelo.

Capítulo 8

Conclusões

Os sistemas de Recuperação de Informação (RI) tradicionais utilizam casamento de padrão entre as palavras-chave da consulta do usuário e dos documentos da coleção para recuperar informação. Isso implica que o usuário tem de traduzir sua necessidade de informação em uma consulta composta por palavras-chave. A dificuldade aumenta à medida que o usuário tradicionalmente constrói consultas pequenas, com duas ou três palavras-chave somente, o que gera consultas ambíguas, especialmente em coleções genéricas que tratam de vários assuntos. Esse fato implica que o sistema gera uma longa lista de documentos como resposta, e o problema do usuário passa a ser então selecionar aqueles que são de seu interesse.

Nesse trabalho foi proposto um modelo teórico para lidar com o problema de ordenação dos resultados em um sistema de RI, o qual utiliza o arcabouço teórico das redes Bayesianas, as redes de crenças. O novo modelo, denominado *Sistema de Recuperação Vertical de Informação*, propõe uma extensão ao modelo de rede de crenças para RI [5], no qual o vocabulário de um tesauro é utilizado como evidência semântica suplementar. A idéia foi utilizar o conhecimento codificado no tesauro para melhorar a especificação do conteúdo das consultas e dos documentos.

Embora nosso modelo de rede de crenças seja genérico e possa ser utilizado com tesauros de diferentes áreas, nesse trabalho verificamos experimentalmente a qualidade do modelo para uma área específica de conhecimento, a área jurídica. O tesauro utilizado foi o Tesauro Jurídico da Justiça Federal, construído para representar os conceitos mais importantes da área jurídica brasileira. Os experimentos foram executados sobre uma coleção de documentos jurídicos advindos dos tribunais superiores brasileiros, o STF, o STJ e dos tribunais federais, os TRF's.

Os resultados indicam que o Tesauro da Justiça Federal é uma fonte importante de informação evidencial. Se consideramos consultas compostas somente com os conceitos do tesauro relacionados às consultas originais, obtemos um resultado com

precisão média de 57%, que é 18% melhor que os resultados obtidos com as consultas originais. Mais importante, os resultados gerados pela combinação Bayesiana entre as consultas orignais e os conceitos a ela vinculados são ainda melhores, com precisão média de 63%, o que representa um ganho de precisão da ordem de 31% em relação aos resultados gerados pelas consultas originais. Esse ganho em precisão se deve ao nosso modelo de rede de crenças que permite combinar informação proveniente do conhecimento do usuário, especificado na consulta através das palavras-chave, com informação proveniente do conhecimento dos especialistas da área jurídica, especificado nos conceitos do tesauro. Esse resultado confirma a primeira hipótese de trabalho.

As demais fontes de evidência — sinônimos, termos específicos, termos genéricos e termos relacionados — também permitem ganhos em precisão média em relação às consultas originais, confirmando a segunda hipótese de trabalho. No entanto, esses relacionamentos não permitem ganhos adicionais em precisão em relação à combinação de consultas originais e conceitos.

Assim, o tesauro da Justiça Federal se mostrou útil como fonte de conhecimento organizado e pode ser utilizado para melhorar a qualidade dos resultados. Temos, pois, um *Sistema de Recuperação Vertical de Informação*.

Como trabalhos futuros, sugerimos pesquisas nas seguintes direções. Primeiro, realizar experimentos com novas consultas para explorar, por exemplo, o efeito do uso do Tesauro da Justiça Federal nos resultados gerados por consultas pequenas. Segundo, realizar experimentos através dos quais o usuário possa interagir com o sistema para selecionar conceitos e relacionamentos do tesauro a serem considerados durante o processamento da consulta. Terceiro, avaliar nosso modelo com tesauros de outras áreas de conhecimento.

Apêndice A

As Famílias de Sistemas de Direito e o Poder Judiciário no Brasil

Neste Apêndice apresentamos uma pequena introdução das maiores famílias de direito vigentes atualmente com o objetivo de situar o sistema de direito brasileiro. A seguir introduzimos a organização do Poder Judiciário no Brasil.

A.1 As Famílias de Sistemas de Direito

Cada país elabora ao longo de sua história um corpo de princípios e regras jurídicas que regem as relações entre pessoas, de forma a garantir os direitos individuais e coletivos. Tais princípios e regras constituem o sistema jurídico ou o sistema de direito. Por razões históricas, os vários países adotam distintos sistemas de direito. Apesar de tal diversidade, os principais sistemas de direito podem ser categorizados em quatro grandes famílias [15, 54]: a família romano-germânica ou *Civil Law*, a *Common Law*, a socialista e outras. A quarta família inclui, por exemplo, os sistemas muçulmano e indiano. Nesta Seção, discutiremos brevemente os dois primeiros sistemas de direito que são utilizados no ocidente.

