Universidade Estadual de Campinas

Faculdade de Engenharia Elétrica e de Computação

Maria Angelica de Andrade Leite

Modelo *Fuzzy* para Recuperação de Informação Utilizando Múltiplas Ontologias Relacionadas

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica. Área de concentração: Engenharia de Computação.

Orientador: Ivan Luiz Marques Ricarte

Campinas, SP
2009

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

Leite, Maria Angelica de
L536m Modelo fuzzy para recuperação de informação utilizando múltiplas ontologias relacionadas / Maria Angelica de Andrade Leite. --Campinas, SP: [s.n.], 2009.

Orientador: Ivan Luiz Marques Ricarte.
Tese de Doutorado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Recuperação de Informação. 2. Representação do conhecimento. 3. Ontologia. 4. Sistemas difusos.
I. Ricarte, Ivan Luiz Marques. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: Fuzzy information retrieval model using multiple related ontologies
Palavras-chave em Inglês: Fuzzy Information retrieval, Knowledge representation, Query expansion, Ontology
Área de concentração: Engenharia da Computação
Titulação: Doutor em Engenharia Elétrica
Banca Examinadora: Frederico Luiz Gonçalves de Freitas, Kleber Xavier Sampaio de Souza, Léo Pini Magalhães, Fernando Antônio Campos Gomide
Data da defesa: 13/03/2009
Programa de Pós Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE DOUTORADO

Candidata: Maria Angelica de Andrade Leite

Data da Defesa: 13 de março de 2009

Título da Tese: "Modelo Fuzzy para Recuperação de Informação Utilizando Múltiplas Ontologias Relacionadas"

Prof. Dr. Ivan Luiz Marques Ricarte (Presidente): _____

Prof. Dr. Frederico Luiz Gonçalves de Freitas: _____

Prof. Dr. Kleber Xavier Sampaio de Souza: _____

Prof. Dr. Léo Pini Magalhães: _____

Prof. Dr. Fernando Antônio Campos Gomide: _____

Resumo

Com a crescente popularidade da *World Wide Web* mais pessoas têm acesso à informação cujo volume vem expandindo ao longo do tempo. A área de recuperação de informação ganhou um novo desafio visando buscar os recursos pelo significado da informação neles contida. Uma forma de recuperar a informação, pelo seu significado, é pelo uso de uma base de conhecimento que modela os conceitos de um domínio e seus relacionamentos. Atualmente, ontologias têm sido utilizadas para modelar bases de conhecimento. Para tratar com a imprecisão e a incerteza, presentes no conhecimento e no processo de recuperação de informação, são empregadas técnicas da teoria de conjuntos *fuzzy*. Trabalhos precedentes codificam a base de conhecimento utilizando apenas uma ontologia. Entretanto, uma coleção de documentos pode tratar temas pertencentes a domínios diferentes, expressos por ontologias distintas, que podem estar relacionados. Neste trabalho, uma forma de organização e representação do conhecimento em múltiplas ontologias relacionadas foi investigada e um novo método de expansão de consulta foi desenvolvido. A organização do conhecimento e o método de expansão de consulta foram integrados no modelo *fuzzy* para recuperação de informação utilizando múltiplas ontologias relacionadas. O desempenho do modelo foi comparado com outro modelo *fuzzy* para recuperação de informação e com a máquina de busca Lucene do projeto Apache. Em ambos os casos o modelo proposto apresentou uma melhora nas medidas de precisão e cobertura.

Palavras-chave: Recuperação de informação *fuzzy*, Representação do conhecimento, Expansão da consulta, Ontologia.

Abstract

With the World Wide Web popularity growth, more people has access to information and this information volume is expanding over the time. The information retrieval area has a new challenge intending to search information resources by their meaning. A way to retrieve information, by its meaning, is by using a knowledge base that encodes the domain concepts and their relationships. Nowadays ontologies are being used to model knowledge bases. To deal with imprecison and uncertainty present in the knowledge and in the information retrieval process, fuzzy set theory techniques are employed. Preceding works encode a knowledge base using just one ontology. However a document collection can deal with different domain themes, expressed by distinct ontologies, that can be related. In this work a way of knowledge organization and representation, using multiple related ontologies, was investigated and a new method of query expansion was developed. The knowledge organization and the query expansion method were integrated in the fuzzy model for information retrieval based on mutiple related ontologies. The model performance was compared with another fuzzy-based approach for information retrieval and with the Apache Lucene search engine. In both cases the proposed model improves the precision and recall measures.

Keywords: Fuzzy information retrieval, Knowledge representation, Query expansion, Ontology.

Agradecimentos

Ao meu orientador, Prof. Dr. Ivan Luiz Marques Ricarte, pela sua disponibilidade, pelas discussões teóricas e técnicas e pelas valiosas sugestões no encaminhamento de meu trabalho.

À Faculdade de Engenharia Elétrica e de Computação, que me acolheu como aluna, e a seus professores pela oportunidade de conviver com pessoas sábias que me ajudaram a abrir a mente para novas possibilidades e desafios de aprendizado.

À EMBRAPA que acreditou em meu potencial e investiu na minha formação.

Ao meu conselheiro acadêmico, Dr. Kleber Xavier Sampaio de Souza, por sua atenção e pelo acompanhamento constante das minhas atividades.

Aos meus colegas da Embrapa Informática Agropecuária pelas discussões técnicas e, em especial, à Leila Maria Lenk que atuou como especialista de domínio na montagem do experimento realizado na tese.

Aos meus pais, Francisco e Licínia, pelo seu amor e lições de vida. A eles eu agradeço os valores ensinados e o incentivo e apoio em todas as minhas iniciativas para aprimorar a minha educação.

Ao meu marido, José Mário, e a meus filhos, Bárbara e Mateus, que foram as âncoras que me deram paz e equilíbrio e que me ajudaram a manter o meu curso de vida e o meu foco no trabalho nos inúmeros imprevistos ocorridos no período de meu doutorado.

Ao meu irmão, Francisco Eugênio, com quem pude compartilhar e dividir decisões difíceis.

Às minhas tias, Maria José e Lygia, pelo suporte emocional e pelo carinho e dedicação a meus pais e irmã nas horas em que não pude estar presente. À minha tia Leda pelas conversas e ensinamentos sobre a nossa jornada neste mundo.

Aos meus sogros, Tereza e José Mário, pelo suporte no dia a dia para que eu pudesse me dedicar a estudar.

Aos meus primos pelo seu companheirismo, compreensão e apoio.

*Dedico esta tese à memória de meus pais, **Francisco** e **Licínia** e à memória de minha irmã, **Maria de Fátima**, que passou por esta vida e que, por suas condições frágeis, não pôde usufruir das mesmas oportunidades que eu.*
*Dedico também ao meu marido, **José Mário**, e a meus filhos, **Bárbara** e **Mateus**, que representam o meu presente e futuro e que estiveram ao meu lado em todos os momentos.*

Sumário

Lista de Figuras

Lista de Tabelas

Glossário

API – Application Program Interface

BINAGRI – Biblioteca Nacional de Agricultura

CLEF – Cross-Language Evaluation Forum

DAML – DARPA Agent Markup Language

Embrapa – Empresa Brasileira de Pesquisa Agropecuária

FaCT – Fast Classification of Terminologies

FAO – Food and Agriculture Organization

FLogic – Frame Logic

FROM – Fuzzy Relational Ontological Model

HTML – HyperText Markup Language

IBGE – Instituto Brasileiro de Geografia e Estatística

JTP – Java Theorem Prover

KIF – Knowledge Interchange Format

LSI – Latent Semantic Indexing

MVC – Model View Controller

OIL – Ontology Inference Layer

OWL – Web Ontology Language

SAX – Simple API for XML

Thesagro – Thesaurus Agrícola Nacional

TREC – Text REtrieval Conference

W3C – World Wide Web Consortium

WWW – World Wide Web

Trabalhos Publicados Pela Autora

1. M. A. A. Leite, I. L. M. Ricarte. “Fuzzy Information Retrieval Model Based on Multiple Related Ontologies ”. *20th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI’08), Dayton, Ohio, USA, pg. 309-316, November 2008.

2. M. A. A. Leite, I. L. M. Ricarte. “Document Retrieval Using Fuzzy Related Geographic Ontologies”. *Second International Workshop on Geographic Information Retrieval* (GIR’08) in ACM 17th Conference on Information and Knowledge Management (CIKM 2008), Napa Valley, California, USA, pg. 47-54, October 2008.

3. M. A. A. Leite, I. L. M. Ricarte. “Using Multiple Related Ontologies in a Fuzzy Information Retrieval Model”. *Third Workshop on Ontologies and Their Applications* (WONTO’2008) in 19th Brazilian Symposium on Artificial Intelligence (SBIA), Salvador, Bahia, Brasil, October 2008.

4. M. A. A. Leite, I. L. M. Ricarte. “A Framework for Information Retrieval Based on Fuzzy Relations and Multiple Ontologies”. *11th Ibero-American conference on AI: Advances in Artificial Intelligence*, LNCS 5290/2008, Lisbon, Portugal, pg. 292-301, October 2008.

Capítulo 1

Introdução

A era da informação produziu uma profusão de informação disponível para qualquer pessoa que queira acessá-la. Com a crescente popularidade da WWW (*World Wide Web*) e dos serviços *online* mais pessoas têm acesso à informação do que em qualquer época anterior. Ao longo do tempo o volume de informação vem crescendo. Grande parte da informação, que antes era impressa e que podia ser pesquisada em uma biblioteca, agora encontra-se digitalizada em arquivos e armazenada em bibliotecas digitais podendo ser acessada via WWW. Os arquivos digitais são recursos de informação e podem conter documentos textuais, imagem, vídeo e som. A variedade de tipos de recursos e sua enorme quantidade disponível na WWW trouxeram novos desafios para a área de recuperação de informação.

Existem diferentes linhas de pesquisa buscando aprimoramento, eficiência e eficácia em todos os passos envolvidos no processo de recuperação da informação na WWW. Uma tendência atual é a de realizar a busca de recursos pelo significado da informação. A busca da informação pelo seu significado é denominada busca semântica [37, 90, 134]. Um recurso pode ser recuperado não só pelo seu conteúdo léxico, mas também pelo seu significado, pelo conhecimento associado ao recurso, ou seja, seu conteúdo semântico. Esta é uma preocupação recente e atinge tanto sistemas de recuperação de informação de uma coleção de documentos específica como a WWW. No que se refere à WWW, o objetivo é tornar as páginas da Web processáveis por agentes inteligentes para que a informação nelas contida possa ser manipulada por computadores e não apenas entendidas pelos seres humanos. Esta preocupação culminou no que está sendo chamado de Web Semântica [4, 11, 20] a qual vem defendendo o uso de estruturas conceituais, denominadas ontologias [18, 52], para organizar o conhecimento e expressar o significado semântico.

1.1 Motivação e Contexto do Trabalho

A recuperação de recursos de informação pelo seu conteúdo semântico constitui a motivação para este trabalho de doutorado. Neste trabalho será estudado o problema de recuperação de informação em uma área de conhecimento específica considerando recursos de informação que apresentam uma descrição textual para sua caracterização. Assim são considerados documentos textuais em si e mesmo arquivos de imagem, som e vídeo desde que possuam informação textual associada que possa ser utilizada para caracterização e recuperação. Em todos estes casos a informação a ser recuperada é uma informação textual representada por documentos. Neste trabalho, os recursos de informação serão considerados como documentos.

A literatura mostra que a busca utilizando os conceitos organizados em uma base de conhecimento permite o acesso aos documentos pelo seu significado melhorando a qualidade e relevância da informação recuperada como, por exemplo, em documentos textuais da WWW [64], catálogos *online* [61] e em sistemas de seleção de informação [66].

Um sistema de recuperação de informação armazena e indexa documentos, de forma que quando os usuários expressam sua necessidade de busca o sistema recupera os documentos relacionados associando uma pontuação a cada um deles. Quanto maior a pontuação, maior a relevância do documento para a consulta. Em geral, os sistemas de recuperação de informação não apresentam o comportamento ideal, ou seja, o usuário recebe uma lista muito grande de documentos, como resposta, e nem sempre os documentos relevantes à consulta são os primeiros da lista de documentos. Desta maneira os usuários devem despender um tempo razoável até encontrar aqueles que são realmente relevantes. Além disto, os documentos são recuperados apenas quando eles contêm os termos especificados na consulta. Entretanto, este enfoque vai negligenciar outros documentos que também possam ser relevantes mas que não contêm os termos especificados na consulta. Ao considerar um domínio de conhecimento específico este problema pode ser superado pela incorporação de uma base de conhecimento no processo de recuperação de informação. A base de conhecimento vai modelar os conceitos de um domínio e seus relacionamentos. Ao utilizar uma base de conhecimento o objetivo é utilizar o conhecimento expresso na base para melhorar a qualidade dos documentos recuperados trazendo mais documentos associados à consulta (melhoria da taxa de cobertura) e apresentando estes documentos numa ordem onde os documentos do topo da lista de documentos sejam os mais relevantes à consulta (melhoria da taxa de precisão). As bases de conhecimento podem ser desenvolvidas manualmente por especialistas de domínio ou construídas automaticamente a partir da coleção de documentos como, por exemplo, bases compostas por ontologias *fuzzy* [72, 127] ou ontologias geográficas [17].

Atualmente técnicas de inteligência computacional têm sido utilizadas para lidar com o conhecimento e tratar diversos aspectos envolvidos com recuperação de informação [15] tais como sistema do tipo pergunta e resposta, sistemas de recomendação de produtos [78], recuperação de informação

multimídia e avaliação da qualidade de informação. Entre as técnicas de inteligência computacional a teoria de conjuntos *fuzzy* surgiu como uma forma de representar, tratar e raciocinar com incertezas [99, 100].

No processo de recuperação de informação, a expansão de consulta consiste em adicionar novos termos semanticamente relacionados com os termos presentes na consulta inicial em função do conhecimento contido em uma base de conhecimento. As bases de conhecimento são modeladas por meio de estruturas conceituais que se encarregam de capturar os conceitos relativos a um determinado domínio bem como as relações entre eles. Entre os enfoques para condução da expansão de consulta um dos mais recentes consiste em utilizar ontologias como estruturas conceituais para modelar o conhecimento permitindo inferir os novos termos a serem adicionados à consulta [12]. As ontologias incluem conceitos organizados em taxonomias, relacionamentos entre os conceitos e propriedades que descrevem estes conceitos. Ao longo do tempo a representação do conhecimento utilizando ontologias vem ganhando força e muitas ontologias tem aparecido refletindo diversos domínios como, por exemplo, o domínio da genética [81], da medicina [26] e da classificação de insetos [133]. Particularmente ontologias *fuzzy* [1, 95] têm sido construídas para modelar a incerteza presente no conhecimento do domínio.

Em geral, os sistemas de recuperação de informação utilizam apenas uma ontologia para modelar o conhecimento e compor a base de conhecimento. Mas o conhecimento que indexa uma coleção de documentos pode ser expresso em múltiplos domínios distintos onde cada um é representado por uma ontologia. A construção de uma ontologia para cada domínio de conhecimento possibilita facilidade de manutenção e reuso. Uma mesma ontologia pode ser reusada em diversas aplicações. Cada ontologia pode ser desenvolvida e evoluída de forma independente por especialistas do domínio. Quando se deseja utilizar diferentes ontologias para um fim comum elas devem ser combinadas. Na literatura encontra-se trabalhos onde esta combinação é feita integrando as ontologias [68, 105] para formar uma nova ontologia numa tarefa de *merge* [63] ou de alinhamento [118]. Neste caso, as ontologias não mantêm sua independência e têm sua estrutura modificada.

Esta tese propõe uma organização do conhecimento onde as ontologias são relacionadas mas mantêm a sua independência não tendo sua estrutura alterada. Os relacionamentos entre as ontologias também são mantidos em uma estrutura independente. No caso de haver alterações nas ontologias, os grupos de especialistas podem trabalhar em paralelo para posteriormente verificar se há necessidade de relacionar estas estruturas trabalhando apenas nos relacionamentos entre elas. As ontologias e os relacionamentos entre entre os conceitos das ontologias formam uma base de conhecimento composta de múltiplas ontologias relacionadas.

A base de conhecimento com múltiplas ontologias relacionadas é utilizada em um modelo *fuzzy* para recuperação de informação [76]. Baseado no conhecimento expresso nas ontologias o modelo

executa a expansão automática da consulta. Os documentos são indexados pelos conceitos na base de conhecimento permitindo a recuperação pela informação semântica existente na base. Os documentos não necessitam estar indexados pelos conceitos de cada ontologia. Dada uma consulta com conceitos de um determinado domínio, novos documentos semanticamente relacionados e indexados pelos conceitos das demais ontologias podem ser recuperados baseado nos relacionamentos entre as ontologias.

Para testar o modelo proposto foi criada uma aplicação na área de recuperação de informação geográfica no domínio da agrometeorologia no Brasil [75]. As máquinas de busca convencionais tratam os nomes de lugares nas consultas da mesma forma como as outras palavras-chaves e vão retornar documentos que incluem o nome especificado. Em algumas aplicações este enfoque pode ser adequado mas existem situações onde seria interessante recuperar, também, documentos que apresentam características geográficas parecidas com o lugar especificado. Em alguns contextos uma região é caracterizada por aspectos que ela apresenta e documentos associados a estes aspectos referem-se a este lugar mesmo que o seu nome não esteja presente. Por exemplo, a região Nordeste no Brasil apresenta o clima semi-árido e em muitas situações ela é referenciada apenas como região semi-árida ao invés de seu nome geográfico. Desta forma, documentos contendo informações sobre o clima semi-árido podem ser interessantes de recuperar para uma consulta contendo o termo "Região Nordeste". Além disto, em muitos casos, a informação geográfica requerida pode não estar presente nos documentos mas pode ser representada pela indicação da entidade geográfica mais genérica ou mais específica.

O modelo *fuzzy* para recuperação de informação baseado em múltiplas ontologias relacionadas foi testado utilizando ontologias *fuzzy* [73] e *crisp* [74]. Os resultados obtidos com o modelo proposto foram comparados com os resultados obtidos utilizando apenas os termos iniciais da consulta e com o modelo de rede de conceitos *fuzzy* [19, 56]. Além disto o método de expansão da consulta também foi testado considerando os documentos indexados pela máquina de busca do Apache Lucene [6].

1.2 Objetivo do Trabalho

O objetivo desta tese foi estudar, adaptar e estender as técnicas existentes relativas à tratamento, organização e recuperação da informação para desenvolver um modelo para organizar uma área de conhecimento permitindo a recuperação de informação dos documentos semanticamente associados. Para atingir este objetivo investigou-se uma forma de organização do conhecimento em estruturas conceituais múltiplas relacionadas, representadas por ontologias, e desenvolveu-se um método de expansão da consulta baseado no conhecimento estabelecido pelas estruturas conceituais.

Acredita-se que o modelo proposto possa ser empregado em qualquer área que possua seu conhecimento organizado em domínios distintos através de estruturas conceituais relacionadas. Os docu-

mentos estarão associados aos conceitos presentes nas estruturas conceituais. No caso deste trabalho, pretende-se que ele possa ser utilizado na recuperação de documentos sobre a agropecuária brasileira disponibilizados pela Embrapa (Empresa Brasileira de Pesquisa Agropecuária) [35].

1.3 Organização do Documento

Este trabalho está organizado da seguinte forma: neste capítulo foi apresentada uma visão geral do problema investigado, o contexto no qual o tema está inserido, a motivação e o objetivo a ser atingido. No Capítulo 2 são discutidos os principais tópicos e definições envolvidos com recuperação de informação propiciando o entendimento das discussões nos capítulos seguintes; o Capítulo 3 apresenta a revisão bibliográfica sobre trabalhos que oferecem opções de recuperação de informação semântica utilizando estruturas conceituais para organizar o conhecimento, os ganhos obtidos e os trabalhos futuros para melhoria desta área; o Capítulo 4 apresenta a proposta do modelo de recuperação de informação semântica investigado no doutorado; o Capítulo 5 apresenta os resultados experimentais resultantes da utilização do modelo proposto em uma aplicação de recuperação de informação geográfica na área de agrometeorologia e o Capítulo 6 apresenta as conclusões. O Apêndice A apresenta a preparação dos dados experimentais utilizados nos testes para validação do modelo proposto.

Capítulo 2

Recuperação de Informação

A recuperação de informação é uma área da ciência da computação que considera métodos e técnicas para representação, armazenamento, organização e recuperação de recursos de informação. Seu principal objetivo é facilitar o acesso aos recursos de informação relevantes às necessidades do usuário. Com o advento da WWW e a profusão de uma imensa quantidade de informação disponível, a área de recuperação de informação ganhou novos desafios. O escopo da recuperação de informação pode considerar os recursos de uma área de conhecimento específica como, por exemplo, agricultura, artes, leis e saúde até os recursos de toda a WWW. Neste trabalho será estudado o problema de recuperação de informação em uma área de conhecimento específica considerando recursos de informação que apresentam uma descrição textual para sua caracterização, representada por documentos. Assim são considerados documentos textuais em si e mesmo arquivos de imagem, som e vídeo desde que possuam informação textual associada que possa ser utilizada para caracterização e recuperação. Desta forma, no restante deste trabalho, os recursos de informação passam a ser considerados como documentos.