O sistema *Civil Law* teve origem na Roma antiga. É um sistema jurídico baseado no direito político no qual as leis são escritas por representantes eleitos pelo povo, o Poder Legislativo. A lei é tão universal e abstrata quanto possível, constitui uma sistematização de princípios gerais e possui uma apresentação lógica, racional e formal. Nesse sistema, a primeira fonte do direito é a lei escrita e, se existir alguma lacuna, os processos passados já julgados, as jurisprudências, são utilizados como base legal para julgamento de processos atuais. A lei é a regra e as jurisprudências são a exceção. A via principal é do universal para o particular.

O sistema *Common Law* teve origem na Inglaterra antiga. É um sistema jurídico

baseado no costume e constituído por uma sociedade solidamente construída sob os pilares da história e da tradição. Sua principal característica é a primazia das jurisprudências (*the case law*). Os casos passados são utilizados como precedentes e sustentadores de argumentos nos casos presentes. Se existe alguma lacuna nos processos já julgados, os juízes e advogados utilizam-se das leis escritas como base legal. Nesse sistema, os processos julgados são a regra e a lei escrita é a exceção. A via principal é do particular para o universal.

É importante notar que atualmente observa-se uma confluência dos princípios e práticas dos sistemas *Civil Law* e *Common Law*. Nos Estados Unidos, por exemplo, país que adota o sistema da *Common Law*, problemas sociais críticos sem consenso entre as partes terminam por serem decididos em processos nos tribunais. Em tais processos, os juízes, que julgaram casos passados e que julgam os atuais, exercem grande poder sobre a sociedade. Como resultado, a sociedade americana tem questionado e debatido o poder exercido pelos juízes. Em oposição, nos países de herança latina como o Brasil, a sociedade tem questionado a burocracia excessiva criada pelo sistema da *Civil Law*, no qual o Poder Legislativo escreve um grande número de leis, muitas delas para lidar com aspectos simples e corriqueiros da vida diária ou para situações muito particulares e outras tantas que não são postas em prática.

Por causa de tais questionamentos feitos pelas sociedades, existe uma tendência no sentido de que o sistema *Civil Law* seja alterado para adotar alguns princípios e práticas do *Common Law*. Também existe a tendência inversa, ou seja, que o sistema *Common Law* seja alterado para adotar alguns princípios e práticas do *Civil Law*.

Para atender ao cumprimento dos princípios e regras que compõem um sistema legal, cada país adota uma organização judiciária. A organização judiciária brasileira é discutida brevemente na Seção A.2.

A.2 O Poder Judiciário no Brasil

A organização do Poder Judiciário brasileiro é definida pela Constituição Federal promulgada em 1988 [9, 67, 36] e por leis escritas posteriormente. É organizado por tribunais de primeira e segunda instância e por tribunais superiores. A Figura A.1 ilustra o organograma do Poder Judiciário brasileiro e apresenta suas respectivas siglas[1].

Ao STF compete a guarda da Constituição, ou seja, compete processar e julgar as causas que envolvam a Constituição, como as ações de inconstitucionalidade e ações

[1] A direção das setas indica a seqüência em que os recursos ocorrem.