2.1 Visão lógica dos documentos

Um documento refere-se a um texto constituído por palavras organizadas segundo regras da linguagem natural escrita. Os documentos carregam informação que está associada a seu conteúdo. O significado da informação descrita em um documento refere-se à sua semântica. Na linguagem escrita algumas palavras carregam mais significado do que outras. Assim, em geral, os sistemas clássicos de recuperação de informação representam os documentos por meio das palavras que expressam melhor o seu conteúdo. Este conjunto de palavras é extraído dos documentos e forma uma estrutura denominada índice. Cada palavra presente no índice é considerada um termo de indexação. Um termo de indexação constitui em uma palavra cujo significado ajuda a lembrar os temas principais discutidos

no documento. Os termos de indexação constituem a visão lógica do documento.

Dado um conjunto de termos de indexação de um documento deve-se notar que alguns deles descrevem melhor o conteúdo do documento em questão do que os outros. Em uma coleção de documentos, um termo que apareça em todos os documentos da coleção não carrega muita informação pois, quando um usuário utiliza este termo para a busca, todos os documentos da base são retornados. Por outro lado um termo que apareça em apenas alguns dos documentos ajuda a reduzir o número de documentos retornados em uma busca feita pelo usuário. Este termo é um bom índice da coleção. Este efeito é capturado associando-se um grau de relevância para cada termo dado por um peso a ele associado. O cálculo do peso de cada termo não é uma tarefa trivial e, em geral, está associado à freqüência do termo no documento e na coleção como um todo. Seja k_i um termo de indexação, d_j um documento e $w_{i,j}$ o peso associado ao par (k_i, d_j), onde $w_{i,j}$ é um número real e $w_{i,j} \geq 0$. Este peso quantifica a importância do termo de indexação em descrever o conteúdo do documento.

2.2 Busca de informação

Quando um usuário busca uma informação em um sistema de recuperação de informação, a forma mais usual dele executar esta tarefa é expressando sua necessidade de informação em uma expressão de busca na linguagem provida pelo sistema. Isto significa especificar um conjunto de palavras que traduza sua necessidade de informação. Para conseguir este objetivo os usuários necessitam ter conhecimento do tema de interesse, em particular do vocabulário do domínio.

O processo básico de recuperação de informação [108] pode ser visto na Fig. 2.1. Primeiramente, antes mesmo do processo de recuperação de informação ser iniciado, é preciso possuir a base dos documentos que se deseja disponibilizar e construir um arquivo de índices para esta base. A construção deste arquivo de índices vai depender do modelo escolhido para representar tanto os documentos como as consultas a serem executadas sobre eles. Várias estruturas podem ser utilizadas sendo que a mais comum é o arquivo invertido. Uma vez que a base de documentos está indexada começa o processo de recuperação de informação propriamente dito.

Inicialmente o usuário deve especificar sua necessidade para o sistema, utilizando a interface disponível, por meio de uma expressão de busca que, em geral, é composta de palavras-chave. Antes de submeter a consulta ao sistema pode-se adotar um mecanismo de confirmação da busca a ser realizado junto ao usuário. Ao final do processo de captura e transformação da consulta tem-se uma representação da expressão de busca traduzida nos objetos de busca, num formato compreensível pela máquina em função da estratégia de busca utilizada no modelo. Por meio destes objetos de busca é realizada uma consulta no arquivo de índices e os documentos que satisfizerem às condições estabelecidas são recuperados. Antes de apresentar o resultado ao usuário os documentos são ordenados de

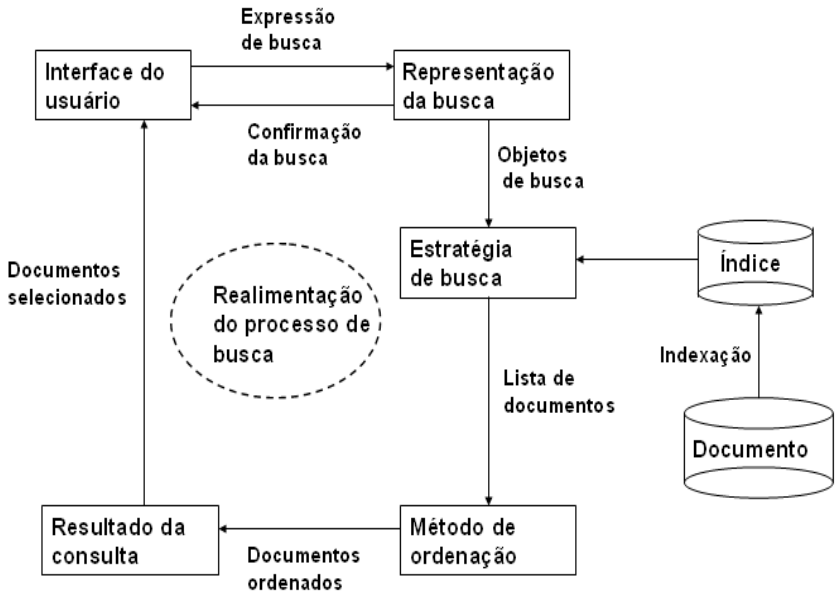

Fig. 2.1: Modelo básico de recuperação de informação. Adaptado de [108]

acordo com o critério de relevância adotado. Na apresentação de resultados o usuário vai examinar o resultado obtido e pode iniciar um ciclo de realimentação da busca.

A representação de documentos utilizada para gerar o arquivo de índices, os objetos de busca, a estratégia de busca e o esquema de ordenação dos resultados estão intimamente ligados e dependem do modelo escolhido para implementá-los.

2.3 Modelos de Recuperação de Informação

Um sistema de recuperação de informação armazena documentos juntamente com dados de indexação destes documentos de tal forma que, quando um usuário expressa sua necessidade, o sistema seleciona os documentos relacionados atribuindo um valor para cada um. Este valor depende do modelo de recuperação de informação utilizado e indica a relevância do documento com relação à consulta. Quanto mais alto o valor maior é a relevância do documento. Os documentos são apresentados ordenados por sua relevância.

Um sistema de recuperação de informação é definido pela quádrupla $[D, Q, F, R(q_i, d_j)]$ [9] onde:

1. D é um conjunto das representações (visão lógica) dos documentos, formando o conjunto de documentos da coleção.

2. Q é um conjunto das representações das consultas dos usuários.

3. F é o *framework* para modelar as representações dos documentos, consultas e seus relacionamentos.

4. $R(q_i, d_j)$ é uma função que associa a cada consulta $q_i \in Q$ e a cada documento $d_j \in D$ um número real que representa relevância do documento para a consulta.

Existem três modelos clássicos de representação de documentos que são conhecidos por modelo booleano, modelo vetorial e o modelo probabilístico que serão apresentados nas próximas seções.

2.3.1 Modelo Booleano

O modelo booleano é um modelo de recuperação de informação baseado na teoria dos conjuntos e na álgebra booleana. Devido à sua simplicidade o modelo booleano foi adotado por muitos sistemas de recuperação de informação bibliográficos. Uma consulta é uma expressão booleana convencional composta de termos de indexação e dos conectivos AND, OR ou NOT. Por meio da utilização de equivalências lógicas como as leis de De Morgan, a lei da eliminação da dupla negação e a lei distributiva é possível reduzir qualquer expressão booleana à forma conjuntiva normal que é a soma do produto de termos ou à forma disjuntiva normal que é o produto da soma de termos [21]. Por exemplo, a consulta $q = k_a \wedge (k_b \vee \neg k_c)$ pode ser escrita na forma disjuntiva normal como $q_{fdn} = (1,\, 1,\, 1) \vee (1, 1, 0) \vee (1, 0, 0)$ podendo ser representada pelo conjunto $Q_{fdn} = \{(1,\, 1,\, 1),\ (1,\, 1,\, 0),\ (1,\, 0,\, 0)\}$. Neste conjunto cada um dos componentes é um vetor binário associado à tupla $(k_a,\, k_b,\, k_c)$. Estes vetores binários, denotados por $\overrightarrow{q}_{cc}$, são chamados de componentes conjuntivos da consulta q. A transformação da consulta q para a forma disjuntiva normal é mostrada a seguir:

$$k_a \wedge (k_b \vee \neg k_c) \Rightarrow \begin{array}{c} k_a \wedge k_b \\ \vee \\ k_a \wedge \neg k_c \end{array} \begin{cases} k_a \wedge k_b \wedge k_c \quad (1\,1\,1) \\ k_a \wedge k_b \wedge \neg k_c \quad (1\,1\,0) \end{cases} \begin{cases} k_a \wedge k_b \wedge \neg k_c \quad (1\,1\,0) \\ k_a \wedge \neg k_b \wedge \neg k_c \quad (1\,0\,0) \end{cases}$$

Definição 2.1 *Seja t o número total de termos de indexação do sistema de recuperação e k_i um termo de indexação genérico. Seja $K = \{k_1, \cdots, k_t\}$ o conjunto de todos os termos de indexação. Um peso $w_{i,j} \geq 0$ é associado a cada termo de indexação k_i do documento d_j. Um termo de indexação que não está presente no documento possui $w_{i,j} = 0$. A cada documento d_j é associado um vetor de termos de indexação representados por $\overrightarrow{d}_j = (w_{1,j},\, w_{1,j},\, \cdots, w_{t,j})$. Neste contexto define-se g_i como uma função que retorna o peso associado ao termo de indexação k_i em qualquer vetor t-dimensional, ou seja, $g_i\left(\overrightarrow{d}_j\right) = w_{i,j}$.*

No modelo booleano, o peso w_{ij} dos termos de indexação assumem valores binários, isto é, $w_{ij} \in \{0,\ 1\}$. Seja Q_{fdn} o conjunto com os componentes conjuntivos, $\overrightarrow{q}_{cc}$, da consulta q. A relevância de um documento d_j para a consulta q é definida pela expressão 2.1. Como o conjunto Q_{fdn} representa a forma disjuntiva normal $q_{fdn} = (1,\ 1,\ 1) \vee (1,\ 1,\ 0) \vee (1,\ 0,\ 0)$, que é uma soma de produtos, então se um documento for relevante para um dos componentes conjuntivos $\overrightarrow{q}_{cc} \in Q_{fdn}$ então ele é relevante para a consulta q. Neste caso, para todo $k_i \in K$, se o peso de k_i em d_j é igual ao peso de k_i em $\overrightarrow{q}_{cc}$ então o valor de r$(d_j,\ q) = 1$ e o documento d_j é relevante para a consulta do usuário.

$$\text{r}(d_j,\ q) = \begin{cases} 1 & \text{se } \exists \overrightarrow{q}_{cc} \mid (\overrightarrow{q}_{cc} \in Q_{fdn}) \wedge (\forall k_i,\ g_i(\overrightarrow{d}_j) = g_i(\overrightarrow{q}_{cc})) \\ 0 & \text{caso contrário} \end{cases} \tag{2.1}$$

O modelo booleano prediz se um documento é relevante, r$(d_j, q) = 1$, ou não relevante, r$(d_j, q) = 0$. A relação de relevância no modelo booleano segue a definição de relação matemática clássica que descreve apenas a presença (1) ou ausência (0) de associação entre uma consulta e um documento. Não existe a noção da relevância parcial com relação à consulta. Assim não há como estabelecer uma ordem de relevância entre os documentos. As principais vantagens do modelo booleano são a transparência do seu formalismo e sua simplicidade. A principal desvantagem é que a pertinência binária pode resultar na recuperação de um número muito alto ou muito pequeno de documentos apresentando um resultado aquém do esperado.

Para lidar com a limitação de pertinência binária, imposta pelo modelo booleano, surgiu o modelo *fuzzy*. A estrutura básica para representação formal dos relacionamentos neste modelo é a relação *fuzzy* a qual estende o conceito matemático de relação. Enquanto as relações matemáticas clássicas descrevem apenas a presença (1) ou ausência (0) de associação entre elementos de dois conjuntos, as relações *fuzzy* permitem expressar o grau da relação. Considere por exemplo uma relação *fuzzy* $R : D \times T \rightarrow [0,\ 1]$ definida sobre um conjunto de documentos D e um conjunto de termos T. Para cada elemento $d \in D$ e cada termo $t \in T$ a relação $R(d,\ t)$ pode ser interpretada como o grau de relevância do documento d para o termo t. O modelo *fuzzy* estende o modelo booleano herdando deste o formalismo simples, baseado em relações, mas ao mesmo tempo permitindo estabelecer a relevância parcial entre os documentos e os termos de uma consulta.

2.3.2 Modelo Vetorial

O modelo vetorial [85] considera pesos não binários para os termos tanto na indexação dos documentos como nas consultas. Desta forma ele permite estabelecer relevância parcial por meio do cálculo do grau de relevância entre um documento armazenado no sistema e a consulta do usuário. Por meio da ordenação dos documentos recuperados pelo seu grau de relevância é possível apresentar primeiramente os documentos que são mais similares à consulta melhorando a qualidade da

informação recuperada.

Para o modelo vetorial, o peso w_{ij}, associado ao par (k_i, d_j), é um número real positivo. Além disto os termos da consulta também possuem pesos. Seja w_{ij} o peso associado ao par (k_i, d_j), onde $w_{ij} \geq 0$. O vetor da consulta $\overrightarrow{q}$ é definido como $\overrightarrow{q} = (w_{1q}, w_{2q}, \cdots w_{tq})$ onde t é o número total de termos de indexação do sistema. O vetor $\overrightarrow{d_j}$ de representação de um documento d_j é dado por $\overrightarrow{d_j} = (w_{1j}, w_{2j}, \cdots w_{tj})$. Um documento $\overrightarrow{d_j}$ e uma consulta $\overrightarrow{q}$ são representados como vetores em um espaço de dimensão t, como ilustra a Fig. 2.2.

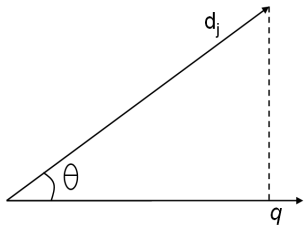

Fig. 2.2: O cosseno de θ representa a equação de relevância r(d_j, q). Adaptado de [9].

O modelo vetorial propõe avaliar o grau de relevância do documento $\overrightarrow{d_j}$ com relação à consulta $\overrightarrow{q}$ como a compatibilidade entre os vetores $\overrightarrow{d_j}$ e $\overrightarrow{q}$. Esta compatibilidade pode ser quantificada pelo cosseno do ângulo entre os dois vetores dado pela Eq. 2.2.

$$\text{r}(d_j, q) = \frac{\overrightarrow{d_j} \cdot \overrightarrow{q}}{\left\|\overrightarrow{d}_j\right\| \left\|\overrightarrow{q}\right\|} = \frac{\sum_{i=1}^{t} w_{ij}\, w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2}\ \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \tag{2.2}$$

Dado que $w_{ij} \geq 0$ e $w_{iq} \geq 0$ então r$(d_j, q) \in [0, 1]$. Ao invés de apenas predizer se um documento é relevante ou não o modelo vetorial gradua os documentos de acordo com o seu grau de relevância com a consulta. Um documento é recuperado mesmo que ele atenda à consulta apenas parcialmente.

O peso dos termos de indexação podem ser calculados de formas diferentes sendo que a mais comum é denominada $tf - idf$. O peso $tf - idf$ é dado por uma equação que faça o balanço entre o número de vezes em que um termo k_i aparece nos documentos d_j da coleção, denotado por tf_{ij}, e o número de vezes em que os documentos d_j (que contêm o termo k_i) aparecem na coleção, denotado por idf_i.

Seja N o número total de documentos na coleção e n_i o número de documentos nos quais o termo de indexação k_i aparece. Seja $freq_{ij}$ o número de vezes em que o termo k_i é mencionado no texto do documento d_j. A freqüência normalizada tf_{ij} do termo k_i no documento d_j é dada por $tf_{ij} = freq_{ij} / (\max_l freq_{lj})$ onde $\max_l$ é calculado sobre todos os termos que são mencionados no texto do documento d_j. Se o termo k_i não aparece no documento d_j então $tf_{ij} = 0$.

Considere n_i o número de documentos na coleção em que o termo k_i aparece ou seja a freqüência de documentos que contém o termo k_i. Esta freqüência é calculada pela equação logarítmica

$idf_i = \log\ (N/n_i)$. Se um termo aparece em apenas um documento a equação atribui valor máximo ($\log\ N - \log\ n_i = \log\ N - \log\ 1 = \log\ N$). Se um termo ocorre em todos os documentos a equação atribui valor zero ($\log\ N - \log\ n_i = \log\ N - \log\ N = 0$). Esta equação é denominada inverso da freqüência do documento.

O valor final do peso w_{ij}, do termo k_i no documento d_j, é dado por $w_{ij} = tf_{ij}\ idf_i$ onde tf_{ij} é a freqüência do termo no documento d_j e o idf_i é o inverso da freqüência dos documentos da coleção que contém o termo k_i.

O modelo vetorial é um modelo simples e rápido que permite estabelecer a relevância parcial de um documento com relação a uma consulta por meio da sua equação do cosseno. Desta forma é possível ordenar os documentos sendo que os mais relevantes ficam no topo da lista. Em termos teóricos o modelo vetorial assume que os termos de indexação são mutuamente independentes. Para melhorar os resultados do modelo vetorial é necessário empregar técnicas de expansão da consulta ou de realimentação da consulta [9].

2.3.3 Modelo Probabilístico

Dada uma consulta q e um documento d_j da coleção, o modelo probabilístico estima a probabilidade de que o documento d_j seja relevante para o usuário. O modelo assume que a probabilidade da relevância depende somente das representações da consulta e dos documentos. O modelo também assume que existe um subconjunto entre os documentos da coleção que o usuário prefira como resposta à consulta q. Este conjunto resposta ideal é rotulado como R.

No modelo probabilístico o peso dos termos de indexação é binário, isto é, $w_{ij} \in \{0,\ 1\}$ e $w_{iq} \in \{0,\ 1\}$. Uma consulta é um subconjunto dos termos de indexação (k_i). Seja R o conjunto de documentos considerados relevantes e $\overline{R}$ o complemento de R, ou seja, o conjunto de documentos não relevantes. $P\left(k_i|R\right)$ é a probabilidade do termo k_i estar presente em um documento selecionado aleatoriamente do conjunto R. $P\left(k_i|\overline{R}\right)$ é a probabilidade do termo k_i estar presente em um documento selecionado aleatoriamente do conjunto $\overline{R}$. A relevância de um documento d_j para a consulta q, considerando os t termos de indexação da coleção, é dada pela Eq. 2.3.

$$\mathrm{r}(d_j,\ q)\ \approx\ \sum_{i=1}^{t}\ w_{iq}\ w_{ij}\ \ \left(\log \frac{P\left(k_i|R\right)}{1\ -\ P\left(k_i|R\right)} + \log \frac{1\ -\ P\left(k_i|\overline{R}\right)}{P\left(k_i|\overline{R}\right)}\right) \tag{2.3}$$

Para aplicar o modelo probabilístico é necessário, inicialmente, calcular as probabilidades de $P\left(k_i|R\right)$ e $P\left(k_i|\overline{R}\right)$. O cálculo destas probabilidades é refinado por meio de aplicações recursivas do algoritmo de cálculo das probabilidades e pela observação dos resultados retornados. Uma discussão mais detalhada sobre o modelo probabilístico e sobre o cálculo das probabilidades de $P\left(k_i|R\right)$ e $P\left(k_i|\overline{R}\right)$ pode ser encontrada em [9, 22, 109].

2.4 Medidas de Desempenho de Sistemas de Recuperação de Informação

Existem duas medidas para avaliar o desempenho de um sistema de recuperação de informação: precisão (*precision*) e cobertura (*recall*) [9]. Para que estas medidas tenham significado é necessário ter domínio da coleção que está sendo testada, ou seja, é preciso ter conhecimento do conteúdo dos documentos da coleção. Para definição destas medidas considere a aplicação da consulta I, expressa pelo usuário, em uma coleção de referência contendo o conjunto R de documentos relevantes. Seja $|R|$ o número de documentos neste conjunto. Assuma que a estratégia de busca a ser testada processe a consulta I e retorne um conjunto de documentos A como resposta. Seja $|A|$ o número de documentos no conjunto A. Seja $|R_a|$ o número de documentos na interseção dos conjuntos R e A. A Fig. 2.3 ilustra estes conjuntos.

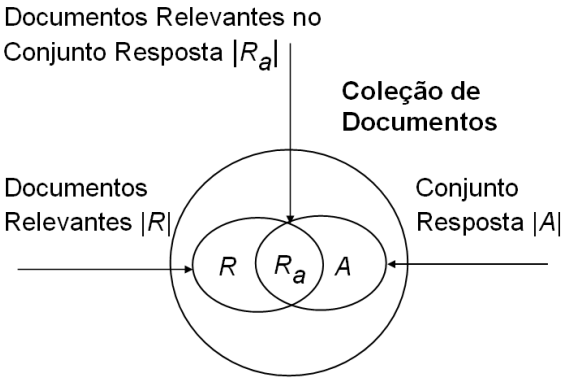

Fig. 2.3: Conjuntos R, A e R_a. Adaptado de [9]

As medidas de cobertura e precisão são definidas como se segue:

- Cobertura: é a fração do número de documentos relevantes recuperados $|R_a|$ pelo número total de documentos relevantes na coleção $|R|$, ou seja, Cobertura $= |R_a| \,/\, |R|$
- Precisão: é a fração do número de documentos relevantes recuperados $|R_a|$ pelo número total de documentos recuperados $|A|$, ou seja, Precisão $= |R_a| \,/\, |A|$

Para que estas medidas possam ser avaliadas o sistema ou usuário, que estão analisando os resultados, devem saber quantos documentos relevantes à consulta existem na base.

2.4.1 Gráficos de Precisão *versus* Cobertura

Os sistemas de recuperação de informação pretendem sempre melhorar as taxas de cobertura e precisão no conjunto de documentos recuperados. Para calcular as medidas de cobertura e precisão

os documentos do conjunto A devem ser examinados pelo usuário. Entretanto este conjunto de documentos não é apresentado ao usuário de uma única vez. Os documentos são ordenados de acordo com um critério de relevância e são apresentados ordenados por este critério em uma lista de documentos. O usuário examina esta lista de documentos começando pelo topo da mesma. Desta forma o valor das medidas de cobertura e precisão variam à medida em que o usuário examina o conjunto resposta A. Estas duas medidas são compatibilizadas em gráficos denominados precisão *versus* cobertura que são baseadas na representação por 11 níveis padrões de cobertura (*11 standard recall levels*). Estes gráficos são comumente utilizados na literatura para comparar algoritmos de recuperação de informação. A construção do gráfico de precisão *versus* cobertura será mostrada através de um exemplo [9].

Considere o conjunto $R_1 = \{d_3,\ d_5,\ d_9,\ d_{25},\ d_{39},\ d_{44},\ d_{56},\ d_{71},\ d_{89},\ d_{123}\}$ constituído pelos documentos relevantes para a consulta q_1. Assim existem dez documentos relevantes para a consulta. Considere um novo algoritmo de recuperação de informação que retorna, para a consulta q_1, o conjunto L de documentos ordenados por relevância como segue:

1. d_{123} •	6. d_9 •	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} •	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25} •	15. d_3 •

Os documentos de L relevantes para a consulta estão marcados com um • depois do número do documento. Ao examinar a ordem dos documentos deve-se notar alguns pontos. O documento d_{123} é relevante, está localizado na posição 1 e representa 10% de todos os documentos relevantes do conjunto R_1. Neste caso a precisão é de 100% para uma cobertura de 10%. O próximo documento relevante, d_{56}, está localizado na posição 3 da lista. Neste caso a precisão é de $66,6\%$ (dois documentos relevantes em um total de três) para 20% de cobertura (dois documentos recuperados dos dez presentes no conjunto R_1. O terceiro documento relevante, d_9 encontra-se na posição 6 resultando em uma precisão de 50% para 30% de cobertura. Na seqüência tem-se o documento relevante d_{25} (posição 10) com precisão de 40% para 40% de cobertura e o documento d_3 (posição 15) com precisão de $33,3\%$ para 50% de cobertura. Estes valores são representados pelo gráfico na Fig. 2.4 baseado nos 11 níveis padrões de cobertura dados por $0\%,\ 10\%,\ 20\%,\ 30\%,\ \cdots,\ 100\%$. O valor da precisão no nível de cobertura 0% é calculado por interpolação de acordo com a Eq. 2.5.

Os algoritmos de recuperação de informação são avaliados pelo seu desempenho em executar várias consultas. Para cada uma das consultas q_i é gerada uma curva de precisão *versus* cobertura. Para avaliar o desempenho do algoritmo, sobre o conjunto de consultas, calcula-se a média entre as medidas de precisão de cada consulta para cada nível de cobertura como dado na Eq. 2.4. Na Eq. 2.4

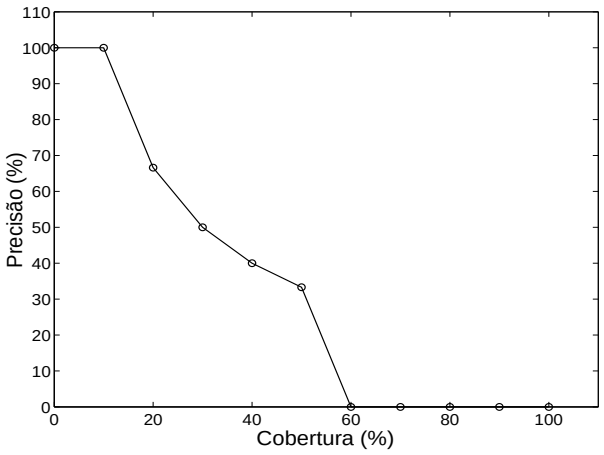

Fig. 2.4: Gráfico de precisão *versus* cobertura para os 11 níveis padrões de cobertura.

$\overline{P}(r)$ é a média dos valores de precisão para o nível de cobertura r, N_q é o número de consultas consideradas e $P_i(r)$ é o valor da precisão para a consulta q_i no nível de cobertura r.

$$\overline{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q} \tag{2.4}$$

Como os níveis de cobertura para cada consulta podem ser diferentes dos 11 níveis padronizados torna-se necessário adotar um procedimento de interpolação. Considere, por exemplo, uma consulta q_2 e seu conjunto de documentos relevantes dado por $R_2 = \{d_3,\ d_{56},\ d_{129}\}$. O conjunto de documentos retornados para a consulta q_2 é o mesmo conjunto L referente à consulta q_1. Seguindo o mesmo raciocínio empregado nas medidas de precisão da consulta q_1, o documento d_{56} (posição 3) possui precisão de $33,3\%$ para $33,3\%$ de cobertura. O segundo documento relevante d_{129} (posição 8) possui precisão de 25% para $66,6\%$ de cobertura e o terceiro documento relevante d_3 (posição 15) possui precisão de 20% para 100% de cobertura. O método de interpolação para calcular os valores para os 11 níveis é dado pela Eq. 2.5. O valor de precisão no j-ésimo nível de cobertura é dado pelo maior valor de precisão no intervalo $(j,\ j+1)$.

$$P(r_j) = \max_{r_j \le r \le r_{j+1}} P(r) \tag{2.5}$$

A Fig. 2.5 mostra o gráfico de precisão *versus* cobertura interpolado para considerar os 11 níveis de cobertura.

O gráfico resultante da média dos gráficos de precisão *versus* cobertura para as consultas q_1 e q_2 é mostrado na Fig. 2.6. O comportamento ideal aconteceria se o valor de precisão fosse igual a 100% para todos os níveis de cobertura indicando que todos os documentos relevantes foram recuperados, para todas as consultas, e foram apresentados no topo da lista de resultados. Infelizmente este não é

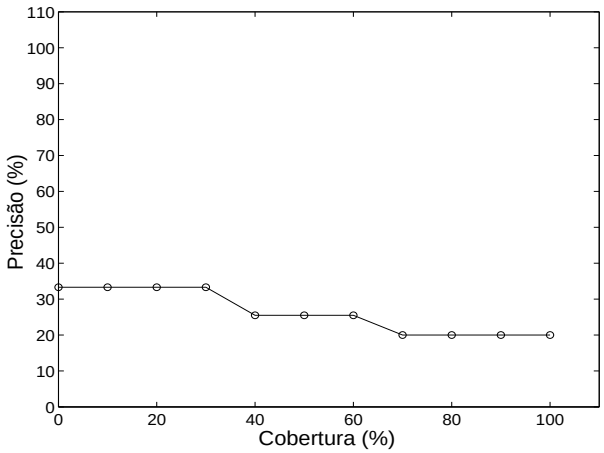

Fig. 2.5: Interpolação no gráfico de precisão *versus* cobertura para os 11 níveis padrões de cobertura para a consulta q_2.

o comportamento usual dos sistemas de recuperação de informação.

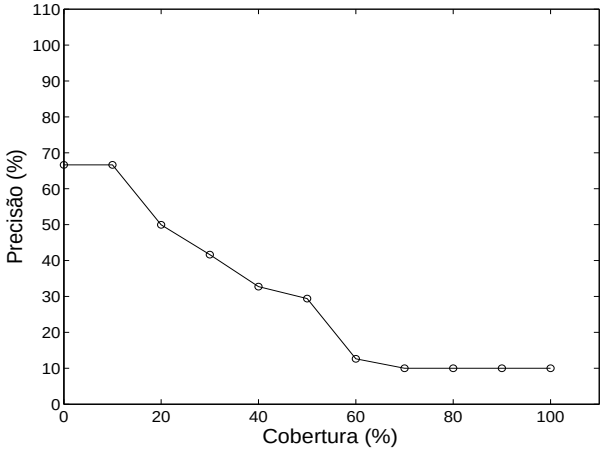

Fig. 2.6: Gráfico de precisão média *versus* cobertura para as consultas q_1 e q_2.

2.5 Problemas Relacionados à Recuperação de Informação

Em geral os sistemas de recuperação de informação não apresentam o comportamento ideal, ou seja, o usuário recebe uma lista muito grande de documentos como resposta e nem sempre os documentos relevantes à consulta são os primeiros da lista de documentos. Desta maneira os usuários devem despender um tempo razoável até encontrar aqueles que são realmente relevantes. Além disto os documentos são recuperados apenas quando eles contêm os termos especificados na consulta. Entretanto este enfoque vai negligenciar outros documentos que também possam ser relevantes mas

que não contêm os termos especificados na consulta. Ao considerar um domínio de conhecimento específico este problema pode ser superado pela incorporação de uma base de conhecimento no processo de recuperação de informação. A base de conhecimento é formada por uma ou mais estruturas conceituais que organizam o conhecimento de um domínio capturando os conceitos que descrevem este conhecimento e seus relacionamentos. Os documentos são indexados pelos conceitos da base de conhecimento ao invés do seu conteúdo léxico apenas. A partir do conhecimento existente na base é possível recuperar os documentos relevantes a uma consulta do usuário.

Ao utilizar uma base de conhecimento o objetivo é utilizar o conhecimento expresso na base para melhorar a qualidade dos documentos recuperados trazendo mais documentos associados à consulta (melhoria da taxa de cobertura) e apresentando estes documentos numa ordem onde os documentos do topo da lista de documentos sejam os mais relevantes à consulta (melhoria da taxa de precisão).

2.6 Técnicas para Recuperação de Documentos por Conceitos

A preocupação de se prover uma estrutura de representação dos documentos, que capture os temas ou conceitos expressos no mesmo, vem desde o processo de seleção dos termos de indexação. Conforme discutido na Seção 2.1 os pesos associados aos termos de indexação indicam quais deles exprimem de forma diferenciada os principais assuntos tratados em um documento. Quanto maior o peso associado a um termo de indexação maior é a importância deste termo para recuperar os documentos a ele associados.

Algumas técnicas procuram representar os conceitos dos documentos por meio de operações na representação lógica dos mesmos. São técnicas automáticas e não usam conhecimento externo providos por um especialista. Entre estas técnicas está a LSI (*Latent Semantic Indexing*) [24, 43] que possibilita que um documento seja recuperado quando ele compartilha conceitos com outros documentos que sejam relevantes à consulta do usuário. Ela é uma variação do modelo de recuperação vetorial.

Uma outra abordagem para tratar os conceitos associados aos documentos é pela utilização de estruturas conceituais utilizadas para fazer o mapeamento dos conceitos. Estas estruturas podem ser extraídas automaticamente da coleção, como por exemplo por meio de operações relativas à mineração de textos, ou construídas com a ajuda de um especialista do domínio. O objetivo destas estruturas é representar os conceitos do domínio e as relações entre eles. No caso de uso de uma estrutura conceitual os documentos deve ser classificados e indexados pelos conceitos desta estrutura. Por outro lado a consulta também deve ser composta com os conceitos da estrutura. Por meio da exploração das relações expressas na estrutura conceitual o sistema de recuperação de informação pode recuperar documentos contendo conceitos que não foram diretamente especificados pelo usuário mas que

podem estar relacionados com a sua consulta. O uso de estruturas conceituais para representar os conceitos de um domínio e suas relações pode ser explorado de duas formas para ajudar no processo de recuperação de informação.

A primeira forma é a expansão da busca que consiste em reformular a expressão de busca inicial adicionando novos conceitos que sejam relacionados semanticamente àqueles utilizados originalmente. Quando a expressão de busca é formada pelos conceitos de uma estrutura conceitual pode-se utilizar as relações existentes entre os conceitos da estrutura para selecionar novos conceitos a serem adicionados à expressão. A expressão de busca expandida é submetida ao mecanismo de recuperação de informação na expectativa de que sejam retornados mais documentos relevantes e que sejam semanticamente associados à expressão de busca original.

A segunda forma é através da navegação nos conceitos da estrutura conceitual. Na navegação o usuário pode interagir com a estrutura que vai ajudá-lo a focar na sua necessidade. Nesta atividade o usuário se depara com a estrutura que organiza o conhecimento existente na coleção permitindo que ele procure por conceitos que traduzam sua necessidade de busca. Ao selecionar um ou mais conceitos, na estrutura conceitual, estes vão formar a expressão de busca a ser submetida ao sistema. A navegação na estrutura conceitual permite uma visão dos conceitos ou temas tratados na coleção de documentos.

2.7 Resumo do Capítulo

Este capítulo apresentou uma visão geral do problema de recuperação de informação abordando o processo de busca da informação e os aspectos envolvidos na mesma, ou seja, como representar os documentos e as consultas e os modelos clássicos de recuperação de informação.

Também foi mostrado que a forma mais comum de recuperar informação é por meio de uma consulta composta de palavras-chave. O problema do uso de palavras-chave é que muitas vezes os documentos retornados não apresentam relevância para o usuário. Além disto os documentos são recuperados apenas quando eles contem os termos especificados na consulta. Entretanto este enfoque vai negligenciar outros documentos que também possam ser relevantes mas que não contem os termos especificados na consulta. Uma estratégia para a resolução deste problema está em indexar os documentos pelos conceitos neles expressos ao invés apenas das palavras presentes no seu texto. Uma das soluções, dentro desta estratégia, é utilizar estruturas conceituais tanto para indexar os documentos como para especificar ou expandir a consulta.

Neste trabalho acredita-se que o uso de estruturas conceituais pode ser de grande valia para ajudar o usuário a definir sua necessidade de busca. Ao selecionar os conceitos estes vão compor a expressão de busca. O uso das relações existentes na estrutura conceitual ou entre estruturas conceituais

distintas, para expandir a consulta, pode melhorar a qualidade de busca no sentido de recuperar um número maior de documentos semanticamente relevantes com relação à consulta original.

Capítulo 3

Estruturas Conceituais na Recuperação de Informação

Este capítulo apresenta as principais estruturas conceituais utilizadas na literatura para classificar e indexar os documentos pelo seu significado, ou seja, seu conteúdo semântico. Estas estruturas também podem ser utilizadas para realizar expansão da consulta e para proporcionar um ambiente de navegação para o usuário. Para cada tipo de estrutura serão apresentadas sua descrição e aplicações de recuperação de informação que empregam a estrutura para melhorar o processo de recuperação. Ao utilizar o conhecimento expresso em uma estrutura conceitual procura-se recuperar documentos que sejam mais relevantes para a consulta inicial do usuário.

3.1 Tesauro

3.1.1 Descrição

Os instrumentos de representação da informação para indexação, armazenamento e recuperação de documentos são considerados como linguagens documentárias [47]. As linguagens documentárias mais conhecidas são o tesauro e os sistemas de classificação bibliográfica. São linguagens artificiais por não resultarem de um processo evolutivo e por necessitarem de regras explícitas para seu uso como será apresentado ao longo do texto. Como todas as linguagens artificiais, não comportam exceções.

Palavra é a menor unidade léxica, cujo significado se depreende do contexto em que ela figura mas que, tomada isoladamente, pode ter vários significados. O uso de palavras na indexação e na recuperação é inadequado pela ambigüidade que elas carregam. Neste ponto começa a artificialidade da linguagem documentária: a partir de alguns princípios, escolhe-se uma determinada palavra ou

expressão para representar um único conceito, ou idéia. Quando isto se dá tem-se então não mais uma "palavra" mas um "termo". O controle dos termos é, portanto, necessário para que a cada um deles não se atribua mais do que um conceito e, também, para que a cada conceito não se atribua mais de um termo. Os termos escolhidos para nomear um conceito (entidade, objeto ou processo) são também chamados de descritores. Os outros são não-descritores e formam o conjunto das remissivas. A partir de um termo que o usuário conhece, o tesauro, por meio de sua estrutura hierárquica, mostra diversos outros que podem ser tão oportunos ou mais do que aquele que lhe veio à mente.

O tesauro organiza um domínio de conhecimento, na forma de termos selecionados e relacionados, sendo utilizado para representar os assuntos dos documentos e das solicitações de busca. A representação do assunto é feita no momento da indexação: o documento é analisado, seu conteúdo identificado, e devidamente "traduzido", de acordo com os termos do tesauro e com a política de indexação estabelecida. A representação da solicitação de busca é feita no momento em que o usuário busca uma informação no sistema: seu pedido é analisado, seu conteúdo identificado e devidamente "traduzido" nos termos do tesauro. As principais relações definidas entre os termos de um tesauro são dadas por:

- Termo mais geral ou BT (*broader term*): se refere ao termo acima na hierarquia e que possui um significado mais amplo. Na prática podem ocorrer vários BT para um único termo. A relação inversa é dada pelo termo mais específico ou NT (*narrower term*). Uma estrutura conceitual que consiste em uma taxonomia possui apenas estas duas relações definidas.

- USE: refere-se a um outro termo, o descritor, que é preferencial com relação ao termo dado e que é o indicado para o uso. Implica relação de sinonímia entre os termos. A relação inversa é dada por UF (*used for*) que indica os termos que não constituem em descritores autorizados e que devem ser substituídos pelo termo corrente.

- Termo relacionado ou RT (*related term*): refere-se a um termo que é relacionado semanticamente ao termo corrente mas que não se constitui em um sinônimo ou um termo mais geral ou mais específico.

Um exemplo que ilustra as relações de um tesauro é apresentado na Fig. 3.1. Neste exemplo os termos são extraídos do Thesagro (Thesaurus Agrícola Nacional) [14] desenvolvido pela BINAGRI (Biblioteca Nacional de Agricultura) [13]. Este exemplo mostra que o termo "Nutrição Animal" deve ser usado ao invés de "Alimentação Animal" indicado pela relação USE. Para o termo "Nutrição Animal" tem-se, como termo mais geral, o termo "Nutrição" indicado pela relação BT. Um termo mais específico é o termo "Alimentação suplementar", indicado pela relação NT, e um termo relacionado é o termo "Eficiência Nutricional", indicado pela relação RT.

Um tesauro pode ser empregado de várias formas [86] incluindo:

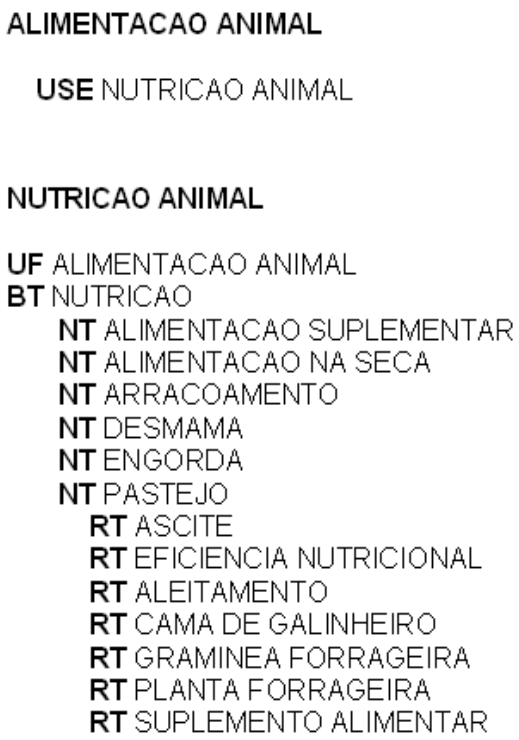

Fig. 3.1: Termos extraídos do Thesaurus Agrícola Nacional.

- Organização de conhecimento: um tesauro provê uma hierarquia de conceitos e seus relacionamentos organizando o conhecimento específico de um domínio.

- Normalização de terminologia: por meio da seleção de um único termo ou frase para representar cada conceito do domínio e por meio de relações de sinonímia, o tesauro reforça consistência terminológica.

- Expansão da consulta: um tesauro facilita a adição de termos a uma consulta provendo relações de hierarquia e associações explícitas entre termos.

3.1.2 Aplicações

Nesta seção são apresentadas algumas aplicações que usam o tesauro como estrutura conceitual.

DOPE

Um exemplo de uso de tesauro para indexação e navegação para recuperação de informação é o do projeto de medicamentos da Elsevier denominado DOPE [123]. Este projeto pesquisa formas de

prover acesso a múltiplas fontes de informação na área de ciências biológicas. O sistema desenvolvido é orientado pelo uso do tesauro EMTREE da Elsevier ([28] apud [123]). Neste projeto foram indexados, automaticamente, 5 milhões de resumos da base Medline e 500.000 artigos completos da coleção da Science Direct utilizando os termos do tesauro EMTREE. O sistema DOPE permite ao usuário realizar tanto a atividade de navegação como a atividade de busca.

Para recuperar informação, inicialmente o usuário deve fornecer uma palavra de busca. O sistema então indica os termos do tesauro relacionados à palavra inicial para que o usuário selecione um dos termos. Ao selecionar o termo o sistema permite ao usuário navegar no tesauro para selecionar um termo mais abrangente, mais específico ou relacionado ao termo escolhido para expandir sua busca. Uma vez que o usuário escolha o termo do tesauro ele se torna o termo de foco inicial e o sistema recupera todos os documentos associados a este termo. O sistema também lista todos os outros termos associados aos documentos recuperados. Ele apresenta uma hierarquia de dois níveis dos termos que co-ocorrem com o termo de foco. Esta hierarquia de termos é organizada como uma árvore e os termos são agrupados pela raiz da árvore (no tesauro) ao qual eles pertencem. A árvore vai constituir em um espaço de navegação dos termos associados. O usuário pode navegar nesta árvore e selecionar um ou mais termos para gerar uma visualização de como eles se relacionam e quais são os documentos associados.