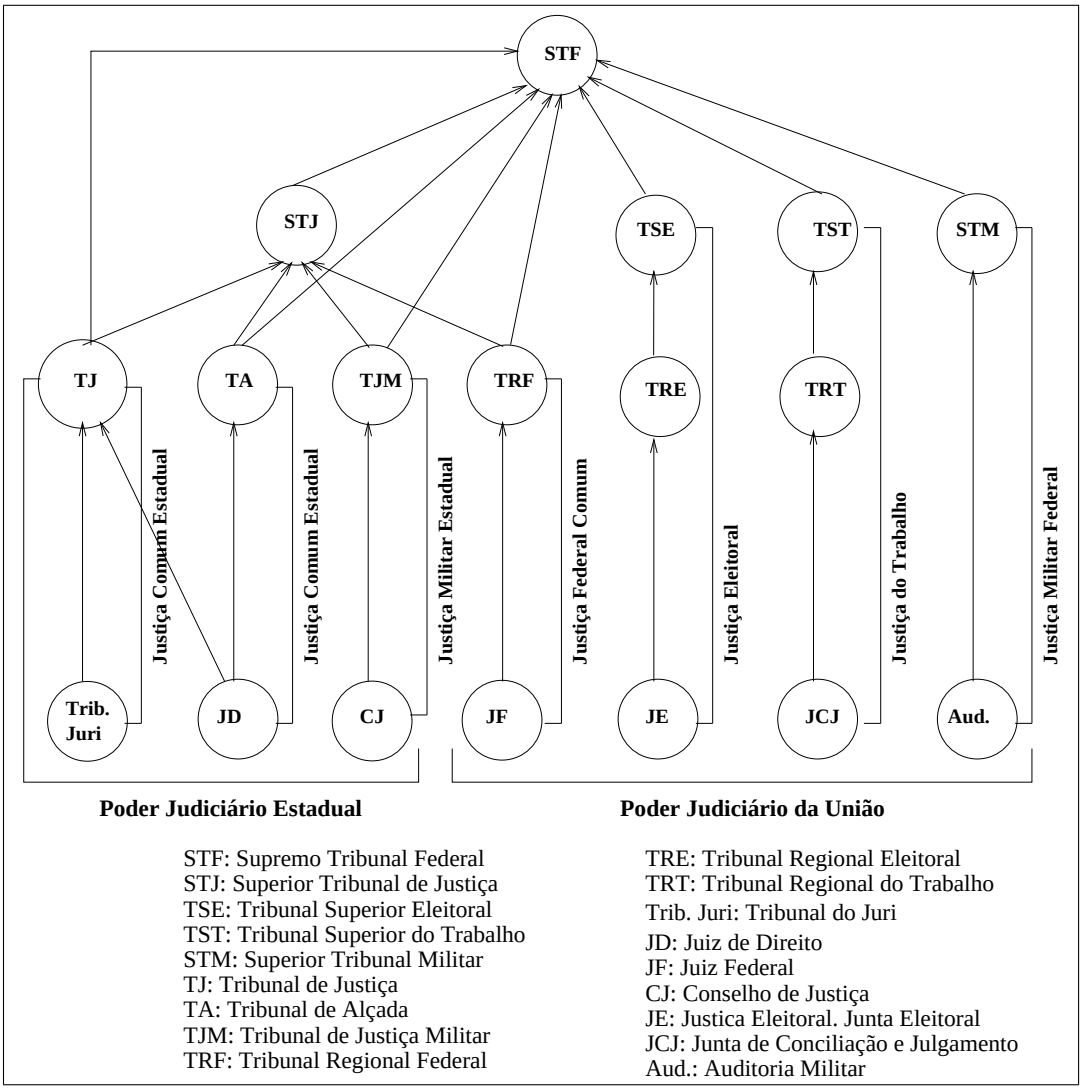

Figura A.1: Organograma do Poder Judiciário Brasileiro (de [67])

declaratórias de constitucionalidade, e ainda ações que envolvam o Presidente da República, o Vice-Presidente, membros do Congresso Nacional e do Senado Federal, os Ministros de Estado, os membros dos Tribunais Superiores, dentre outras.

Ao STJ compete processar e julgar, originariamente[2], em ações de crimes comuns, os Governadores de Estados e do Distrito Federal, os membros dos Tribunais de Contas dos Estados e do Distrito Federal, dos Tribunais Regionais Federais, Eleitorais e do Trabalho, dentre outras. Compete ainda julgar, em recurso ordinário e em recurso especial[3], decisões denegatórias dos Tribunais Regionais Federais e Tri-

[2]Instância ou local de entrada do processo.

[3]Os pedidos de recurso se dão pela não concordância, pela parte perdedora, com as decisões tomadas em instâncias inferiores.

bunais Estaduais, causas que envolvam órgãos internacionais, ou ainda causas que contrariem leis federais. O Conselho de Justiça Federal, CJF, é ligado ao STJ e tem a função de coordenação administrativa e orçamentária da Justiça Federal de primeira e segunda instâncias.

A Justiça Federal é composta, em segunda instância[4], pelos Tribunais Regionais Federais (TRF's) e pelos Juízes Federais (JF), em primeira instância. Os Tribunais Regionais Federais são compostos para atender aos processos de cinco regiões da federação brasileira. Ao TRF compete processar e julgar, originariamente, nos crimes comuns e de responsabilidade, os juízes federais da sua área de jurisdição e os membros do Ministério Público da União. Compete ainda julgar, em grau de recurso, decisões de juízes federais e estaduais, assim como os conflitos de competência. Aos juízes federais compete processar e julgar causas que envolvam a União, autarquias e empresas públicas federais, causas que envolvam órgãos internacionais ou tratado e convenção internacional, dentre outras.

Ao TSE, aos Tribunais Regionais Eleitorais (TRE), aos Juízes Eleitorais (JE) e às Juntas Eleitorais compete julgar as causas eleitorais. Ao TST, aos Tribunais Regionais do Trabalho (TRT) e aos Juízes do Trabalho compete julgar as causas trabalhistas. Ao STM e Juízes Militares compete processar e julgar os crimes militares definidos em lei. A organização da Justiça estadual é definida pela Constituição do Estado e não será aqui abordada.

Os Tribunais de instância superior e de segunda instância são órgãos colegiados, nos quais os julgamentos são decisões tomadas em consenso por turmas de juízes. Por essa razão, os documentos que representam a sentença proferida na seção de julgamento é denominada Acórdão. Tais documentos são de interesse público e por isso recebem um tratamento e uma organização com vistas a facilitar seu acesso pelos interessados.