O sistema DOPE permite tanto a expansão da consulta, através da navegação no tesauro EMTREE, como também utiliza a co-ocorrência de termos nos documentos para criar um espaço de navegação organizado de acordo com a estrutura do tesauro. Na avaliação realizada, os usuários acharam o sistema útil para explorar o espaço de informação. O uso da co-ocorrência de termos para refinar a consulta também foi visto como um ponto positivo. Futuras melhorias no projeto incluem tratar o conhecimento existente em mais de um tesauro ou múltiplas ontologias e explorar melhor o conhecimento existente nas ontologias para criar um mecanismo de pergunta e resposta.

Deja Vu

Deja Vu [49] é uma interface para recuperação em bibliotecas digitais contendo material digital catalogado utilizando um tesauro. Deja Vu permite aos usuários navegar pelos termos do tesauro para encontrar aqueles que representem sua necessidade de busca. Através da seleção de um ou mais termos o usuário compõe a consulta que será submetida ao sistema. Inicialmente o usuário deve selecionar um termo do tesauro através de uma listagem alfabética de todos os termos ou digitando uma palavra que causa o retorno de todos os termos que contêm esta palavra. Dada a lista de termos o usuário seleciona aquele que se aproxima de sua necessidade de busca. Uma vez que o usuário selecione um termo este passa a ser o termo de foco e o espaço de navegação vai incluir os termos mais gerais, os mais específicos e os termos relacionados ao termo de foco baseado nas relações

expressas no tesauro. Neste espaço os termos do tesauro que co-ocorrem com o termo de foco, nos recursos da coleção, são destacados possibilitando ao usuário selecionar termos para compor uma consulta que realmente retorne algum material. À medida que o usuário vai selecionando os termos o sistema automaticamente destaca os demais para mostrar aqueles que podem ser compostos com os já selecionados para refinar a consulta. Também é apresentada uma lista de recursos associados aos termos.

O sistema Deja Vu foi avaliado em duas coleções de imagens digitais: a coleção de 11.000 imagens capturadas de um vídeo em um disco laser na Universidade do Estado de Dakota do Norte e a coleção de 25.000 fotos e imagens da Biblioteca do Congresso em Washington, ambos nos Estados Unidos. Na avaliação do sistema, a navegação em uma interface que permita a visualização da integração dos termos de catalogação com os recursos foi visto como um ponto forte pois permite que o usuário tenha uma visão imediata dos recursos disponíveis à medida que ele navega pelos termos do tesauro. Uma das necessidades levantadas na avaliação do sistema é que ele considere outros tesauros no espaço de navegação uma vez que um mesmo recurso digital pode ter sido catalogado usando termos de mais de um tesauro. Também foi apontado que o sistema não oferece ajuda no caso do usuário desejar pesquisar pelo nome de autor ou por outra informação que esteja relacionada com a produção do recurso.

Phind

Phind [96, 97] é um sistema que utiliza uma combinação de termos de tesauro e frases extraídas automaticamente dos documentos para montar o espaço de navegação apresentado na interface. O sistema é aplicado à coleção de documentos da FAO (*Food and Agriculture Organization*) [38] catalogados utilizando os termos do tesauro AGROVOC [40] construído manualmente para sistemas de informação agrícola desta organização. O sistema identifica todas as frases que ocorrem no texto completo dos documentos e as organiza em uma hierarquia baseada na inclusão léxica, isto é, uma frase referencia outras frases mais longas e mais específicas nas quais ela está incluída. Uma frase é uma seqüência de palavras que ocorrem mais de uma vez no texto. Por exemplo, a palavra “floresta” está incluída na frase “floresta sustentável” que por sua vez está incluída na frase “gerenciamento de floresta sustentável” e assim por diante.

A interface apresenta esta coleção de frases organizadas de forma hierárquica começando por frases compostas por menos palavras até frases compostas por mais palavras sendo estas mais específicas. O sistema é integrado ao tesauro, para que possam ser exploradas as relações existentes entre os termos do tesauro quando o usuário estiver navegando na interface. Os usuários podem examinar as frases extraídas da coleção de documentos e ter acesso aos documentos em que elas estão contidas digitando uma palavra inicial. O sistema vai mostrar as frases que contém a palavra inicial

e também as entradas do tesauro que contém esta palavra. As entradas do tesauro sugerem novos relacionamentos e novos termos de pesquisa relativos ao termo inicial. Para navegar pelas frases da coleção, os usuários podem alternar entre selecionar as frases extraídas pelo sistema ou as entradas no tesauro para que o sistema retorne um novo conjunto de frases contendo sua seleção. Ao mostrar as frases extraídas o sistema apresenta uma lista com os títulos dos documentos que contém as frases retornadas.

Rede de Conceitos *Fuzzy*

No sistema de Rede de Conceitos *Fuzzy* [19] a estrutura conceitual consiste em uma rede onde os conceitos estão representados pelos nós e as relações entre os conceitos são dadas pelas ligações entre os nós. A força da relação entre os conceitos é dada por um coeficiente *fuzzy* com valor no intervalo [0,1] associado à relação. Valores iguais a zero indicam a ausência da relação, valores iguais a 1 indicam uma relação forte e valores intermediários indicam os graus da relação situados entre inexistente (valor 0) e fortemente relacionados (valor 1). Neste sistema cada nó pode estar associado a um outro nó por três tipos distintos de relação dadas por:

- Associação Positiva *Fuzzy*: relaciona conceitos que possuem significados similares. Exemplo: pessoa $\leftrightarrow$ indivíduo.

- Generalização *Fuzzy*: um conceito é uma generalização de outro conceito se ele consiste daquele conceito $(\text{máquina} \rightarrow \text{parafuso})$ ou se ele inclui aquele conceito $(\text{veículo} \rightarrow \text{carro})$.

- Especialização *Fuzzy*: é o inverso da relação de generalização.

A rede de conceitos é construída automaticamente baseada na co-ocorrência sintática na coleção e neste trabalho, em função dos tipos de relacionamentos existentes, é considerada similar a um tesauro. Os documentos estão associados aos conceitos por uma relação *fuzzy* com valor no intervalo [0,1] que indica o grau de associação entre o conceito e o documento. Valores iguais a zero indicam a ausência de associação entre o documento e o conceito, valores iguais a 1 indicam uma relação forte de associação entre o documento e o conceito. Valores intermediários indicam os graus de associação situados entre ausente (valor 0) e fortemente associados (valor 1). Através dos relacionamentos existentes na rede de conceitos o sistema infere novos relacionamentos explorando as relações implícitas entre os conceitos, mesmo que elas não tenham sido especificadas inicialmente pelo especialista do domínio. Desta forma um documento que antes não era associado a um determinado conceito pode vir a ter um grau de associação maior que 0 em função das inferências realizadas na rede. Deve-se observar que a associação dos documentos a novos conceitos passa a existir em função das inferências realizadas

na rede de conceitos. Também é importante notar que cada documento possui um peso associado ao conceito relacionado significando o quanto o conceito é importante no documento.

Na consulta o usuário especifica os conceitos que ele deseja que estejam representados nos documentos e associa um valor no intervalo [0,1] a cada conceito. Este peso vai significar o quanto o conceito é importante nos documentos a serem recuperados. Quando a consulta é submetida ao sistema ele verifica a similaridade entre os pesos dos conceitos na consulta e os pesos dos mesmos conceitos nos documentos. Os documentos que possuírem o valor de similaridade acima de um limite pré-estabelecido são retornados ao usuário.

Tesauro Automático

O sistema Tesauro Automático [103] propõe um método para a criação automática de um tesauro a partir de um conjunto de termos extraído de uma coleção, sobre um domínio de conhecimento, e de tesauros construídos manualmente. Este tesauro é utilizado para expansão da consulta possibilitando uma melhora nas medidas de cobertura e precisão. O conjunto inicial de termos é obtido a partir da seleção dos termos em uma coleção de documentos sobre o assunto de interesse. Esta seleção de termos é realizada por meio pré-processamento linguístico, como *stemming* e lista de *stopwords*, e pelo uso das medidas de indexação $tf - idf$ para selecionar os termos candidatos. A seguir um tesauro inicial é construído considerando as relações dos termos do conjunto inicial em tesauros já existentes. No caso foram utilizados três tesauros: Spines (contém informação sobre a área de ciência e tecnologia), Eurovoc (contém conceitos relativos a atividades realizadas na União Européia) e ISOC-Economy (contém informação na área de economia). A seguir o tesauro inicial é enriquecido com novas relações de equivalência, hierarquia ou associação a partir do conhecimento extraído nos documentos da coleção por meio de medidas de co-ocorrência e, também, pela consulta ao dicionário Real Academia Española, na língua espanhola.

O tesauro automático foi utilizado na expansão da consulta considerando o *framework* de testes provido pelo CLEF (*Cross-Language Evaluation Forum*) [125], no ano de 2001, na língua espanhola. Um total de 50 consultas foram expandidas e executadas considerando o tesauro gerado e os mesmos tesauros utilizados para gerar o tesauro inicial. Nos testes realizados as consultas expandidas, utilizando o tesauro gerado automaticamente, apresentaram melhores resultados de precisão para as medidas de cobertura quando comparado com a expansão utilizando os demais tesauros e a consulta sem expansão.

Tesauro Sócio-Político

O sistema Tesauro Sócio-Político [2, 82] apresenta o desenvolvimento do Tesauro Sócio-Político nas línguas russa e inglesa e uma aplicação do mesmo na recuperação de informação. O tesauro

foi desenvolvido ao longo de 10 anos e é considerado uma ontologia linguística como uma rede hierárquica de conceitos. A principal unidade do tesauro é o conceito. Um conceito possui um nome em inglês e russo, um conjunto de expressões lingüísticas usadas para referenciar o conceito nos textos, relações de taxonomia com outros conceitos, *related-term* (RT), relações do tipo "parte-de" e relações ontológicas de dependência (por exemplo o conceito *FOREST* depende do conceito *TREE* mas o inverso não é verdade pois uma árvore pode existir sem estar em uma floresta).

O tesauro foi utilizado para indexação de textos e expansão da consulta podendo-se utilizar termos em russo ou inglês. O processo de expansão da consulta considera os termos presentes no título, descrição ou na narrativa dos documentos de forma diferenciada. Os experimentos foram realizados considerando documentos providos pelo CLEF em 2005. As consultas foram construídas utilizando os conceitos do tesauro. Os resultados de melhoria de precisão para as mesmas taxas de cobertura, obtidos com o uso do tesauro sócio-político, ficaram entre os quatro melhores quando comparado com os mesmos experimentos realizados com outros tesauros.

3.2 Facetas

3.2.1 Descrição

A classificação por facetas trata os recursos ou grupos de recursos como coleções que possuem características claramente definidas e mutuamente exclusivas. Estas características são chamadas de facetas [25, 130]. No domínio da arqueologia, por exemplo, os artefatos podem ser classificados por múltiplas dimensões ortogonais: período de tempo, cultura, material e localização geográfica. Cada uma destas dimensões constitui-se em uma faceta. Assim as facetas podem ser vistas como eixos distintos ao longo dos quais os recursos podem ser classificados. Uma faceta está associada a uma dimensão de classificação e indica uma visão ou perspectiva distinta dos recursos. Cada faceta possui uma quantidade de termos. A forma como os termos estão arranjados nas facetas pode variar mas é usual que uma estrutura do tipo tesauro seja utilizada. Geralmente um termo pertence a apenas uma faceta. A idéia é selecionar um termo em cada faceta para descrever o recurso ao longo dos eixos ou dimensões. A classificação dos recursos nas facetas indica como os conceitos co-ocorrem na coleção. As facetas podem ser utilizadas para organizar qualquer domínio. Dentro deste tópico um dos desafios é estabelecer o número de facetas a ser utilizado e o conceito que será refletido em cada uma.

A classificação por facetas foi proposta originalmente por Ranganathan [120] para classificação de documentos. Sua proposta original, também conhecida como classificação Colon, consistia de cinco facetas:

- Personalidade: é considerada a faceta mais importante e está relacionada ao tema principal do documento.
- Matéria: a matéria ou substância descrita no documento.
- Energia: os processos ou atividades descritas no documento.
- Espaço: as localizações descritas no documento.
- Tempo: o perído de tempo descrito no documento.

Para Ranganathan qualquer conceito sobre o qual um livro ou documento pudesse ser escrito poderia ser representado considerando sua descrição nestas cinco facetas. Dado este conjunto de facetas, um documento que tratasse do "projeto de mobília em madeira na América no século XVIII" seria classificado como: Personalidade (mobília), Matéria (madeira), Energia (projeto), Espaço (América) e Tempo (século XVIII).

3.2.2 Aplicações

Nesta seção são apresentadas algumas aplicações que usam a estrutura de facetas para que os recursos possam ser classificados nas suas dimensões.

DocCube

O sistema DocCube [8, 89] oferece um espaço semântico, descrito por hierarquias de conceitos dependentes do domínio, para busca e exploração da informação. Cada hierarquia de conceitos é vista como uma dimensão ou uma faceta através da qual os documentos da coleção são classificados automaticamente. O número de dimensões e a sua especificação dependem do domínio e da necessidade de informação dos usuários. Para o domínio de economia, por exemplo, as dimensões podem ser indicadores econômicos, tipo de indústria e regiões geográficas envolvidas. Cada hierarquia pode se constituir em uma ontologia.

O sistema corresponde a um portal onde domínios diferentes estão disponíveis. O usuário pode selecionar um domínio, as hierarquias e, dentro de cada hierarquia, os conceitos que lhe interessam visualizar. Para explorar o espaço de informação os usuários navegam nas hierarquias de conceitos selecionadas especificando e refinando suas necessidades de informação. Na interface principal do sistema as hierarquias são apresentadas como eixos e a dispersão dos documentos em relação aos conceitos é mostrada ao longos destes eixos. O topo das hierarquias mostram os conceitos mais gerais. À medida que o usuário atravessa os níveis das hierarquias ele tem acesso aos conceitos mais específicos. Os conjuntos de documentos relativos aos conceitos são representados como esferas. As

dimensões das esferas estão relacionadas com o número de documentos associados. Ao selecionar uma ou mais esferas os documentos correspondentes são apresentados. As hierarquias selecionadas e os conceitos presentes em cada uma constituem o conhecimento pelo qual os documentos são classificados. Uma das vantagens apontada pelos autores é a facilidade de navegação pelas estruturas conceituais guiando o usuário pelo espaço de informação evitando que ele perca o contexto semântico da sua busca ou de seu interesse.

Coleção de Artes

Este sistema foi implementado para realizar busca em 35 mil imagens de uma coleção de artes do Museu de Artes Finas de São Francisco [131] através da navegação e seleção de termos organizados em hierarquias que representam facetas das imagens. As facetas são extraídas semi-automaticamente dos metadados e das descrições associados às imagens. O sistema permite tanto a atividade de navegação, nas hierarquias das facetas, como a atividade de busca pela entrada de um termo de busca. Inicialmente o sistema apresenta todas as possíveis facetas do domínio e suas categorias mais gerais possibilitando vários pontos de partida para a navegação. O usuário pode selecionar as categorias das facetas ou entrar com um termo de busca. Em função das categorias selecionadas ou do termo de busca fornecido o sistema reorganiza o espaço de busca, para apresentar os tópicos relacionados com a seleção inicial, juntamente com os grupos de imagens correspondentes. Este processo de seleção e reorganização do espaço de busca vai sendo refinado à medida que o usuário selecione valores mais específicos nas facetas até que ele escolha uma imagem para ser apresentada.

O sistema é implementado utilizando os recursos de um banco de dados relacional. O sistema utiliza a hierarquia das facetas para organizar as imagens e para ajudar na visualização das características a serem exploradas. Na avaliação realizada o comportamento padrão dos usuários foi de achar a interface organizada permitindo visualizar sobre o que procurar. A interface recupera as imagens e as agrupa nas facetas realizando o trabalho de classificação para o usuário. Um dos pontos para melhoria é o tempo de resposta do sistema.

Navegação Personalizada

O sistema de Navegação Personalizada [126] permite a busca e navegação no domínio multimídia. O sistema propõe um navegador semântico facetado adaptativo construído a partir do paradigma de navegação por facetas, uso de uma ontologia de domínio e adaptação baseada em um modelo de usuário capturado automaticamente. Através da navegação no sistema o usuário pode fazer busca de recursos multimídia. Para facilitar a modelagem automática das preferências do usuário, os eventos que ocorreram durante a iteração do usuário são armazenados para uso na próxima sessão e os caminhos percorridos pelo usuário na sessão corrente também são utilizados para guiar a adaptação do

sistema. A ontologia de domínio modela o conjunto de características e atributos de recursos multimídia sendo utilizada para a geração automática de facetas à medida em que o usuário navega no sistema. O usuário pode selecionar restrições sobre os atributos para especificar melhor sua consulta. Pelo uso destas características o sistema permite visões adaptativas do espaço de busca, prevenção de sobrecarga de informação, refinamento da consulta, orientação no espaço de busca e apresentação e navegação visual. Um conjunto de facetas para o domínio multimídia relacionado a fotografias é composto por data de criação, local relacionado à fotografia e o tema da fotografia.

Experimentos iniciais indicaram que o mecanismo de adaptação reduziu o tempo de processamento e de refresh devido ao reduzido número de facetas ativas para disponibilizar. Entretanto o número de clicks do usuário aumentou uma vez que nem sempre as facetas corretas estavam ativas sendo necessário habilitá-las manualmente. O uso de restrições nos atributos reduziram o tempo de busca e o número de clicks permitindo que a busca fosse mais especializada. Trabalhos futuros incluem experimentos com maior número de usuários e uso de redes de relacionamento dos usuários para capturar suas preferências.

Fragmentos de Documentos

A recuperação de fragmentos de documentos evita que os usuários tenham que procurar a informação percorrendo todo o texto do documento. O sistema de Gerenciamento de Fragmentos de Documentos [79, 80] recupera a informação correta pelo gerenciamento de componentes ou fragmentos de documentos. O sistema integra tecnologias de marcação de texto, anotação de texto, extração de fragmentos de documentos e mecanismo de navegação e classificação por facetas para recuperação de fragmentos de documentos. Ele permite a recuperação do texto e dos metadados dos documentos por meio de diferentes perspectivas e granularidades considerando um ou múltiplos documentos.

O esquema de classificação por facetas é uma proposta de solução para o fato de que o conhecimento possui muitas dimensões e o número de conceitos nestas dimensões é muito grande. Cada uma das dimensões pode ser representada por uma faceta sendo cada faceta organizada por meio de uma estrutura conceitual composta de conceitos e relacionamentos entre eles. O topo da árvore de classificação de facetas possui duas partes. A primeira parte relaciona os diversos tipos de decomposição de documentos, como por exemplo, decomposição física (cabeçalho, capítulo, seção) e decomposição lógica (experimento, revisão, discussão). A segunda parte está associada ao conteúdo da informação, no caso do sistema este conteúdo é referente a projetos de engenharia. Os fragmentos de documentos são anotados com informações tanto da faceta relacionada ao tipo de decomposição quanto da faceta relacionada ao conteúdo da coleção. O sistema permite que outros tipos de classificações sejam acopladas na árvore de representação por facetas.

O usuário pode pesquisar os documentos de três formas: navegação pelo tipo de esquema de

decomposição, navegação pelo conteúdo dos documentos ou pelo uso de palavras-chave. Estas três formas podem ser utilizadas separadamente ou em conjunto permitindo a realização de consultas complexas e precisas. Os autores consideram que o uso de classificação por facetas permite ao usuário inexperiente conhecer e explorar o espaço de busca evitando que ele tenha que adivinhar o termo de busca correto para que sua pesquisa tenha sucesso. O uso de fragmentos de documentos permite que o usuário explore a informação precisa que ele necessita sem precisar pesquisar todo o texto do documento. Trabalhos futuros incluem explorar outras formas de decomposição de documentos e utilizar os fragmentos retornados para compor novos documentos facilitando o reuso de informação.

3.3 Ontologia

3.3.1 Descrição

Ontologia, na visão deste trabalho, representa o ápice em termos de estruturas conceituais para descrever os assuntos em um documento. As estruturas citadas anteriormente possuem uma linguagem definida para a sua construção. Em uma ontologia esta linguagem é aberta e é construída em função do domínio. Por exemplo, as taxonomias possuem apenas as relações de hierarquia para associar os termos, os tesauros possuem as relações de termo mais geral ou específico, sinônimos e termos relacionados. A classificação por facetas constitui mais em uma forma consistente de aplicar as estruturas conceituais. Em uma ontologia o criador pode especificar tanto os termos do vocabulário como diferentes tipos de relações. Ele não precisa se prender a relações pré-definidas como acontece no tesauro com suas relações BT, NT, e USE. O criador pode construir outros tipos de relações, existentes entre os termos, em função da necessidade do domínio que está sendo modelado. Neste sentido considera-se que a linguagem para especificar as relações é aberta.

Na área de inteligência artificial uma ontologia se refere a um artefato de engenharia constituído por um vocabulário utilizado para descrever uma determinada realidade juntamente com um conjunto de asserções considerando o significado a ser expresso [53]. Este vocabulário captura os conceitos que se pretende tratar.

A comunidade de ontologias distingue entre ontologias que são principalmente uma taxonomia, denominadas ontologias *lightweight*, de ontologias que modelam o domínio de uma maneira mais profunda fornecendo mais restrições sobre a semântica do domínio, denominadas ontologias *heavyweight*. Ontologias *lightweight* incluem conceitos, taxonomias dos conceitos, relacionamentos entre conceitos e propriedades que descrevem os conceitos. Ontologias *heavyweight* adicionam axiomas e restrições às ontologias *lightweight* [48]. Uma vez que se construa uma ontologia para um domínio, o conhecimento, traduzido pela ontologia, pode ser compartilhado por outras aplicações que

tratem do mesmo domínio [18, 42].

Na área da computação uma ontologia compreende um conjunto de definições de conceitos, propriedades, relações, restrições, axiomas, processos e eventos que descrevem um certo domínio ou universo de discurso. Provendo este corpo de definições sobre um domínio, uma ontologia capacita aplicações e agentes de software a usar uma semântica precisa, clara e formal para processar a informação descrita e para usar esta informação em aplicações inteligentes.

Para utilizar ontologias é necessário definir como o conhecimento vai ser representado especificando os tipos de conceitos e as relações entre os conceitos. Tanto os conceitos quanto as relações podem ser representados por taxonomias. Os conceitos podem ser instanciados em indivíduos. Existem paradigmas distintos para representar o conhecimento de uma ontologia como *frames*, cálculo de predicados de primeira ordem e lógica descritiva. Para cada paradigma existem linguagens de representação específicas e seus respectivos mecanismos de inferência [48]. Por exemplo, para o paradigma de *frames* existe a linguagem FLogic (*Frame Logic*) [67] associada ao mecanismo de inferência Ontobroker [23, 94]. Para o paradigma de cálculo de predicados de primeira ordem existe a linguagem KIF (*Knowledge Interchange Format*) [45] associada ao mecanismo de inferência JTP (*Java Theorem Prover*) [41]. Para o paradigma de lógica descritiva existem linguagens como DAM+OIL [55] associada ao mecanismo de inferência FaCT (*Fast Classification of Terminologies*) [58] e a linguageem OWL (*Web Ontology Language*) [114] associada ao mecanismo de inferência RACER [88]. A linguagem OWL representa uma evolução da linguagem DAM+OIL. OWL é a linguagem desenvolvida e recomendada pelo consórcio W3C (*World Wide Web Consortium*) para ser utilizada para descrever classes e relações que representam recursos disponíveis na WWW.

Um dos campos de aplicação de ontologias é em processamento de textos [86]. Existe uma ligação entre ontologias e texto escrito em linguagem natural. As ontologias tentam capturar a representação de um domínio em uma linguagem formal. Os seres humanos normalmente formulam suas descrições de um domínio utilizando linguagem natural textual. Esta correspondência entre ontologias e textos, enquanto formas de expressar significado, pode ser explorada aplicando ontologias em aplicações de processamento de textos. Sistemas de processamento de textos baseados em ontologias são de grande utilidade para processar dados relacionados a um domínio específico onde as máquinas de busca baseadas em palavras-chave, embora robustas, muitas vezes retornam resultados imprecisos uma vez que elas retornam muitos documentos irrelevantes para uma dada consulta de usuário. O maior problema deste enfoque é a negligência do contexto associado aos documentos deixando de recuperar documentos úteis associados à consulta. Uma das formas de contornar este problema é prover os sistemas de recuperação de informação com capacidade de inferência permitindo-os buscar documentos associados à consulta pelo seu significado e por conceitos relacionados. A combinação de recuperação de informação e ontologias surge como uma alternativa para processar textos utilizando informações

do contexto e constitui uma oportunidade para pesquisa e desenvolvimento. Nesta área podem ser adotados três enfoques:

- Expansão de consulta baseada em ontologia: a consulta do usuário é melhorada adicionando conceitos relacionados que estão organizados em uma ontologia.

- Sistemas de recuperação de informação semânticos: o corpo de documentos é previamente anotado de acordo com ontologia do domínio durante o processo de indexação. O processo de recuperação envolve um procedimento de inferência para checar o resultado ou melhorá-lo utilizando informação contextual presente na ontologia. O conhecimento extraído da ontologia, pelo procedimento de inferência, é utilizado pelo mecanismo de recuperação de informação para recuperar mais documentos relevantes à consulta.

- Sistemas de coleta de informação baseados em ontologia: desempenham funções relativas a processamento de texto como classificação, extração e busca. O conhecimento existente nas ontologias permite que estes tipos de sistemas realizem inferências e possam qualificar e interpretar o conteúdo semântico dos textos processados.

Quando dois textos são anotados considerando ontologias distintas é possível tirar conclusões significativas sobre as relações semânticas dos textos apenas se for possível caracterizar os relacionamentos entre os conceitos da duas ontologias [86]. Todas as dificuldades existentes na manipulação de ontologias afetam a área de interseção entre ontologias e recuperação de texto. Neste contexto o alinhamento de ontologias é particularmente importante para recuperação de textos porque a promessa na recuperação de textos é de que o uso de ontologias vai permitir a busca ao nível semântico ao invés de ser confinada ao nível léxico [91]. Não se deve esperar que os textos sejam anotados considerando uma única ontologia universal. Assim para pesquisar em coleções que contenham anotação que se refira a mais de uma ontologia é necessário a habilidade de mapear conceitos expressos em uma ontologia para conceitos equivalentes ou semanticamente relacionados em outras ontologias e explorar este mapeamento na recuperação dos textos.

3.3.2 Aplicações

Nesta seção são apresentadas algumas aplicações que usam ontologia como estrutura conceitual para apoiar a recuperação de informação quando se usa palavras-chave ou quando se navega na ontologia para selecionar os conceitos pelos quais os documentos devem ser recuperados.

CIRI

No sistema CIRI [3] são utilizadas ontologias para seleção de conceitos, criação de uma consulta a partir dos conceitos, expansão da consulta utilizando o conhecimento das ontologias e submissão da consulta à máquina de busca. As relações nas ontologias podem ter pesos atribuídos. Estes pesos podem ser utilizados para controlar a expansão da consulta. A máquina de busca a ser utilizada pode ser escolhida pelo usuário e o sistema expande a consulta considerando as necessidades da máquina de busca selecionada.

Inicialmente o usuário escolhe a ontologia em que deseja navegar para selecionar os conceitos e compor a consulta. A consulta vai ser composta de conceitos da ontologia escolhida. Quando o usuário termina a seleção dos conceitos da consulta o sistema automaticamente expande a consulta, considerando as associações da respectiva ontologia, e a submete à máquina de busca. Os documentos retornados são listados na interface. O sistema utiliza banco de dados relacional para modelar as relações nas ontologias. Apesar das ontologias serem construídas utilizando o editor Protégé [119], e exportadas em OWL, elas são transformadas em tabelas de um banco de dados relacional. Este sistema explora ontologias para recuperação de informação utilizando as relações de hierarquia e associação para expandir os conceitos e reformular a consulta.

Para validar o sistema foram realizados testes junto ao usuário [124]. Para efeitos de comparação, além do sistema CIRI, foi utilizado um sistema auxiliar, baseado em palavras-chave, que não possuía apoio de uma ontologia. Para a validação foi utilizada uma ontologia da área de indústria de alimentos e uma coleção de artigos de jornais sobre o mesmo tema. O sistema auxiliar apresentou resultados um pouco melhores com relação à facilidade de uso e nas taxas de precisão e cobertura. Um dos motivos foi o fato dos usuários não encontrarem o conceito desejado na ontologia e utilizarem vários conceitos secundários na consulta influenciando a ordem de apresentação dos resultados na qual o conceito principal não aparecia no topo. Muitos usuários apontaram que eles precisariam mais prática com o uso do sistema CIRI para avaliar melhor sua utilidade.

OnAir

O sistema OnAIR [98] é um sistema de recuperação de informação que utiliza uma ontologia para auxiliar na recuperação de clips de vídeos utilizando consultas em linguagem natural. A ontologia é utilizada para expandir a consulta do usuário. O sistema foi desenvolvido para recuperar trechos do vídeo da entrevista com a artista brasileira Ana Teixeira. Para tanto foi desenvolvida uma ontologia com conceitos e relações sobre arte contemporânea utilizando o editor Protégé e a linguagem OWL. Os trechos de vídeo são indexados através de palavras chave atribuídas por um especialista do domínio e por palavras contidas nas transcrições da entrevista relativa ao vídeo. A importância das palavras

na coleção é dada associando um peso às mesmas.

Quando o usuário faz uma consulta, em linguagem natural, o sistema trata esta consulta para considerar apenas os termos relevantes da consulta na expressão de busca eliminando as palavras que não carregam significado. O sistema associa pesos aos termos da expressão de busca em função da sua frequência na coleção e da sua presença ou ausência na ontologia. Em seguida é feita a expansão da consulta pelos termos de indexação em função do conhecimento existente na ontologia. Pelos testes realizados o trabalho conclui que ao considerar o uso da ontologia para expansão da consulta o sistema melhora a precisão e a cobertura da recuperação.

OWLIR

O sistema OWLIR [39, 112] recupera documentos que contém texto livre e anotação semântica. A anotação semântica consiste em adicionar conhecimento aos documentos em forma de marcação (*tags*). A anotação semântica ocorre antes da fase de indexação dos documentos e a informação contida na anotação é indexada permitindo que os conceitos da anotação sejam associados aos documentos como se fossem índices. A idéia é que a indexação da anotação semântica melhore a performance da recuperação de informação. Neste caso o conhecimento utilizado na anotação semântica é extraído do próprio texto podendo ser acrescido com conhecimento obtido pela inferência em uma ontologia.

O sistema utiliza uma ontologia sobre eventos de uma universidade e foi aplicado sobre uma coleção de 1540 páginas de anúncios de eventos desta mesma universidade. Inicialmente são extraídas as palavras e frases das páginas de eventos visando identificar os tipos de eventos tratados na coleção. No próximo passo o sistema anota as páginas utilizando a informação extraída do texto acrescida do conhecimento inferido na ontologia. Em seguida é feita a indexação dos documentos anotados. Quando uma consulta é processada ela é acrescida com conhecimento obtido da ontologia.

Os testes de recuperação foram realizados sobre três tipos diferentes de base: uma contendo páginas de eventos indexadas apenas pelo seu texto, outra contendo páginas indexadas por texto e anotação semântica contendo apenas a informação extraída do texto e a outra contendo texto e anotação semântica com conhecimento adicional obtido por inferência sobre o conhecimento existente na ontologia. A recuperação na base contendo documentos indexados com conhecimento adicional existente na ontologia exibiu melhor precisão para uma mesma taxa de cobertura.

OntoSeek

O sistema OntoSeek [54] é um sistema de recuperação de informação baseada em conteúdo, aplicado em páginas amarelas e catálogo de produtos, visando a aumentar a precisão e a cobertura na recuperação de informação dos produtos. Ontoseek combina um mecanismo de equiparação (*matching*) de conteúdo suportado por uma ontologia com um formalismo para representação tanto dos produtos

como das consultas. Este formalismo é constituído por estruturas denominadas grafos conceituais léxicos simplificados onde os nomes dos nós e dos arcos expressam os produtos e suas relações. Para criar os grafos conceituais simplificados é utilizado o conhecimento da ontologia.

Ao utilizar grafos conceituais, tanto para representar recursos como para representar consultas, o problema de recuperação por conteúdo se reduz à equiparação de grafos. Os nós e arcos individuais de um grafo, representando uma consulta, devem se equiparar aos nós e arcos de um grafo representando um produto considerando se a ontologia indica que existe uma relação de isomorfismo ou de subsunção entre eles. Os produtos recuperados são apresentados como uma página expressa na linguagem HTML (*HyperText Markup Language*). No projeto foi utilizada a ontologia Sensus ([71] apud [54]) que consiste numa extensão e reorganização do tesauro Wordnet ([87] apud [54]). O artigo indica que o sistema Ontoseek, bem como outros sistemas baseados em ontologias linguísticas, podem ser utilizados em aplicações multilíngues. Dado que as ontologias línguísticas trabalham com conceitos e são independentes do idioma (léxico) então, uma vez que as aplicações indexem o conceito, basta indicar os idiomas associados à ontologia de conceitos e o sistema pode lidar com consultas baseadas em idiomas distintos.

Guerreiros de Terracota

Este sistema de recuperação de informação [115, 116] é aplicado na recuperação de imagens de guerreiros de terracota do Primeiro Imperador da China. Trata-se de um sistema de recuperação de imagens com anotação semântica. Neste sistema a descrição de cada imagem é analisada e, através de uma ontologia de domínio e um tesauro de Chinês Mandarin, é automaticamente transformada em anotação semântica associada à respectiva imagem. Em seguida o conjunto de imagens é indexado automaticamente considerando a anotação semântica associada.

Para fazer a recuperação, o usuário especifica uma consulta em linguagem natural. Esta consulta é submetida a um *parser* e, através da mesma ontologia e tesauro, ela é transformada em um esquema que traduz a sua semântica. Em seguida o sistema faz a equiparação da consulta com cada uma das imagens indexadas computando o grau de similaridade entre elas através de uma heurística. As imagens que apresentarem maior similaridade são retornadas para o usuário. Nos experimentos realizados considerou-se um conjunto de 49 imagens e 30 consultas. Foi observada uma melhora na medida de precisão para uma mesma avaliação da medida de cobertura quando comparado com a busca por palavras-chaves.

FROM

O sistema FROM [100, 101, 102] implementa o modelo ontológico relacional *fuzzy* para recuperação de informação textual. O sistema faz a expansão da consulta considerando as relações existentes

na estrutura conceitual. Para tanto ele utiliza uma ontologia de dois níveis compostos por categorias e palavras-chaves que representam os conceitos do domínio. As categorias denotam os conceitos mais gerais e as palavras-chaves denotam os conceitos mais específicos. As palavras-chaves e as categorias estão associados por relações *fuzzy* que determinam o grau de associação entre elas. Não existem associações entre duas palavras-chaves ou duas categorias. Os documentos estão associados tanto às categorias como às palavras chaves de forma independente sendo que o grau de associação entre documentos e palavras-chaves ou categorias também é dado por relações *fuzzy*.

Uma consulta do usuário pode ser composta apenas por palavras-chaves, por categorias ou por ambas. Quando o usuário faz uma consulta, o sistema utiliza o conhecimento existente na ontologia para fazer a expansão da mesma. Ao expandir a consulta o sistema pode adicionar novas categorias e palavras-chaves, em função das conexões existentes na ontologia, pois as categorias e palavras-chaves fornecidas pelo usuário podem estar associadas a outras, que não faziam parte da consulta original, mas que são consideradas importantes ao se analisar as relações entre elas na ontologia. Em seguida a consulta é submetida ao sistema sendo calculada a similaridade dos documentos em relação à consulta através de operações *fuzzy*. Os documentos que apresentarem a medida de similaridade acima de um limite estabelecido são retornados ao usuário

O sistema foi testado utilizando uma base de 100 documentos na área de Inteligência Computacional classificados em uma estrutura composta por 61 termos (6 categorias e 55 palavras-chaves) organizados por um especialista do domínio. Nos testes realizados o sistema apresentou uma melhora na precisão, para uma mesma taxa de cobertura, quando comparado com outros algoritmos *fuzzy* clássicos de recuperação de informação ([56], [93] apud [101]).

Índice Geográfico

Muitos documentos armazenados em bibliotecas digitais ou bancos de documentos incluem referências geográficas nos seus textos. Em geral, os sistemas de índices ou algoritmos de recuperação de informação não consideram a natureza espacial das referências geográficas existentes nos documentos. Técnicas de busca textuais focam no léxico dos documentos e técnicas de busca espaciais focam nos aspectos geográficos dos documentos. Nenhuma delas é totalmente adequada para um enfoque conjunto para recuperação de informação pois uma técnica negligencia a outra. O sistema Índice Geográfico [83, 84] considera as características destes dois tipos de informação para recuperação de documentos. Para tanto ele é composto de duas estruturas de índices: uma ontologia espacial que modela a organização espacial de lugares e um índice invertido para indexar as palavras nos documentos para busca textual.

Na ontologia, as classes representam os lugares e elas são conectadas baseado na relação de inclusão espacial entre os lugares que elas representam. Por exemplo, a Galícia é incluída na Espanha então

existe uma conexão entre as classes que representam estes lugares. As informações armazenadas na ontologia são o nome do lugar, as referências geográficas associadas, sua representação geométrica, uma lista dos documentos associados ao lugar e os lugares relacionados por inclusão espacial.

O tipo de organização de conhecimento proposto suporta consultas puramente textuais pelo uso de palavras-chave, consultas puramente espaciais pela indicação de uma região geográfica em um mapa, consultas textuais com nomes de lugares e consultas textuais que se aplicam a uma área geográfica.

Quando as consultas utilizam informação de lugares geográficos então a ontologia espacial é utilizada para extrair conhecimento relativo à região geográfica citada na consulta. Este conhecimento é utilizado para expandir a consulta do usuário. O sistema foi testado utilizando a coleção de documentos dada pela TREC (*Text REtrieval Conference*) FT-91 (*Financial Times*, ano de 1991). O TREC é um forum internacional que possui coleções de documentos para avaliação de sistemas de recuperação de informação. Nos experimentos realizados foi medido apenas o tempo de resposta das consultas. Mais experimentos estão sendo conduzidos para determinar as medidas de cobertura e precisão. Trabalhos futuros incluem o uso de ontologias diferentes para compor o índice e a definição de outros tipos de relacionamentos entre as localizações geográficas na ontologia, como por exemplo o relacionamento de adjacência, além do relacionamento de inclusão já existente.

Mineração de Ontologia

O modelo de Mineração de Ontologia [72] procura construir uma ontologia *fuzzy* a partir da mineração de textos e aplicar as relações na ontologia expandir a consulta do usuário. Uma vez que as relações de taxonomia, descobertas por meio de mineração de texto, sempre envolvem incertezas então uma ontologia *fuzzy* é construída pelo método proposto. Neste trabalho uma ontologia é definida pela quádrupla $Ont = < X,\ C,\ R_{XC},\ R_{CC} >$ onde X é o conjunto de termos e C é o conjunto de conceitos. A relação *fuzzy* $R_{XC} : X \times C \mapsto [0,1]$ mapeia o conjunto de termos ao conjunto de conceitos associando os valores de pertinência e a relação *fuzzy* $R_{CC} : C \times C \mapsto [0,1]$ denota as relações de taxonomia *fuzzy* entre o conjunto de conceitos C. Utilizando combinações léxico-sintáticas e técnicas de aprendizado estatístico os conjuntos X, C e as relações R_{XC} e R_{CC} são mineradas a partir dos textos dos documentos da coleção.

A ontologia *fuzzy* é utilizada no processo de expansão da consulta onde cada termo da consulta inicial é expandido para considerar o termo equivalente, o mais geral e o mais específico. O modelo proposto é testado utilizando o subconjunto Lewis-Split da coleção Reuters-21578 que contém 19.813 documentos. Este conjunto de documentos é parte da TREC. Inicialmente foi gerada a ontologia *fuzzy* para a coleção e em seguida foram testadas 15 consultas. O sistema mostrou uma melhora na precisão para as mesmas taxas de cobertura quando comparado com a consulta utilizando apenas as palavras-chaves.

Portal Semântico

O Portal Semântico [132] integra capacidade de recuperação de informação textual e técnicas de inferência e especificação de consulta, utilizando lógica descritiva, para recuperar tanto a informação textual quanto a informação semântica disponível no portal. A informação do portal é organizada como uma ontologia de domínio e armazenada em uma base de conhecimento. Enquanto os métodos de recuperação de informação calculam o grau de relevância de um documento para uma consulta, os métodos formais e as máquinas de inferência, baseados em lógica descritiva, fazem um julgamento binário. Para solucionar este problema o modelo propõe o uso de uma lógica descritiva *fuzzy* ([122] apud [132]). O sistema permite tanto estabelecer condições lógicas, para construir uma consulta para fazer a recuperação na base de conhecimento, como utilizar os métodos de recuperação de informação para recuperar as descrições textuais de forma conjunta. Ao utilizar a lógica descritiva *fuzzy* os resultados vão possuir um grau de importância associado permitindo estabelecer uma ordenação para apresentar os mais relevantes para o usuário.

Algumas limitações do modelo são a possível dificuldade do usuário em montar a consulta como uma forma lógica e a disponibilização de uma máquina capaz de realizar inferências considerando uma lógica *fuzzy*. O modelo proposto é testado pela implementação do sistema SportS (*Semantic+Portal+Service*) [77] que trata das publicações disponíveis na Intranet do Departamento de Ciência da Computação e Engenharia da Universidade de Shangai Jiao Tong. Para fazer a recuperação de informação textual utilizou-se a máquina de busca Apache Lucene e para as buscas envolvendo expressões lógicas a máquina de inferência alc-F, que se encontra em um estágio inicial de desenvolvimento. Duas consultas foram realizadas e um grupo de estudantes do laboratório fizeram a validação dos resultados, retornados pelo sistema, para coletar as medidas de cobertura e precisão. As consultas foram especificadas utilizando tanto o modelo de lógica descritiva *fuzzy* proposto no portal semântico quanto uma combinação simplificada de lógica descritiva e recuperação textual. Os resultados obtidos com as consultas especificadas no modelo do portal semântico obtiveram resultados de precisão melhores para os mesmos valores de cobertura.