No processo de organização dos documentos, que ocorre após o julgamento e antes da liberação ao público, os Tribunais geram uma nova seção para cada documento, denominada Seção de Indexação. Tal Seção é construída por profissionais de direito e seu objetivo básico é padronizar a linguagem utilizada para descrever o conteúdo do documento, com vistas na sua recuperabilidade. A construção da Seção de Indexação é baseada no vocabulário fornecido pelo Tesauro Jurídico da Justiça Federal, cuja descrição encontra-se na Seção 3.4.

A coleção de referência utilizada nos experimentos desse trabalho, descrita na Seção 6.1, é composta por documentos advindos dos TRF's, do STJ e do STF.

[4]Instância é o grau de jurisdição em que a ação ocorre ou hierarquia jurídica.

Apêndice B

Glossário

ANSI: *American National Standard Institute.*

CJF: Conselho de Justiça Federal.

Conceito: Unidade de pensamento formada mentalmente pela combinação de algumas ou todas as características de um objeto concreto ou abstrato, real ou imaginário [3].

Descritor: Tipo de cabeçalho que representa um termo que foi escolhido como expressão preferencial de um conceito em um tesauro [3].

Documento: Um item, impresso ou em outra forma, disponível para catalogação e indexação [3].

Cabeçalho de Assunto (*subject heading*): Palavra ou frase, ou uma combinação de palavras e frases utilizadas para indicar o conteúdo de um documento [3].

CJ: Conselho de Justiça.

Garantia Literária (*literary warrant*): Justificativa para a representação de um conceito em uma linguagem de indexação ou para a seleção como termo preferido porque sua ocorrência é freqüente na literatura [3].

Garantia do Usuário (*user warrant*): Justificativa para a representação de um conceito em uma linguagem de indexação ou para a seleção como termo preferido porque ele é freqüente em consultas de usuários em um sistema de Recuperação de Informação.

Homógrafo: Uma de duas ou mais palavras escritas da mesma forma, mas com significado e origem diferentes [3].

Indicador de Relacionamento: Uma palavra, frase, abreviatura ou símbolo que indica o relacionamento semântico entre termos. Exemplos: TE, TG, USE, UP (Usado Para), TR.

JCJ: Junta de Conciliação e Julgamento.

JD: Juiz de Direito.

JE: Justiça Eleitoral. Junta Eleitoral.

JF: Juiz Federal.

Linguagem de indexação: Um vocabulário controlado ou sistema de classificação e regras para sua aplicação [3].

Lista de Cabeçalho de Assunto (*subject heading list*): Lista alfabética de Cabeçalho de Assunto com as referências cruzadas entre termos não-preferenciais e preferenciais e termos relacionados [3].

NISO: *National Information Standards Organization.*

Palavra: Menor forma que pode ocorrer sozinha como expressão completa, elemento de significado.

Palavra-chave: Uma palavra que ocorre na linguagem natural de um documento que é considerada significante para indexação e recuperação [3].

Pós-coordenação: Combinação de descritores na etapa de pesquisa e não na de construção do tesauro ou na indexação dos documentos.

Pré-coordenação: Combinação de descritores na construção do tesauro ou na etapa de indexação.

Sinônimo: Uma palavra ou termo que possui o significado exato ou muito próximo de outra palavra ou termo.

***Stopwords*:** São termos de um idioma, considerados no âmbito de uma aplicação específica, que agregam pouca ou nenhuma semântica à sentença, tais como artigos, conjunções, preposições e conectivos.

STF: Supremo Tribunal Federal.

STJ: Superior Tribunal de Justiça.

TA: Tribunal de Alçada.

Termo: Uma ou mais palavras que designa um conceito [3].

Termo de entrada: Termo não-preferido em uma referência cruzada que leva a um descritor em um tesauro.

Termo Específico (TE): Um descritor que é subordinado a outro descritor ou a múltiplos descritores em uma hierarquia.

Termo Genérico (TG): Um descritor do qual outro descritor ou múltiplos descritores são subordinados em uma hierarquia.

Termo Preferido: Um de dois ou mais sinônimos selecionados como descritores.

Termo Não-preferido: Um de dois ou mais sinônimos ou variação léxica que serve como um termo de entrada.

Termo Relacionado (TR): Descritor que é associativamente, mas não hierarquicamente, ligado a outro descritor do tesauro [3].

Tesauro: Um Vocabulário Controlado organizado em uma ordem conhecida na qual os relacionamentos de equivalência, hierarquia e associatividade entre termos são claramente mostrados e identificados por indicadores de relacionamento padronizados, que têm que ser empregados reciprocamente [3].

TJ: Tribunal de Justiça.

TJM: Tribunal de Justiça Militar.

TRE: Tribunal Regional Eleitoral.

TRF: Tribunal Regional Federal.

TRT: Tribunal Regional do Trabalho.

TSE: Tribunal Superior Eleitoral.

TSM: Tribunal Superior Militar.

TST: Tribunal Superior do Trabalho.

UNESCO: *United Nations Educational, Scientific and Cultural Organization.*

Vocabulário: Conjunto de termos únicos de uma coleção de documentos.

Vocabulário Controlado: Subconjunto de palavras da linguagem natural. Uma lista de termos preferidos e não-preferidos produzidos pelo processo de controle de vocabulário. Tipos de vocabulários controlados incluem lista de cabeçalhos de assunto e tesauros [3].

Apêndice C

Texto das Consultas Utilizadas

1 EM ACOES ORDINARIAS DE NULIDADE DE MARCAS OU PATENTES, O INPI DEVE FIGURAR COMO REU, ASSISTENTE LITISCONSORCIAL OU ASSISTENTE SIMPLES?

2 PAGAMENTO DE HONORARIOS PARA ADVOGADOS DATIVOS E PERITOS EM ACOES SOB O PALIO DA JUSTICA GRATUITA.

3 ACAO DECLARATORIA DE INEXISTENCIA DE RELACAO JURIDICA TRIBUTARIA E EVENTUAL CONEXAO COM ACAO DE EXECUCAO FISCAL.

4 IPI - INSUMOS DESTINADOS A PRODUTOS SUJEITOS A ALIQUOTA ZERO.

5 O MINISTERIO PUBLICO E A DEFESA DO MEIO AMBIENTE EM JUIZO.

6 COMPETENCIA NOS CRIMES CONTRA A FAUNA - PRINCIPIO DA INSIGNIFICANCIA NO CRIME AMBIENTAL.

7 DIREITO ADMINISTRATIVO. ACAO DE REINTEGRACAO DE POSSE DE BEM PUBLICO PROPOSTA CONTRA PARTICULAR. DISTINCAO ENTRE POSSE NOVA E POSSE VELHA. CABIMENTO.

8 PRESCRICAO DE ACAO DE EXECUCAO CONTRA A FAZENDA PUBLICA.

9 CITACAO E INTIMACAO DA FAZENDA PUBLICA - UNIAO, ESTADO, DISTRITO FEDERAL E AUTARQUIAS.

10 CASSACAO DA APOSENTADORIA DE SERVIDOR PUBLICO APOSENTADO COMO EFEITO DA CONDENACAO.

11 CONTRIBUICOES DE INTERVENCAO NO DOMINIO ECONOMICO SESC, SENAC, SEBRAE, SENAI.

12 EXTINCAO DE ACAO DE AVERBACAO DE TEMPO DE SERVICO SEM JULGAMENTO DO MERITO POR AUSENCIA DE DOCUMENTO INDISPENSAVEL A PROPOSITURA DA ACAO INEXISTENCIA DE INICIO DE PROVA MATERIAL.

13 TRIBUTARIO. EXIGENCIA DA TAXA SELIC COMO JUROS DE MORA INCIDENTES SOBRE CREDITO TRIBUTARIO.

14 COMPETENCIA PARA PROCESSAMENTO DE CRIMES DE TRAFICO ILICITO DE ENTORPECENTES E DA JUSTICA FEDERAL OU ESTADUAL?

15 PLANO DE EQUIVALENCIA SALARIAL EM CONTRATOS DO SISTEMA FINANCEIRO DE HABITACAO.

16 ACAO ORDINARIA - CONTRA O INSS.

17 COMPETENCIA JUSTICA FEDERAL OU ESTADUAL NOS CRIMES ENVOLVENDO O SUS, NOTADAMENTE DELITO DE CONCUSSAO PRATICADO POR MEDICOS CREDENCIADOS AO SISTEMA UNICO DE SAUDE.

18 NECESSIDADE DE QUE O QUIMICO E/OU ENGENHEIRO QUIMICO TENHA REGISTRO DUPLO NO CREA E NO CRQ.

19 CORRESPONDENCIA ENTRE CONTRIBUICAO E BENEFICIO PARA EFEITOS DO CALCULO DO MENOR VALOR DO TETO EM RELACAO AO SALARIO MINIMO E NAO EM UNIDADE SALARIAL.

20 É POSSIVEL A INCORPORACAO DE METADE DO VALOR DO AUXILIO ACIDENTE A PENSAO POR MORTE NAO DERIVADA DE ACIDENTE DE TRABALHO, TENDO SIDO O EVENTO MORTE OCORRIDO APOS A REVOGACAO DA POSSIBILIDADE DE INCORPORACAO?

21 EXECUCOES FISCAIS - FALTA DE INTERESSE ECONOMICO EM EXECUCOES FISCAIS DE PEQUENO VALOR - MOTIVO PARA EXTINCAO DO PROCESSO POR FALTA DE INTERESSE PROCESSUAL.