3.4 Classificação dos Sistemas de Recuperação de Informação Semântica

Nesta seção é apresentado um levantamento dos critérios observados nos sistemas estudados com relação ao uso de estruturas conceituais na recuperação de informação. O objetivo é levantar os processos de recuperação de informação onde as estruturas conceituais são utilizadas, os ganhos obtidos com o seu uso e os possíveis pontos que podem ser melhorados. Os trabalhos estudados

serão classificados nestes critérios. Nas seções anteriores observou-se que os trabalhos apresentados recuperam documentos, imagens ou vídeos. Aqui usar-se-á apenas a denominação recursos para referenciar os itens constituintes de uma coleção. Os critérios levantados são:

1. Quantidade de estruturas de conhecimento utilizadas: a quantidade de estruturas de conhecimento refere-se ao número de estruturas conceituais utilizadas para tratar a semântica do domínio. Uma vez que uma coleção possa tratar assuntos pertencentes a domínios de conhecimento distintos pode ser necessário que seus recursos sejam indexados em diferentes estruturas conceituais. Este critério procura identificar esta característica e verificar se as aplicações lidam com mais de uma estrutura conceitual.

2. Fases do processo de recuperação em que o conhecimento foi utilizado: o objetivo deste critério é identificar possibilidades de uso de estruturas conceituais nas diversas fases do processo de recuperação de informação.

3. Formas de avaliação dos sistemas: neste critério procura-se levantar como avaliar um sistema que utiliza estruturas conceituais e qual foi o ganho obtido pela utilização das mesmas.

3.4.1 Quantidade de Estruturas de Conhecimento Utilizadas

Neste tópico é verificado o número de estruturas conceituais manipuladas pelos sistemas. A classificação dos sistemas, pelo número de estruturas conceituais, é ilustrada na Fig. 3.2

Uso de Uma Única Estrutura de Conhecimento

O sistema DOPE utiliza o tesauro EMTREE da Elsevier para indexar recursos e expandir as consultas. O sistema Deja Vu utiliza o tesauro da Biblioteca do Congresso para Material Gráfico para indexar os recursos e especificar o termo da consulta. O sistema Phind utiliza o tesauro AGROVOC da FAO para que o usuário possa selecionar termos relacionados ao termo inicial para refazer a consulta. O sistema OnAir utiliza uma ontologia sobre arte contemporânea para indexar os recursos e para expandir a consulta. O sistema OWLIR utiliza uma ontologia de eventos da universidade de Dakota do Norte para anotar os documentos e as consultas. O sistema OntoSeek utiliza a ontologia Sensus para montar os grafos conceituais que representam os recursos e as consultas. O sistema FROM utiliza uma ontologia na área de Inteligência Computacional para indexar os recursos e expandir a consulta. O sistema de Rede de Conceitos *Fuzzy* utiliza uma rede de conceitos desenvolvida pelo especialista do domínio para indexar os recursos. No sistema CIRI o usuário seleciona uma ontologia para compor uma consulta e o sistema expande a consulta com os conceitos da mesma ontologia. No sistema Tesauro Automático um tesauro é construído automaticamente a partir de uma coleção de

documentos sobre um domínio especfífico. No Tesauro Sócio-Político um tesauro é construído nas línguas russa e inglesa. No Índice Geográfico uma ontologia espacial modela a organização espacial de lugares. O modelo Mineração de Ontologia constrói uma ontologia *fuzzy* a partir da mineração de textos. O Portal Semântico utiliza uma ontologia de domínio para descrever a semântica do portal. O sistema de Navegação Personalizada utiliza uma ontologia de domínio para construir as facetas personalizadas para o usuário.

Uso de Múltiplas Estruturas de Conhecimento

O sistema DocCube utiliza várias hierarquias de conceito ou ontologias criadas em função do domínio de aplicação. Cada hierarquia vai constituir em uma faceta onde os recursos serão classificados e as facetas vão constituir os eixos do espaço de navegação semântico. O sistema para Coleção de Artes utiliza hierarquias como facetas para classificar os recursos. O sistema para Guerreiros de Terracota utiliza um tesauro de chinês mandarin e uma ontologia de domínio para anotar os recursos e a especificar a consulta. O sistema Gerenciamento de Fragmentos de Documentos utiliza uma estrutura distinta para cada dimensão na qual os fragmentos de documentos podem estar relacionados.

A maioria dos sistemas utiliza apenas uma estrutura conceitual para modelar o conhecimento. Em geral, os sistemas que utilizam mais de uma estrutura conceitual o fazem em função da indexação por facetas. Apesar de sistemas como DOPE e Deja Vu apontarem a necessidade de considerar mais de uma estrutura conceitual na indexação dos documentos, a maioria dos sistemas ainda não suporta este tipo de funcionalidade.

3.4.2 Fases Onde Ocorre a Exploração da Semântica

Neste tópico é tratado em que fase do processo de recuperação de informação o sistema utiliza a semântica relacionada ao conhecimento expresso nos conceitos e relações da estrutura conceitual.

Indexação

Neste trabalho a indexação conceitual dos recursos consiste em associar os conceitos da estrutura conceitual aos mesmos. Esta associação pode ser feita considerando os conceitos como índices dos recursos. Isto ocorre nos sistemas: Dope, Deja Vu, CIRI, FROM, Rede de Conceitos *Fuzzy*, Tesauro Automático e Tesauro Sócio-Político. O sistema OnAir indexa os recursos por seu texto completo mas quando um termo de indexação expressa um conceito presente na estrutura conceitual este termo recebe um tratamento especial. Os sistemas OWLIR, Guerreiros de Terracota e Fragmentos de Documentos inserem anotação semântica nos recursos e posteriormente indexam a anotação semântica. No sistema Ontoseek não ocorre uma indexação, os conceitos são associados nos grafos conceituais

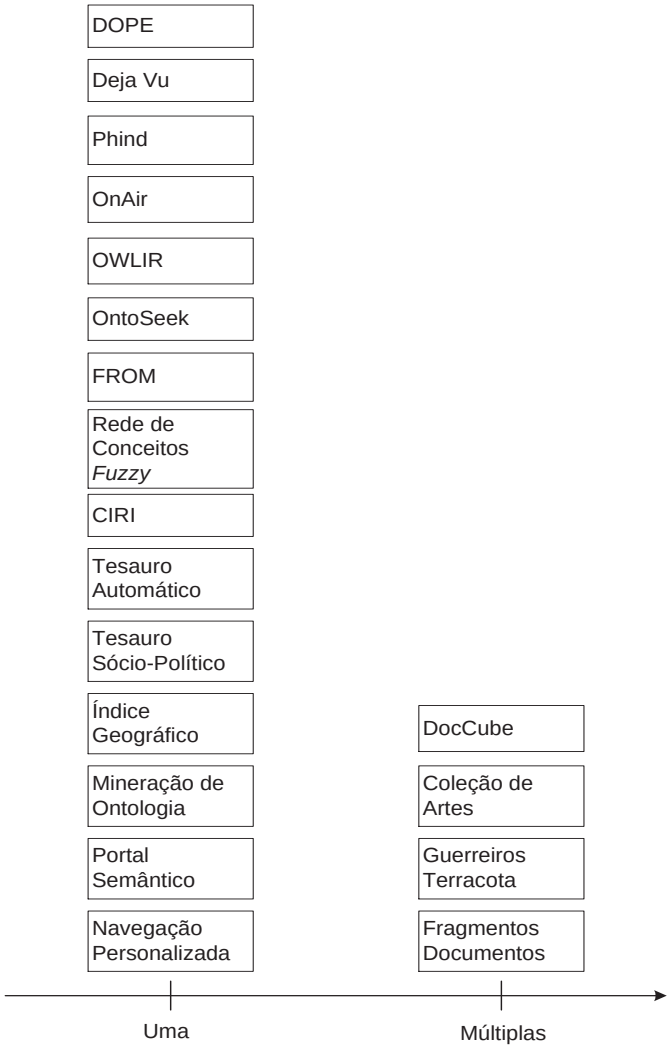

Fig. 3.2: Quantidade de estruturas de conhecimento utilizadas.

que representam os produtos de catálogos de produtos e páginas amarelas. Nos sistemas DocCube, Coleção de Artes e Navegação Personalizada a estrutura conceitual é um conjunto de hierarquias e cada hierarquia é vista como uma faceta. Os recursos são associados aos conceitos das facetas. O sistema Phind associa uma lista de documentos para cada palavra e frase extraída nos respectivos documentos. As palavras e frases extraídas podem estar presentes na estrutura conceitual. O sistema Índice Geográfico associa uma lista de documentos para cada conceito da ontologia espacial. O modelo Mineração de Ontologia minera os termos e conceitos da ontologia a partir do texto. Desta forma os documentos são indexados pelos termos e conceitos da ontologia. O Portal Semântico indexa tanto os documentos quanto os conceitos e indivíduos da ontologia de domínio pelos termos contidos na sua descrição textual.

Especificação da Consulta

Para consultar os recursos da coleção alguns sistemas possuem uma interface de navegação na estrutura conceitual ou oferecem a possibilidade de entrada de palavras chave ou ambas. Como os documentos estão, através de algum mecanismo de indexação, associados aos conceitos da estrutura conceitual então ocorre a recuperação semântica. A Fig. 3.3 mostra a organização dos sistemas com relação às formas de especificação de consulta utilizada. Nos sistemas DOPE, CIRI, Deja Vu, OntoSeek, Coleção de Artes e Tesauro Sócio-Político, o usuário pode selecionar os conceitos na estrutura conceitual para montar a consulta. Nos demais sistemas são utilizadas outras formas para especificação da consulta como expressão em linguagem natural, seleção de conceitos nas facetas, palavras-chave, expressão em lógica, texto estruturado ou seleção de uma região geográfica em um mapa. O sistema Tesauro Automático não apresentou como a consulta foi especificada.

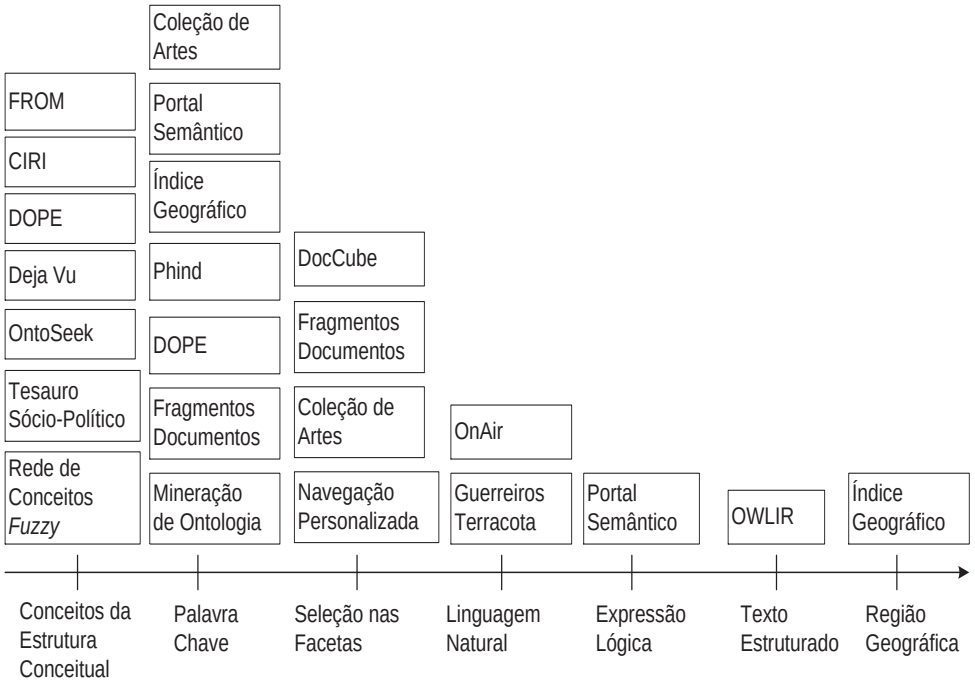

Fig. 3.3: Formas de especificação da consulta.

Em alguns sistemas, antes da consulta ser submetida, ela é expandida com outros conceitos da estrutura conceitual através de suas relações de hierarquia e de associação. Os sistemas OnAir, CIRI, OWLIR, DOPE, FROM, Tesauro Automático, Tesauro Sócio-Político, Índice Geográfico e Mineração de Ontologia realizam a expansão da consulta. A Fig. 3.4 mostra a organização dos sistemas com relação à expansão da consulta. Deve-se lembrar que a representação dos documentos e da consulta estão intimamente ligados ao modelo do processo de recuperação de informação para que possa ser calculada a similaridade entre um documento e a consulta.

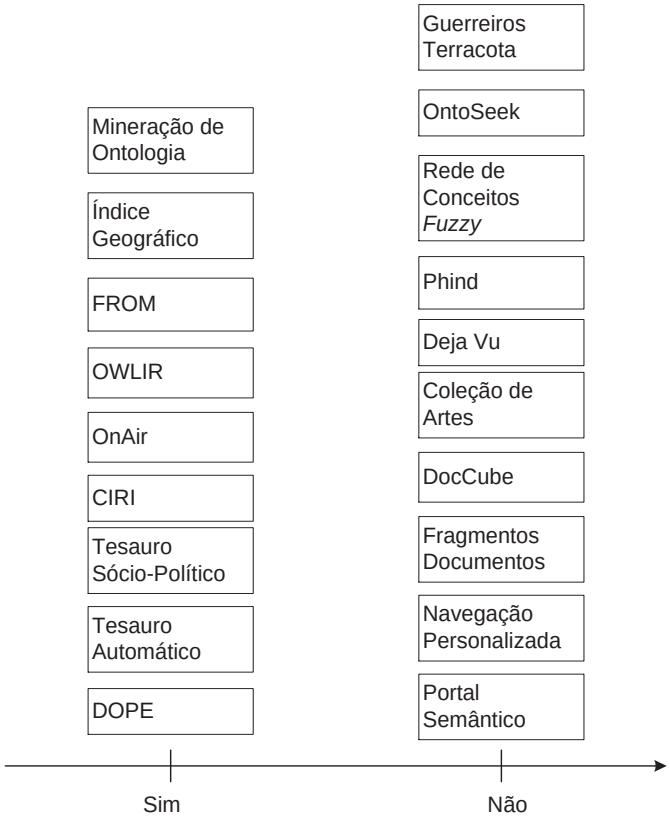

Fig. 3.4: Uso de expansão da consulta.

Forma de Apresentação de Resultados

Alguns sistemas apresentam os resultados organizados pelos conceitos da estrutura conceitual para facilitar a visualização dos mesmos e para mostrar a sua distribuição pelos conceitos.

Sistemas que apresentam os resultados organizados pelas facetas são o DocCube, Coleção de Artes, Navegação Personalizada e Fragmentos de Documentos. No sistema DOPE há a preocupação em apresentar os resultados organizados pelos conceitos aos quais eles estão associados. Os recursos aparecem como agrupamentos que se distribuem entre os conceitos. Para visualizar os recursos deve-se selecionar o agrupamento desejado. O sistema Deja Vu mostra uma lista de documentos associados aos conceitos da estrutura conceitual. Os sistemas OnAir, OWLIR, Phind, CIRI, OntoSeek, Guerreiros de Terracota, FROM, Rede de Conceitos *Fuzzy*, Tesauro Automático, Tesauro Sócio-Político, Mineração de Ontologia e Portal Semântico apresentam uma lista dos recursos ordenados por um critério de relevância. O sistema Índice Geográfico pretende apresentar uma lista dos documentos recuperados por um critério de relevância mas como ele permite executar uma busca por meio de seleção de uma região geográfica em um mapa então é necessário, inicialmente, definir o critério de relevância para ordenação dos documentos recuperados neste tipo de busca.

3.4.3 Formas de Avaliação dos Sistemas

A avaliação dos sistemas de recuperação pode ser realizada através das medidas de precisão e cobertura ou pela validação junto ao usuário. Dentre os sistemas estudados, os sistemas OnAir, OWLIR, Guerreiros de Terracota, FROM, Tesauro Automático, Tesauro Sócio-Político e Mineração de Ontologia realizaram testes para verificar os ganhos obtidos através das medidas de precisão x cobertura. Os sistemas Deja Vu, DOPE, CIRI, Coleção de Artes, Navegação Personalizada e Portal Semântico fizeram uma validação junto ao usuário. Os demais sistemas não deixaram explícito.

Melhoria das Medidas de Precisão e Cobertura

Os sistemas OnAir, FROM, CIRI, Tesauro Automático, Tesauro Sócio-Político, Mineração ontologia e Índice geográfico utilizaram a estrutura conceitual para realizar a expansão da consulta. Os sistemas Guerreiros de Terracota e OWLIR utilizaram o conhecimento da estrutura conceitual para fazer anotação semântica nas páginas dos documentos. A seguir é apresentado uma discussão sobre a particularidade de cada um dos sistemas no que diz respeito ao seu desempenho considerando as medidas de precisão e cobertura.

O sistema OnAir apresentou que a expansão da consulta, utilizando o conhecimento de uma ontologia de domínio, melhorou em média a performance do sistema no que se refere tanto à precisão como à cobertura. O sistema apresentado em Guerreiros de Terracota indicou que ao anotar as páginas com informação semântica extraída da ontologia de domínio e do tesauro de chinês mandarim ocorreu uma melhora na precisão para um mesmo valor na cobertura. Ambos os sistemas compararam o resultado com a busca por palavras-chave. O sistema OWLIR conseguiu obter melhoria na precisão, para uma mesma taxa de cobertura, ao anotar as páginas com o conhecimento de uma ontologia e ao acrescentar conhecimento obtido pela inferência no conhecimento da ontologia a precisão aumentou ainda mais. A comparação foi feita com a busca realizada em textos sem a anotação semântica. O sistema FROM obteve, no geral, uma melhora na precisão para uma mesma taxa de cobertura, quando comparado com outros algoritmos *fuzzy* para recuperação de informação. Na validação do sistema CIRI não houve ganho nas taxas de precisão e cobertura quando comparado com os resultados do sistema auxiliar baseado em palavras-chave. No sistema Tesauro Automático o tesauro proposto é construído a partir de outros tesauros pré-existentes e nas consultas expandidas com os conceitos do tesauro construído obteve-se melhores taxas de precisão quando comparado com os resultados das consultas expandidas com os conceitos dos tesauros pré-existentes. O sistema Tesauro Sócio-Político obteve melhoria das taxas de precisão para uma mesma taxa de cobertura e seus resultados ficaram entre os quatro melhores quando comparado com os resultados obtidos por outros tesauros considerando as mesmas consultas e coleção de documentos. O sistema Índice Geográfico vai realizar

experimentos para determinar as medidas de precisão e cobertura dos resultados obtidos. O sistema Mineração de Ontologia mostrou uma melhora na precisão para as mesmas taxas de cobertura quando comparado com a consulta utilizando apenas as palavras-chaves. As consultas especificadas utilizando o modelo de lógica descritiva *fuzzy* proposto no sistema Portal Semântico obtiveram resultados de precisão melhores, para os mesmos valores de cobertura, quando comparadas com consultas especificadas com uma combinação simplificada de lógica descritiva e recuperação textual. Em geral, os sistemas que utilizaram o conhecimento da estrutura conceitual na expansão da consulta ou na anotação dos recursos conseguiram um ganho na precisão.

Validação do Sistema Junto ao Usuário

Os sistemas que realizaram a validação junto ao usuário são aqueles que utilizam os conceitos da estrutura conceitual para organizar a coleção de documentos. O usuário pode navegar na estrutura conceitual para selecionar os conceitos que vão compor a consulta. Os documentos são apresentados ao usuário organizados em função da estrutura conceitual.

Para os sistemas que utilizam a estrutura conceitual, tanto para selecionar os conceitos das consultas como para apresentar os resultados, a expectativa é de que o usuário tenha maior facilidade de entendimento dos assuntos tratados na coleção. Mesmo os sistemas que não fizeram a validação defendem que o uso de uma estrutura conceitual ajuda no entendimento dos temas tratados nos documentos.

Os sistemas que apresentam a estrutura conceitual, na interface, para dar ao usuário uma visão do domínio para ajudar na especificação da consulta são: DOPE, DocCube, Deja Vu, CIRI, Phind, OntoSeek, Navegação Personalizada, Fragmentos de Documentos e o sistema que trata da Coleção de Artes. Para estes sistemas o usuário sempre tem acesso ao contexto semântico da consulta a ser realizada, de uma forma intuitiva, através da integração e apresentação das estruturas de representação do conhecimento na interface. No sistema Deja Vu um dos pontos positivos, segundo os catalogadores, foi a possibilidade de integração da informação de catalogação com os recursos da coleção através da visualização dos termos do tesauro. Ao navegar no tesauro o usuário já tem a visualização dos temas disponíveis na coleção. O processo de navegar pelo tesauro aumenta o entendimento do usuário sobre o relacionamento entre os recursos e o material de catalogação. No sistema que trata da Coleção de Artes os usuários indicaram que a navegação na interface dava uma idéia geral do que buscar e alguns se sentiram mais confiantes nos resultados obtidos na navegação. Na validação do sistema CIRI os usuários indicaram a necessidade de se habituar ao uso da interface. No sistema Navegação Personalizada o número de clicks do usuário aumentou uma vez que nem sempre as facetas corretas estavam ativas sendo necessário habilitá-las manualmente. O uso de restrições nos atributos reduziram o tempo de busca e o número de clicks permitindo que a busca fosse mais especializada.

3.5 Resumo do Capítulo

Neste capítulo foi apresentada uma relação de trabalhos que tratam da recuperação de recursos baseada na informação semântica neles expressa. Para isto eles empregam estruturas conceituais que são utilizadas para tarefas como indexar os recursos, expandir a consulta, possibilitar a navegação no conhecimento do domínio, expresso na estrutura, e anotar os recursos com informação semântica, considerando as relações entre os conceitos da estrutura.