22 COMPENSACAO DE TRIBUTOS. POSSIBILIDADE DE COMPENSACAO EM LIMINAR OU EM ANTECIPACAO DE TUTELA.

23 SERVIDOR PUBLICO CIVIL / EXONERACAO A PEDIDO / APOSENTADORIA / DIREITO ADQUIRIDO.

24 ACAO CIVIL PUBLICA, DIREITOS INDIVIDUAIS HOMOGENEOS, EXECUCAO INDIVIDUAL, REQUISITOS, INSTRUMENTO.

25 TUTELA ANTECIPADA EM ACAO RESCISORIA.

26 PRINCIPIO DA INSIGNIFICANCIA APLICADO AOS CRIMES.

27 INDENIZACAO POR DANO MORAL CRIMES DA LEI DE IMPRENSA INTERESSE PUBLICO E DIREITO A INTIMIDADE E PERSONALIDADE.

28 APROVEITAMENTO DE CREDITO RELATIVO A IPI - INSUMOS TRIBUTADOS A ALIQUOTA ZERO OU ISENTOS.

29 PRESCRICAO DA EXECUCAO CONTRA A FAZENDA PUBLICA APOS TRANSCORRIDOS MAIS DE CINCO ANOS DO TRANSITO EM JULGADO DA SENTENCA.

30 ERRO MATERIAL EM SENTENCA DADA A ACAO RESCISORIA.

Apêndice D

Exemplo de Conceitos e Relacionamentos do Tesauro da Justiça Federal

(6348) ASSISTENTE
TE1 ASSISTENTE LITISCONSORCIAL
TR ASSISTIDO

(6100) ASSISTENTE LITISCONSORCIAL
TG1 ASSISTENTE
TR LITISCONSORTE

(6096) LITISCONSORTE
TR ASSISTENTE LITISCONSORCIAL
TR LITISCONSÓRCIO

(5282) CONSELHO REGIONAL DE ENGENHARIA ARQUITETURA E AGRONOMIA
USE CONSELHO REGIONAL DE ENGENHARIA ARQUITETURA E AGRONOMIA (CREA)

(5283) CONSELHO REGIONAL DE ENGENHARIA ARQUITETURA E AGRONOMIA (CREA)
UP CONSELHO REGIONAL DE ENGENHARIA ARQUITETURA E AGRONOMIA
UP CREA
UP CREAA
TG1 CONSELHO REGIONAL
TG2 CONSELHO DE FISCALIZAÇÃO PROFISSIONAL

TR	CONSELHO FEDERAL DE ENGENHARIA ARQUITETURA E AGRONOMIA (CONFEA)
TR	CONSELHO REGIONAL DE ADMINISTRAÇÃO E ECONOMIA (CRAE)
TR	CONSELHO REGIONAL DE CONTABILIDADE (CRC)
TR	CONSELHO REGIONAL DE CORRETORES DE IMÓVEIS (CRECI)
TR	ENGENHARIA
(4704)	BEM DE DOMÍNIO PÚBLICO
USE	BEM PÚBLICO DE USO COMUM
CAT	ADM/DAJ
(2339)	BEM PÚBLICO DE USO COMUM
UP	BEM DE DOMÍNIO PÚBLICO
TG1	BEM PÚBLICO
TG2	BENS
TE1	ESTRADA
TE1	MAR
TE2	MAR TERRITORIAL
TE1	PRAÇA PÚBLICA
TE1	RIO
TE1	RUA
TR	BEM DOMINIAL
TR	BEM PÚBLICO DE USO ESPECIAL
TR	VIAÇÃO RODOVIÁRIA
CAT	ADM/DAJ

Referências Bibliográficas

[1] Jean Aitchison and Alan Gilchrist. *Thesaurus Construction.* Aslib, London, 1972.

[2] James Allan. Relevance feedback with too much data. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 337–343. ACM Press, 1995.

[3] American National Standard Institute/National Information Standards Organization - ANSI/NISO. *ANSI/NISO Z39.19 - 1993: Guidelines for the Construction, Format, and Management of Monolingual Thesauri.* ANSI/NISO, 1993.

[4] Rodrigo Tôrres Assumpção. Recuperação de documentos jurídicos baseada em um tesauro. Master's thesis, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 2001.

[5] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, Harlow, 1999.

[6] Tânia C. DAgostini Bueno, Christiane Gresse von Wangenheim, Eduardo da Silva Mattos, Hugo Cesar Hoeschl, and Ricardo M. Barcia. Jurisconsulto: Retrieval in jurisprudencial text bases using juridical terminology. In *Proceedings of the seventh International Conference on Artificial Intelligence and Law*, pages 147–155. ACM Press, 1999.

[7] Pável Calado, Berthier Ribeiro-Neto, Nivio Ziviani, Edleno Moura, and Ilmério Silva. Local versus global link information in the Web. *ACM Transactions on Information Systems (TOIS)*, 21(1):42–63, 2003.

[8] Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS)*, 19(1):1–27, 2001.