De acordo com o estudo realizado pode-se concluir que ao utilizar as relações entre os conceitos da estrutura conceitual na expansão da consulta ou para adicionar mais informação semântica na anotação do recurso ocorreu, em geral, uma melhora na qualidade da informação recuperada representada pelo aumento da precisão. A maioria dos sistemas utiliza apenas uma estrutura conceitual para modelar o conhecimento. Em geral, os sistemas que utilizam mais de uma estrutura conceitual o fazem em função da indexação por facetas. Nenhum dos sistemas considera que as estruturas conceituais distintas podem estar relacionadas entre elas. Apesar de sistemas como DOPE e Deja Vu apontarem a necessidade de considerar mais de uma estrutura conceitual na indexação dos documentos, a maioria dos sistemas ainda não suporta este tipo de funcionalidade. O uso de mais de uma estrutura de conhecimento deve ser explorado pois uma coleção de documentos pode tratar temas pertencentes a domínios diferentes que vão ser expressos em estruturas conceituais distintas.

No que se refere à apresentação de resultados para o usuário, a possibilidade de apresentar os recursos associados ao conhecimento existente na estrutura conceitual é interessante para dar uma visão de onde os recursos se localizam dentro do domínio de conhecimento. A navegação no domínio expresso na estrutura conceitual permite ao usuário entender o espaço de busca e a perceber os conceitos tratados na coleção ajudando-o a especificar melhor a consulta.

Capítulo 4

Modelo *Fuzzy* Utilizando Múltiplas Ontologias Relacionadas

No estudo dos trabalhos que tratam de recuperação de informação, realizado no Cap. 3, concluiu-se que o uso de mais de uma estrutura de conhecimento para indexação e recuperação de informação deve ser explorado pois uma coleção de documentos pode tratar temas pertencentes a domínios diferentes que vão ser expressos em estruturas conceituais distintas. Em geral, os sistemas que utilizam mais de uma estrutura conceitual o fazem em função da indexação por facetas. Nenhum dos sistemas considera que as estruturas conceituais distintas podem estar relacionadas entre elas. Além disto, concluiu-se também que ao utilizar as relações entre os conceitos da estrutura conceitual, para adicionar informação semântica na expansão da consulta, ocorreu uma melhora na qualidade da informação recuperada.

Com base nestas considerações este capítulo apresenta o modelo *fuzzy* que explora o uso de ontologias distintas e relacionadas para indexação e recuperação de documentos de uma coleção. O modelo considera que cada ontologia representa um domínio de conhecimento sendo que as ontologias podem estar relacionadas entre si através de relações *fuzzy*.

O modelo apresentado é baseado nos trabalhos Rede de Conceitos *Fuzzy* e FROM, apresentados no Cap. 3, que empregam a teoria de conjuntos *fuzzy* para modelar bases de conhecimento utilizadas para melhorar a qualidade da informação recuperada no processo de recuperação de informação. O modelo proposto nesta tese estende estes dois trabalhos considerando que o conhecimento pode estar expresso por múltiplas ontologias onde cada ontologia possui relações de especialização e generalização *fuzzy* entre seus conceitos e que as ontologias podem estar relacionadas entre si e através de relações de associação positiva *fuzzy*.

4.1 Teoria de Conjuntos *Fuzzy*

Esta seção apresenta os conceitos da teoria de conjuntos *fuzzy* [69, 70, 99, 100] utilizados para representar bases de conhecimento por meio de relações *fuzzy*. Estes conceitos são úteis no entendimento do modelo de recuperação de informação *fuzzy* apresentado na próxima seção.

4.1.1 Conjuntos *Fuzzy*

Uma das principais motivações para a introdução de conjuntos *fuzzy* é a representação de conceitos imprecisos. O grau de pertinência de um elemento em um conjunto *fuzzy* expressa o grau de compatibilidade do elemento com o conceito representado pelo conjunto. Um conjunto *fuzzy* A, sobre o conjunto universal X, é definido por uma função de pertinência que atribui para cada elemento x de X um número $A(x) \in [0,\ 1]$, ou seja, $A:\ X \rightarrow [0,\ 1]$.

Se o universo X é discreto e finito, com cardinalidade n, então o conjunto *fuzzy* A é representado na forma de um vetor de dimensão n onde as entradas do vetor correspondem aos graus de pertinência dos elementos correspondentes de X. A notação de somatório pode ser utilizada para esta representação. Ela permite representar os elementos de X que possuem graus de pertinência diferentes de zero. Por exemplo, se $X = \{x_1,\ x_2,\ x_3, x_1, \ldots, x_n\}$ então o conjunto *fuzzy* $A = \{(a_i/x_i)\,|x_i \in X\}$, onde $a_i = A\,(x_i)$ e $i = 1,\ \ldots,\ n$ é dada por:

$$A = a_1/x_1\ +\ a_2/x_2\ +\ \ldots\ a_n/x_n = \sum_{i=1}^{n} a_i/x_i$$

Nesta representação o símbolo $\sum$ não deve ser confundido com o somatório algébrico. O uso deste símbolo é para denotar o conjunto de pares ordenados. O conjunto A também pode ser representado como um vetor $A = [a_1\ a_2\ \ldots\ a_n]$. Seja o universo $X = \{1, 2, 3, 4, 5\}$ e os conjuntos definidos em X, $A = 0.1/1\ +\ 0.3/2\ +\ 0.5/3\ +\ 0.8/4\ +\ 0.9/5$ e $B = 0.2/1\ +\ 0.6/2\ +\ 0.7/3\ +\ 0.4/4\ +\ 0.5/5$. A representação, por vetor, dos conjuntos é dada por $A = [0.1\ 0.3\ 0.5\ 0.8\ 0.9]$ e $B = [0.2\ 0.6\ 0.7\ 0.4\ 0.5]$

Interseção de Conjuntos *Fuzzy*

Considerando dois conjuntos *fuzzy* A e B, sobre o universo X, a operação de interseção definida nestes conjuntos é dada por :

$$(A \cap B)(x)\ \ =\ \ \min(A(x),\ B(x))\ \ =\ \ A(x)\ \wedge\ B(x);\ \ \forall x \in X$$

$$A(x) \wedge B(x) = \begin{cases} A(x) & se \quad A(x) \leq B(x) \\ B(x) & se \quad A(x) > B(x) \end{cases}$$

A interseção dos conjuntos *fuzzy* A e B representados pelos vetores $A = [0.1 \ 0.3 \ 0.5 \ 0.8 \ 0.9]$ e $B = [0.2 \ 0.6 \ 0.7 \ 0.4 \ 0.5]$ resulta em $A(x) \wedge B(x) = [0.1 \ 0.3 \ 0.5 \ 0.4 \ 0.5]$

União de Conjuntos *Fuzzy*

Dados dois conjuntos *fuzzy* A e B, definidos no universo X, a operação de união definida nestes conjuntos é dada por :

$$(A \cup B)(x) \quad = \quad \max(A(x), \, B(x)) \quad = \quad A(x) \vee B(x); \quad \forall x \in X$$

$$A(x) \vee B(x) = \begin{cases} A(x) & se \quad A(x) \geq B(x) \\ B(x) & se \quad A(x) < B(x) \end{cases}$$

A união dos conjuntos *fuzzy* A e B representados pelos vetores $A = [0.1 \ 0.3 \ 0.5 \ 0.8 \ 0.9]$ e $B = [0.2 \ 0.6 \ 0.7 \ 0.4 \ 0.5]$ resulta em $A(x) \vee B(x) = [0.2 \ 0.6 \ 0.7 \ 0.8 \ 0.9]$

α-Cut

Dado um conjunto *fuzzy* A, o conjunto α-cut de A, A_α, é um conjunto dos elementos do universo X cujos valores de pertinência excedem o limite dado por α, ou seja, $A_\alpha = \{x \in X \mid A(x) \geq \alpha\}$. Dado o conjunto $A = [0.1 \ 0.3 \ 0.5 \ 0.8 \ 0.9]$, o conjunto α-cut de A, para $\alpha = 0.5$, é $A_{0.5} = [0 \ 0 \ 0.5 \ 0.8 \ 0.9]$

4.1.2 Relações *Fuzzy*

Sejam X e Y dois universos. Uma relação clássica definida em $X \times Y$ é qualquer subconjunto do produto cartesiano destes dois universos, $R: \ X \times Y \rightarrow \{0, \ 1\}$. Se o valor de $R(x, y) = 1$ indica que os dois elementos estão relacionados pela relação R. Caso contrário, $R(x, y) = 0$, indica que os dois elementos não estão relacionados. As relações *fuzzy* generalizam o conceito da relação clássica admitindo a noção de pertinência parcial entre os elementos dos dois universos de discurso, ou seja, $R : X \times Y \rightarrow [0, \ 1]$. A forma mais comum de representar relações *fuzzy* em universos finitos é através de matrizes cujos elementos correspondem aos graus de pertinência entre os elementos dos universos.

Dada uma relação *fuzzy* binária $R : X \times Y$ onde $X = \{x_1,\ x_2,\ \cdots, x_m\}$ e $Y = \{y_1,\ y_2,\ \cdots, y_n\}$ a representação matricial desta relação é uma matriz $m \times n$ como segue:

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{n2} & \cdots & r_{mn} \end{pmatrix}$$

Nesta matriz $r_{ij} = R(x_i,\ y_j) \in [0,\ 1]$, $1 \leq i \leq m$ e $1 \leq j \leq n$ representa o grau com que o elemento x_i está associado ao elemento y_j pela relação R.

As definições das operações de interseção e união para as relações *fuzzy* são semelhantes às operações correspondentes nos conjuntos *fuzzy*. Dadas duas relações *fuzzy* R e W definidas no universo $X \times Y$ tem-se que, $\forall x \in X$ e $\forall y \in Y$:

$$\begin{aligned} (R \cap W)(x,\ y) &= \min\left[R(x,\ y),\ W(x,\ y)\right] \\ (R \cup W)(x,\ y) &= \max\left[R(x,\ y),\ W(x,\ y)\right] \end{aligned}$$

A composição de duas relações *fuzzy* P em $X \times Y$ e Q em $Y \times Z$ é a relação *fuzzy* R em $X \times Z$ dada por:

$$R(x,\ z) = (P \circ Q)(x,\ z) = \max_{y \in Y} \min\left[P(x,\ y),\ Q(y,\ z)\right]$$

Para ilustrar o cálculo da composição de relações *fuzzy* sejam os universos de discurso $X = \{p_1, p_2, p_3, p_4,\}$ contendo o conjunto de pacientes, $Y = \{s_1, s_2, s_3\}$ o conjunto de sintomas e $Z = \{d_1, d_2, d_3, d_4, d_5\}$ o conjunto de doenças.

A relação *fuzzy* $P :\ X \times Y$ indica quão fortes são as manifestações dos sintomas nos pacientes. A relação *fuzzy* $Q :\ Y \times Z$ indica quão fortes os sintomas estão associados às doenças. As relações P e Q são representadas por:

$$P = \begin{array}{c} \\ p_1 \\ p_2 \\ p_2 \\ p_4 \end{array} \begin{array}{c} \begin{array}{ccc} s_1 & s_2 & s_3 \end{array} \\ \begin{pmatrix} 0 & 0.3 & 0.4 \\ 0.2 & 0.5 & 0.3 \\ 0.8 & 0 & 0 \\ 0.7 & 0.7 & 0.9 \end{pmatrix} \end{array} \qquad Q = \begin{array}{c} \\ s_1 \\ s_2 \\ s_2 \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & d_4 & d_5 \end{array} \\ \begin{pmatrix} 0.7 & 0 & 0 & 0.3 & 0.6 \\ 0.5 & 0.5 & 0.8 & 0.4 & 0 \\ 0 & 0.7 & 0.2 & 0.9 & 0 \end{pmatrix} \end{array}$$

Executando a composição entre as relações *fuzzy* P e Q tem-se a relação *fuzzy* $R = P \circ Q$ que expressa a associação entre pacientes e doenças. A representação da composição é mostrada a seguir:

$$\begin{pmatrix} 0 & 0.3 & 0.4 \\ 0.2 & 0.5 & 0.3 \\ 0.8 & 0 & 0 \\ 0.7 & 0.7 & 0.9 \end{pmatrix} \circ \begin{pmatrix} 0.7 & 0 & 0 & 0.3 & 0.6 \\ 0.5 & 0.5 & 0.8 & 0.4 & 0 \\ 0 & 0.7 & 0.2 & 0.9 & 0 \end{pmatrix} = \begin{pmatrix} 0.3 & 0.4 & 0.3 & 0.4 & 0 \\ 0.5 & 0.5 & 0.5 & 0.4 & 0.2 \\ 0.7 & 0 & 0 & 0.3 & 0.6 \\ 0.7 & 0.7 & 0.7 & 0.9 & 0.6 \end{pmatrix}$$

Nesta execução, os valores de r_{11} e r_{43}, por exemplo, são obtidos conforme descrito a seguir. Os demais valores da relação R são obtidos seguindo o mesmo raciocínio.

$$\begin{aligned} r_{11} = 0.3 &= \max\left[\min\left(p_{11}, q_{11}\right), \min\left(p_{12}, q_{21}\right), \min\left(p_{13}, q_{31}\right)\right] \\ &= \max\left[\min\left(0, 0.7\right), \min\left(0.3, 0.5\right), \min\left(0.4, 0\right)\right] \\ &= \max\left[0,\ 0.3,\ 0\right] \end{aligned}$$

$$\begin{aligned} r_{43} = 0.7 &= \max\left[\min\left(p_{41}, q_{13}\right), \min\left(p_{42}, q_{23}\right), \min\left(p_{43}, q_{33}\right)\right] \\ &= \max\left[\min\left(0.7, 0\right), \min\left(0.7, 0.8\right), \min\left(0.9, 0.2\right)\right] \\ &= \max\left[0,\ 0.7,\ 0.2\right] \end{aligned}$$

Da mesma forma como na composição entre relações *fuzzy*, define-se a composição de um conjunto *fuzzy* A em X e uma relação *fuzzy* P em $X \times Y$ como sendo o conjunto *fuzzy* B em Y dado por:

$$B(y) \ = \ (A \circ P)(y) \ = \ \max_{x \in X} \min\left[A(x),\ P(x,\ y)\right]$$

Neste texto três propriedades das relações *fuzzy* devem ser destacadas: reflexividade, simetria e transitividade. Uma relação *fuzzy* R definida em $X \times X$ é reflexiva se e somente se para todo $x \in X$, $R(x,\ x) = 1$, ou seja, todos os elementos da diagonal da matriz são iguais a 1. Uma relação *fuzzy* R é simétrica se e somente $R(y,\ x) = R(x,\ y)$ para todo $x,\ y \in X$. Se R é simétrica então sua transposta é igual a ela mesma, $R^T = R$. Na teoria clássica uma relação é transitiva se dado que um primeiro elemento é relacionado ao segundo elemento e este é relacionado a um terceiro elemento então o primeiro elemento também é relacionado ao terceiro elemento, ou seja, se para quaisquer três elementos $(x,\ y,\ z) \in X$, $(x,\ z) \in R$ sempre que $(x,\ y) \in R$ e $(y,\ z) \in R$ para pelo menos um $y \in X$. A definição de transitividade é baseada no conceito de composição de relações *fuzzy*. Uma relação *fuzzy* é transitiva se:

$$R(x,\ z) \ \geq \ \max_{y \in Y} \min[R(x,\ y),\ R(y,\ z)]$$

Em algumas aplicações uma relação *fuzzy* que deveria ser transitiva não apresenta a propriedade de transitividade. Neste caso é necessário converter uma dada relação *fuzzy* R em uma relação transitiva que seja o mais próxima possível de R. Esta relação é denominada o fecho transitivo de R. O fecho transitivo $R^{\bullet}$ é determinado pelo algoritmo a seguir:

1. Calcular $R' = R \cup (R \circ R)$;
2. Se $R' \neq R$, fazer $R = R'$ e voltar ao passo 1, caso contrário $R^{\bullet} = R'$ e o algoritmo termina;

4.2 Modelo *Fuzzy* Utilizando Múltiplas Ontologias Relacionadas

O modelo *fuzzy* de múltiplas ontologias relacionadas é a proposta desta tese para recuperação de informação utilizando uma base de conhecimento constituída por múltiplas ontologias *lightweight* relacionadas onde cada uma representa um ramo do conhecimento do domínio. Esta abordagem permite que cada ontologia possa ser desenvolvida por grupos de especialistas distintos, de forma independente, sendo que o relacionamento entre elas é realizado através do estabelecimento de relações entre seus conceitos. Este modelo considera que cada ontologia seja constituída por uma taxonomia com relações de especialização e generalização *fuzzy* entre seus conceitos e que as relações entre as ontologias sejam do tipo associação positiva *fuzzy* estabelecidas entre os conceitos das ontologias distintas.

Os documentos estarão indexados pelos conceitos das ontologias. Quando o usuário elabora uma consulta, utilizando os conceitos das ontologias do domínio, o modelo faz a expansão desta consulta para considerar conceitos que não tenham sido relacionados pelo usuário mas que, em função do conhecimento existente na base, estejam relacionados entre si podendo melhorar a qualidade da informação recuperada. Nas próximas seções a proposta do modelo *fuzzy* de ontologias relacionadas será apresentada em detalhes.

4.2.1 Representação do Conhecimento pelas Ontologias Relacionadas

No modelo *fuzzy* de múltiplas ontologias relacionadas a representação do conhecimento é feita utilizando múltiplas ontologias *lightweight*, ou seja, cada ontologia é representada por uma hierarquia de conceitos associados por relações de especialização e generalização. Cada ontologia modela um domínio do conhecimento. Estas ontologias estão relacionadas formando a base de conhecimento do modelo. A Fig. 4.1 mostra uma base de conhecimento composta por duas ontologias relacionadas correspondentes aos domínios D_i e D_j. As setas com linhas tracejadas ilustram a associação entre os conceitos das ontologias distintas. Esta forma de organização e representação do conhecimento é uma das contribuições desta tese.

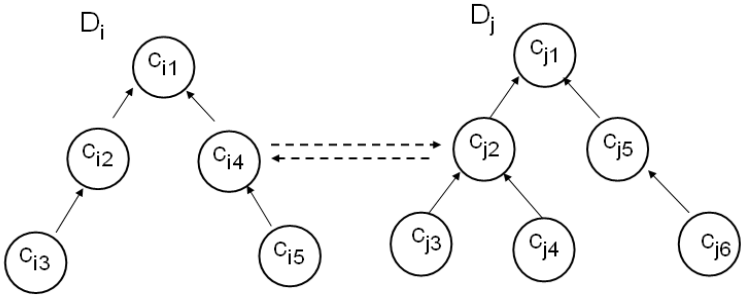

Fig. 4.1: Base de conhecimento com múltiplas ontologias relacionadas.

Para expressar as associações entre os conceitos nas ontologias são considerados três tipos de relacionamentos. Os conceitos pertencentes a uma mesma ontologia estão organizados em uma taxonomia e estão relacionados por dois tipos de relacionamentos: associação de especialização *fuzzy* (S) e associação de generalização *fuzzy* (G). Os conceitos pertencentes a ontologias distintas podem estar relacionados pelo terceiro tipo de relacionamento dado pela associação positiva *fuzzy* (P). A Fig. 4.2 ilustra o esquema da representação do conhecimento utilizando múltiplas ontologias e os possíveis relacionamentos existentes entre elas. Neste esquema cada ontologia é dada pelo conjunto de conceitos do domínio $D_k = \{c_{k1},\, c_{k2},\, \cdots c_{ky}\}$ onde $1 \leq k \leq K$ sendo K o número de domínios e $y = |D_k|$ é o número de conceitos em cada domínio.

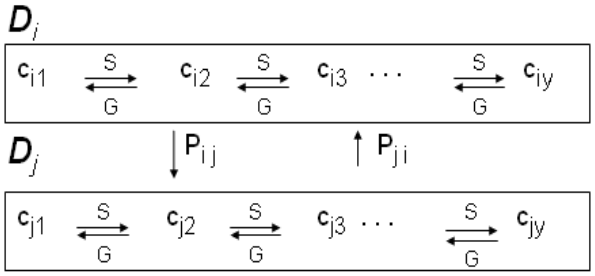

Fig. 4.2: Esquema da base de conhecimento com múltiplas ontologias e seus relacionamentos: especialização *fuzzy* (*S*), generalização *fuzzy* (*G*) e associação positiva *fuzzy* (*P*).

A associação positiva *fuzzy* (P) relaciona conceitos que possuem significados associados como, por exemplo, através de uma relação espacial $(\text{Nordeste} \leftrightarrow \text{Clima Semi-árido})$, causal $(\text{praga} \leftrightarrow \text{vírus})$ ou de similaridade $(\text{armazém} \leftrightarrow \text{silo})$ em alguns contextos. Esta variação no tipo de relacionamento que uma associação positiva *fuzzy* pode representar faz com que ela possibilite associar um significado semântico maior, entre os conceitos, que as relações de termos relacionados (RT) existentes em um tesauro.

Na associação de generalização *fuzzy* (G), um conceito é uma generalização de outro conceito se ele consistir daquele conceito $(\text{ferramenta} \rightarrow \text{martelo})$ ou se ele incluir aquele conceito no sentido

partitivo (cereal $\rightarrow$ milho). A associação de especialização *fuzzy* (S) é o inverso da relação de generalização *fuzzy*.