[9] Presidência da República - Casa Civil. http://www.presidencia.gov.br/ccivil_03 /Constituicao,2002.

[10] Chia-Hui Chang and Ching-Chi Hsu. Enabling concept-based relevance feedback for information retireval on the WWW. *IEEE Transactions on Knowledge and Data Engeneering*, 11(4):595–609, 1999.

[11] Conselho de Justiça Federal - CJF. http://www.cjf.gov.br, abril, 2002.

[12] Peter Clark, John Thompson, Heather Holmback, and Lisbeth Duncan. Exploiting a thesaurus-based semantic net for knowledge-based search. In *Proceedings of the 12th Conference on Innovative Applications of AI (AAAI/IAAI00)*, pages 988–995, 2000.

[13] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. Mit Press, McGraw-Hill, New York, 1997.

[14] W. Bruce Croft, Robert Cook, and Dean Wilder. Providing government information on the internet: Experiences with THOMAS. In *Proceedings of Digital Libraries Conference*, pages 19–25, 1995.

[15] René David. *Os Grandes Sistemas do Direito Contemporâneo*. Martins Fontes, Rio de Janeiro, 1996. Tradução do francês por Hermínio A. Carvalho. Título Original: Les Grands Systèmes du Droit Contemporains (Droit Comparé).

[16] Bert J. Dempsey, Robert C. Vreeland, Robert G. Sumner Jr, and Kiduk Yang. Design and empirical evaluation of searh software for legal professionals on the WWW. *Information Processing and Management*, 36:253–273, 2000.

[17] Judith P. Dick. Conceptual retrieval and case law. In *Proceedings of the first International Conference on Artificial Intelligence and Law*, pages 106–115. ACM Press, 1987.

[18] William B. Frakes and Ricardo Baeza-Yates. *Information Retrieval: Data Structures & Algoritms*. Prentice Hall PTR, New Jersey, USA, 1992.

[19] Alan Gilchrist. *The Thesaurus in Retrieval*. Aslib, London, 1971.

[20] Jane Greenberg. Automatic query expansion via lexical-semantic relationships. *Journal of the American Society for Information Science and Technology*, 52(5):402–415, 2001.

[21] Jane Greenberg. Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. *Journal of the American Society for Information Science and Technology*, 52(6):487–498, 2001.

[22] David Haines and W. Bruce Croft. Relevance feedback and inference networks. In *Proceedings of the 16th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–11. ACM Press, 1993.

[23] E. Ide. *G. Salton, editor, The SMART Retrieval System - Experiments in Automatic Document Processing*, chapter New Experiments in Relevance Feedback, pages 337–354. Prentice-Hall, Inc., Englewood Cliffs,NJ, 1971.

[24] International Standards Organization - ISO. *ISO 2788-1986: Guidelines for the Establishment and Development of Monolingual Thesauri*. ISO, 1986.

[25] Y. Jing and W. Bruce Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 146–160, New York, US, 1994.

[26] A. Kent, M. M. Berry, L. V. Luehrs Jr, and J. W. Perry. Machine literature searching VIII: Operational criteria for designing information retrieval systems. *American Documentation*, 6(2):93–101, 1955.

[27] Jaana Kristensen. Expanding end-users' query statements for free text searching with a search-aid thesaurus. *Information Processing & Management*, 29(6):733–744, 1993.

[28] Frederick Wilfrid Lancaster. *Vocabulary Control for Information Retrieval*. Information Resources Press, Arlington, Virginia, 1986.

[29] Frederick Wilfrid Lancaster. *Indexing and Abstracting in Theory and Practice*. Library Association Publishing, London, 1991.

[30] Frederick Wilfrid Lancaster and Amy J. Warner. *Information Retrieval Today*. Information Resources Press, Arlington, Virginia, 1993.

[31] David E. Losada and Alvaro Barreiro. A homogeneous framework to model relevance feedback. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 422–423. ACM Press, 2001.

[32] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–197, 1999.

[33] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. Query expansion using heterogeneous thesauri. *Information Processing & Management*, 36(2000):361–378, 2000.

[34] Charles T. Meadow. *Text Information Retrieval Systems*. Academic Press, Inc., San Diego, California, 1992.

[35] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–214. ACM Press, 1998.

[36] Juarez de Oliveira. *Constituição da República Federativa do Brasil*. Editora Saraiva, São Paulo, 1995. 12^a edição.

[37] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Los Altos, California, 1988.

[38] Yonggang Qiu and H. P. Frei. Concept based query expansion. In *Proceedings of the sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, 1993.

[39] Berthier Ribeiro-Neto and Rodrigo Tôrres Assumpção. Recuperação de documentos jurídicos baseada em um tesauro. *XVI Simpósio Brasileiro de Banco de Dados, Brasília-DF, Brasil*, outubro 2001.

[40] Berthier Ribeiro-Neto and Richard R. Muntz. A belief network model for IR. In *ACM Conference on Research and Development in Information Retrieval - SIGIR96*, pages 253–260, 1996.

[41] Berthier Ribeiro-Neto, Ilmério Silva, and Richard Muntz. *Soft Computing in Information Retrieval Techniques and Applications*, chapter Bayesian Network Models for IR, pages 259–291. Springer Verlag, 2000.

[42] J. J. Rocchio. *G. Salton, editor, The SMART Retrieval System - Experiments in Automatic Document Processing*, chapter Relevance feedback in information retrieval, pages 313–323. Prentice-Hall, Inc., Englewood Cliffs,NJ, 1971.

[43] D. E. Rose and R. K. Belew. Legal information retrieval a hybrid approach. In *Proceedings of the second International on Conference on Artificial Intelligence and Law*, pages 138–146. ACM Press, 1989.

[44] Stuart J. Russel and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, New Jersey, 1995.

[45] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Co., New York, 1983.

[46] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of theACM*, 18(11):613–620, november 1997.

[47] Gerard Salton and Chris Buckley. Term-weighting approaches in automatic retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[48] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society of Information Science*, 41(4):288–297, 1990.

[49] Ali Asghar Shiri and Crawford Revie. Thesauri on the Web: Current developments and trends. *Online Information Review*, 24(4):273–279, 2000.

[50] De Plácido e Silva. *Vocabulário Jurídico*. Editora Forense, Rio de Janeiro, 2000. Atualizadores: Nagib Slaibi Filho e Geraldo Magela Alves.

[51] Ilmério Silva. *Bayesian Networks for Information Retrieval Systems*. PhD thesis, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil, 2000. Orientador: Berthier Ribeiro-Neto.

[52] Ilmério Silva, Berthier Ribeiro-Neto, Pável Calado, Edleno Moura, and Nivio Ziviani. Link-based and content-based evidential information in a belief network model. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR2000*, pages 96–103. ACM Press, 2000.

[53] Maria de Lourdes da Silveira, Berthier Ribeiro-Neto, Rodrigo de Freitas Vale, and Rodrigo Tôrres Assumpção. Vertical searching in juridical digital libraries. In *Proceedings of the 25th European Conference on Information Retrieval Research ECIR 2003*, pages 491–501, Abril,2003.

[54] Guido Fernando Silva Soares. *Common Law: Introdução ao Direito dos EUA*. Editora Revista dos Tribunais, São Paulo, 2000. 2^a edição.

[55] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

[56] Supremo Tribunal Federal - STF. http://www.stf.gov.br, abril, 2002.

[57] Superior Tribunal de Justiça - STJ. http://www.stj.gov.br, abril, 2002.

[58] Robert S. Taylor. The process of asking questions. *American Documentation*, 13(4):391–396, 1962.

[59] Richard M. Tong, Clifford A. Reid, Gregory J. Crowe, and Peter R. Douglas. Conceptual legal document retrieval using the RUBRIC system. In *Proceedings of the first International Conference on Artificial Intelligence and Law*, pages 28–34. ACM Press, 1987.

[60] Tribunal Regional Federal - Primeira Região - TRF1. http://www.trf1.gov.br, abril, 2002.

[61] Tribunal Regional Federal Segunda Região - TRF2. http://www.trf2.gov.br, abril, 2002.

[62] Tribunal Regional Federal Terceira Região - TRF3. http://www.trf3.gov.br, abril, 2002.

[63] Tribunal Regional Federal Quarta Região - TRF4. http://www.trf4.gov.br, abril, 2002.

[64] Tribunal Regional Federal Quinta Região - TRF5. http://www.trf5.gov.br, abril, 2002.

[65] Howard R. Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *Transaction on Information Systems*, 9(3):187–222, 1991.

[66] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, Lodon, 1979.

[67] Carlos Mário da Silva Velloso. Do poder judiciário: Organização e competência. *Revista de Direito Administrativo*, 200(Abr/Jun):1–19, 1995.

[68] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69. Springer-Verlag New York, Inc., 1994.

[69] Ellen M. Voorhees and Donna Harman. Overview of the fifth Text REtrieval Conference. In *Proceedings of the fifth Text REtrieval Conference (TREC-5)*, pages 1–28. National Institute of Standards and Technology, Gaithersburg, MD 20899, 1996.

[70] Ellen M. Voorhees and Donna Harman. Overview of the seventh Text REtrieval Conference. In *Proceedings of the seventh Text REtrieval Conference (TREC-8)*, pages 1–23. National Institute of Standards and Technology, Gaithersburg, MD, 1999.

[71] Rosina Weber. Intelligent jurisprudence research: A new concept. In *Proceedings of the seventh International Conference on Artificial Intelligence and Law*, pages 164–172. ACM Press, 1999.

[72] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11. ACM Press, 1996.

[73] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS)*, 18(1):79–112, 2000.