Definição 4.1 *Sejam D_i e D_j dois domínios constituídos de conjuntos distintos de conceitos.*

- *Associação Positiva Fuzzy é uma relação fuzzy:* $(R_{ij}^P : D_i \times D_j \rightarrow [0,1])$ *que é não simétrica, não reflexiva e não transitiva.*
- *Generalização Fuzzy é uma relação fuzzy:* $(R_i^G : D_i \times D_i \rightarrow [0,1])$ *que é não simétrica, não reflexiva e transitiva.*
- *Especialização Fuzzy é uma relação fuzzy:* $(R_i^S : D_i \times D_i \rightarrow [0,1])$ *que é não simétrica, não reflexiva e transitiva.*

Considerando-se $K = 2$ domínios dados por $D_1 = \{c_{11}, c_{12}, \cdots c_{1x}\}$ onde $1 \leq x \leq m$, $m = |D_1|$ e $D_2 = \{c_{21}, c_{22}, \cdots c_{2y}\}$ onde $1 \leq y \leq n$, $n = |D_2|$ e, de acordo com a Def. 4.1, tem-se que:

- A relação R_{12}^P indica que existe uma associação positiva *fuzzy* entre os conceitos dos domínios D_1 e D_2 sendo que o valor da relação $R_{12}^P(c_{1x},\ c_{2y}) = r_{xy} \in [0,1]$ indica o grau em que o conceito c_{1x} está associado positivamente ao conceito c_{2y}. O valor 0 indica que não existe associação positiva entre os conceitos. A relação R_{12}^P pode ser representada por uma matriz $m \times n$.
- A relação R_1^G indica os graus de generalização entre os conceitos do domínio D_1 sendo que o valor da relação $R_1^G(c_{1x},\ c_{1z}) = r_{xz} \in [0,1]$ indica o grau em que o conceito c_{1x} generaliza o conceito c_{1z}, neste caso, $1 \leq x,\ z \leq m$. A relação R_1^G pode ser representada por uma matriz $m \times m$.
- A relação R_1^S indica os graus de especialização entre os conceitos do domínio D_1 sendo que o valor da relação $R_1^S(c_{1x},\ c_{1z}) = r_{xz} \in [0,1]$ indica o grau em que o conceito c_{1x} especializa o conceito c_{1z}, neste caso, $1 \leq x,\ z \leq m$. A relação R_1^S pode ser representada por uma matriz $m \times m$. A relação R_1^S é o inverso da relação R_1^G.

De acordo com o esquema apresentado na Fig. 4.2 cada domínio possuirá uma relação de especialização e uma relação de generalização num total de K relações de cada tipo. No que se refere às relações de associação entre domínios cada domínio possuirá uma relação de associação positiva com cada um dos outros domínios gerando um total de relações positivas dadas pelo arranjo $A_s(k,2) = k!/(k-2)! = k(k-1)$. Cada uma destas relações será representada por uma matriz.

Para obter o grau dos relacionamentos implícitos entre os conceitos de um mesmo domínio calcula-se o fecho transitivo ponderado das relações de especialização e generalização *fuzzy* baseado no algoritmo de fecho transitivo apresentado na seção 4.1.2. O cálculo dos fechos transitivos ponderados das relações R_i^G e R_i^S, onde $1 \leq i \leq K$ e K é o número de domínios, resulta nas relações R_{Gi}^* e R_{Si}^* respectivamente.

Definição 4.2 *O fecho transitivo ponderado R^* de uma relação fuzzy R é definido utilizando-se o algoritmo iterativo consistindo dos seguintes passos:*

1. *Calcular $R' = R \cup [we_t(R \circ R)]$ onde $we_t \in [0,1]$, $t = \{G, S\}$;*
2. *Se $R' \neq R$, fazer $R = R'$ e voltar ao passo 1; caso contrário $R^* = R'$ e o algoritmo termina.*

A Fig. 4.3 mostra as relações implícitas entre os conceitos das ontologias através de setas tracejadas. O fecho transitivo ponderado vai permitir calcular o valor destas relações entre os conceitos das ontologias. O peso $we_t \in [0,1]$ penaliza a força da associação entre conceitos distantes na ontologia. Isto significa que conceitos mais próximos na taxonomia possuem um valor de associação mais alto. À medida que a distância entre os conceitos aumenta o seu valor de associação diminui. Considera-se que os conceitos com valor de associação alto possuem uma associação semântica maior.

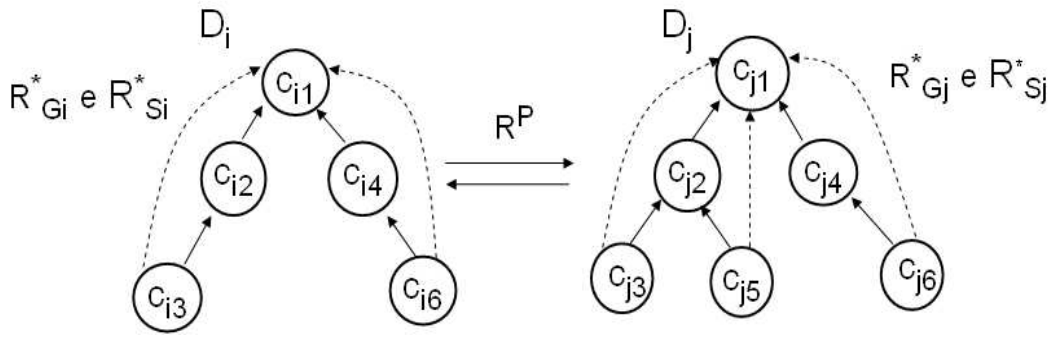

Fig. 4.3: Cálculo das relações implícitas nas ontologias pelo fecho transitivo ponderado.

Na literatura [51] é discutido que, no processo de expansão da consulta, o uso de estruturas conceituais como tesauros ou ontologias pode deteriorar a qualidade da recuperação de documentos. Isto se dá pois a adição indiscriminada de novos termos na consulta original tende a gerar ruído, aumentando a cobertura mas diminuindo a precisão. Para controlar a adição de novos conceitos na expansão da consulta pode-se adotar um valor limite b que estabelece o valor mínimo para que a associação de generalização ou especialização *fuzzy* entre dois conceitos seja considerada. Valores de associação menores que b são considerados como zero. Este procedimento é aplicado depois do cálculo do fecho transitivo ponderado e permite manter apenas as associações que possuem maior significado semântico.

4.2.2 Representação dos Documentos

Nos modelos para recuperação de informação, os documentos são representados por um conjunto de termos indicativos de seu conteúdo denominados índices. Dada uma coleção de documentos os conjuntos de termos representativos de toda a coleção formam o conjunto de índices da coleção. Neste esquema um documento é representado por um vetor formado pelo conjunto de índices da coleção. O vetor de índices pode assumir valores booleanos indicando a presença (1) ou ausência (0) do índice no documento ou valores que indicam a importância do índice na representação do documento.

No modelo *fuzzy* de múltiplas ontologias relacionadas deve-se associar cada documento aos conceitos das ontologias que representam os domínios. O conjunto de documentos é dado por $DOC = \{d_1, d_2, d_3, \cdots, d_l\}$ onde $1 \leq l \leq N$ e N é o número de documentos da coleção. Os domínios são representados pelos conjuntos $D_k = \{c_{k1},\ c_{k2},\ \cdots\ c_{ky}\}$, onde $1 \leq k \leq K$, sendo K o número de domínios e $y = |D_k|$ é o número de conceitos em cada domínio.

Para cada domínio D_k existe uma relação $U_k : DOC \times D_k$ que associa o conjunto de documentos da coleção, dado por DOC, ao conjunto de conceitos do domínio dado por D_k. O valor da relação U_k indica o quanto os conceitos da ontologia do domínio D_k são relevantes para descrever os documentos DOC da coleção. O valor da relação é calculado seguindo o esquema *tf-idf* [9, 110].

Definição 4.3 *Seja N o número total de documentos no sistema, c_{ky} um conceito do domínio D_k onde $1 \leq y \leq |D_k|$ e n_y o número de documentos nos quais o conceito c_{ky} aparece. Seja $freq_{ly}$ a frequência no documento d_l para o conceito c_{ky}(o número de vezes em que o conceito c_{ky} é mencionado no texto do documento d_l). Então a frequência normalizada f_{ly}, no documento d_l, para o conceito c_{ky} é dada pela Eq. 4.1.*

$$f_{ly} = \frac{freq_{ly}}{\max_t freq_{lt}} \tag{4.1}$$

O máximo é calculado sobre todos os termos que são mencionados no texto do documento d_l. Se o conceito c_{ky} não aparece no documento d_l então $f_{ly} = 0$. Seja a frequência inversa do documento idf_y para o conceito c_{ky} dada pela Eq. 4.2.

$$idf_y = \log \frac{N}{n_y} \tag{4.2}$$

O peso tf-idf u_{ly}, no documento d_l, para o conceito c_{ky} é dado pela Eq. 4.3.

$$u_{ly} = f_{ly}\ \log \frac{N}{n_y} \tag{4.3}$$

Neste contexto, o valor da relação $U_k\,(d_l,\ c_{ky}) = u_{ly}$, onde $1 \leq l \leq |DOC|$, $1 \leq k \leq K$ e $1 \leq y \leq |D_k|$ indica o grau em que o conceito $c_{ky} \in D_k$ representa o documento $d_l \in DOC$. As

relações U_k, $1 \leq k \leq K$, são representadas por matrizes $l \times m$ onde $l = |DOC|$ e $m = |D_k|$.

4.2.3 Representação da Consulta

Nos sistemas de recuperação de informação as consultas são expressas por termos conectados por operadores booleanos, em geral E, OU e NOT. Através da utilização de equivalências lógicas como as leis de De Morgan, a lei da eliminação da dupla negação e a lei distributiva é possível reduzir qualquer expressão booleana à forma conjuntiva normal que é o produto da soma dos termos. Uma expressão booleana está na forma conjuntiva normal se e somente se ela é uma conjunção de uma ou mais disjunções compostas de um ou mais termos. As expressões $A \wedge B$ e $((A \vee B) \wedge (C \vee D))$ são exemplos de expressões na forma conjuntiva normal.

No modelo *fuzzy* de múltiplas ontologias uma consulta deve ser expressa utilizando os conceitos das ontologias de domínio. Considera-se que esta consulta esteja na forma conjuntiva normal representada por um conjunto de sub-consultas conectadas pelo operador lógico E sendo que, em cada sub-consulta, os conceitos estarão conectados pelo operador lógico OU. Uma consulta do usuário, q_{user}, é representada como $q_{user} = \bigwedge_{h=1}^{S} q_h$, onde $1 \leq h \leq S$, S é o número de sub-consultas e q_h é composta por conceitos das ontologias conectados por OU. Dados os domínios $D_1 = \{c_{11}, c_{12}, c_{13}, c_{14}, c_{15}\}$ e $D_2 = \{c_{21}, c_{22}, c_{23}, c_{24}, c_{25}, c_{26}\}$ exemplos de consultas válidas seriam:

$$\begin{aligned} q_{user1} &= (c_{11} \vee c_{22}) \wedge (c_{13} \vee c_{25}) \\ q_{user2} &= (c_{14} \wedge c_{22}) \\ q_{user3} &= (c_{16} \vee c_{21}) \end{aligned}$$

As consultas do tipo q_{user1} constituem uma consulta composta por mais de uma sub-consulta. Neste exemplo uma sub-consulta é dada por $q_h = (c_{11} \vee c_{22})$. As consultas dadas por q_{user2} e q_{user3} são os casos onde a sub-consulta é a própria consulta final. Neste caso $q_{user} = q_h$.

Como no modelo proposto os documentos estão associados aos conceitos das ontologias por relações distintas então as sub-consultas, q_h, também devem ser particionadas para considerar os conceitos de cada domínio separadamente. Cada partição será constituída por um vetor $\overrightarrow{q}_i = (w_{i1}, w_{i2}, \cdots, w_{it})$ onde $1 \leq t \leq |D_i|$, D_i é o domínio associado ao conceito c_i, presente na sub-consulta, e $w_{it} \in \{0, 1\}$ indica a presença (1) ou ausência (0) do respectivo conceito, c_i, na consulta do usuário. No caso do exemplo anterior a sub-consulta $q_h = (c_{14} \wedge c_{22})$ será particionada em $q_1 = [0\,0\,0\,1\,0]$ e $q_2 = [0\,1\,0\,0\,0\,0]$ resultando $q_h = ([0\,0\,0\,1\,0] \wedge [0\,1\,0\,0\,0\,0])$.

Uma vez que a consulta esteja expressa na forma conjuntiva normal cada sub-consulta será tratada separadamente e retornará um conjunto de documentos recuperados, V_h, onde $1 \leq h \leq S$ e S é o número de sub-consultas que compõem a consulta do usuário q_{user}. A interseção dos conjuntos de

documentos V_h, retornados por cada sub-consulta, resultará o conjunto final de documentos, V, para a consulta do usuário, dada por q_{user}, conforme mostra a Eq. 4.4. Desta forma o restante deste trabalho irá tratar os aspectos relacionados às sub-consultas.

$$V = \bigwedge_{h=1}^{S} V_h \tag{4.4}$$

4.2.4 Expansão da Consulta

Ao se utilizar uma base de conhecimento para indexar os documentos de um sistema de recuperação de informação o objetivo é utilizar o conhecimento expresso na base para melhorar a qualidade dos documentos recuperados trazendo mais documentos associados à consulta (melhoria da taxa de cobertura) e apresentando estes documentos numa ordem onde os documentos do topo da lista de documentos sejam os mais relevantes à consulta, sendo ordenados em ordem decrescente de relevância (melhoria da taxa de precisão).

Através da base de conhecimento é possível explorar as relações entre os conceitos dos domínios para expandir a consulta do usuário com novos conceitos que, apesar de não estarem presentes na consulta inicial, sejam relacionados a estes. Pela expansão da consulta os novos conceitos serão incorporados à mesma permitindo a recuperação de documentos que sejam semanticamente relacionados à consulta original em função do conhecimento contido na base.

Esta seção apresenta o método de expansão da consulta a ser utilizado no modelo *fuzzy* de múltiplas ontologias relacionadas para considerar o conhecimento existente na base. O método de expansão da consulta é uma das contribuições desta tese.

O método de expansão da consulta é realizado em duas etapas. Na primeira etapa cada partição q_i, da sub-consulta inicial q_h, é expandida para considerar as relações existentes entre o domínio relativo à partição, dado por D_i, e os outros domínios da base. Para cada partição q_i serão gerados K novos conjuntos cada um se referindo aos conceitos dos outros domínios da base de conhecimento dados por $D_j, j \neq i, 1 \leq i, j \leq K$ que sejam associados aos conceitos presentes em q_i. Este processo gera uma nova consulta expandida entre domínios denominada $qent$.

A primeira etapa da expansão é dada pela Eq. 4.5. Na Eq. 4.5 a variável i se refere ao domínio da partição q_i e a variável j se refere aos demais domínios da base que serão considerados para expandir a partição.

$$qent = \bigvee_{i=1}^{K} \bigvee_{j=1}^{K} \begin{cases} q_i & j = i \\ w_P \left(q_i \circ R_{ij}^P \right) & j \neq i \end{cases} \tag{4.5}$$

Para fazer a expansão da consulta, entre os domínios, é utilizada a relação de associação positiva *fuzzy*, R_{ij}^P, entre os conceitos dos domínios dados por D_i e D_j. O modelo permite associar um valor

$w_P \in [0,1]$, que define um peso para a associação. Desta forma a influência da expansão positiva *fuzzy* no processo de expansão pode ser ajustada. O valor de $w_P = 0$ indica que a associação positiva não será considerada. A partição não é expandida com relação ao seu próprio domínio quando $j = i$. Quando se trata de domínios distintos, ou seja $j \neq i$, a partição da consulta é expandida através da composição entre a partição q_i e a relação positiva *fuzzy*, R_{ij}^P. Cada expansão vai gerar uma nova partição contendo os conceitos do domínio D_j. Os valores da nova partição indicam o grau em que os conceitos do domínio D_j estão associados aos conceitos da partição q_i da consulta pela associação positiva *fuzzy*.

Considere dois domínios D_1 e D_2 e uma sub-consulta $q_h = q_1 \vee q_2$, representada por vetores, e particionada nos dois domínios. A Fig. 4.4 representa, esquematicamente, o resultado do processo de expansão. Pela Fig. 4.4 pode-se ver que cada partição q_i, $1 \leq i \leq 2$, da sub-consulta inicial é expandida nos outros domínios gerando outras q_{ij} partições, $1 \leq j \leq 2$ num total de $ij = 4$ partições. Neste contexto $q_{12(P)}$, significa a expansão da partição q_1, da sub-consulta inicial, para o domínio D_2 através da associação *fuzzy* dada por R_{12}^P. Isto significa que os conceitos presentes em q_1, considerando a associação positiva *fuzzy* R_{12}^P, levaram os conceitos associados do domínio D_2 estarem presentes na sub-consulta também. A partição $q_{12(P)}$ é um vetor composto dos conceitos pertencentes ao domínio D_2. Os valores de $q_{12(P)} \in [0,1]$ indicam o grau de associação entre os conceitos do domínio D_2 com os conceitos do domínio D_1, presentes em q_1.

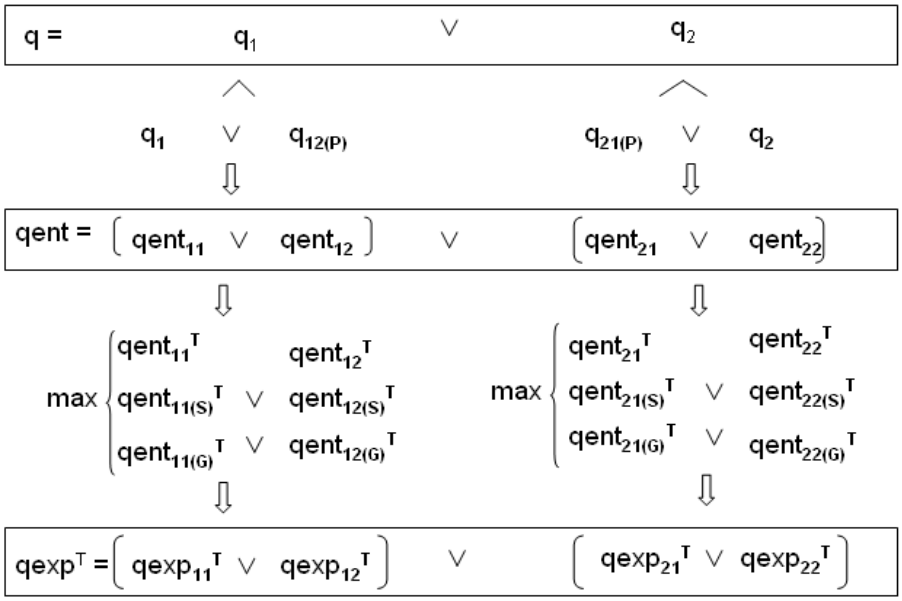

Fig. 4.4: Esquema de expansão da sub-consulta inicial em dois domínios considerando as relações de associação positiva *fuzzy* (P), especialização *fuzzy*(S) e generalização *fuzzy*(G)

A Fig. 4.5 ilustra a primeira fase da expansão da consulta considerando as bases de conhecimento

formada pelos domínios D_1 e D_2. Nesta base de conhecimento os conceitos c_{14} e c_{22} estão associados pela associação positiva. Considere os valores de associação entre os conceitos iguais a 1.0 e o valor do peso da associação positiva $w_P = 0.7$. Para a consulta $q = q_h = c_{14}$, composta por apenas um conceito, tem-se a partição da consulta dada por $q_1 = ([0\ 0\ 0\ 1\ 0])$. Depois da primeira fase da expansão tem-se o valor de $qent = (c_{14}\ or\ c_{22})$. A representação da consulta na forma particionada é dada por $qent = ([0\ 0\ 0\ 1\ 0] \vee [0\ 0.7\ 0\ 0\ 0\ 0])$. Nesta fase da expansão o conceito c_{22} é adicionado à expansão do conceito c_{14}. Na forma particionada os pesos com que os conceitos são adicionados à consulta são armazenados na consulta expandida.

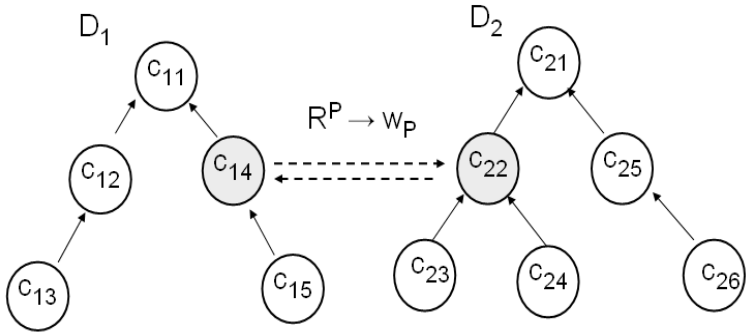

Fig. 4.5: Primeira fase da expansão da sub-consulta, em dois domínios, considerando as relações de associação positiva *fuzzy* (P).

Uma vez que se tenha a sub-consulta $qent$ expandida entre os domínios é realizada a segunda etapa da expansão. Esta etapa visa realizar a expansão da sub-consulta $qent$ intra domínios, isto é, considerando o fecho transitivo ponderado das relações de especialização *fuzzy*, R^*_{Si} e generalização *fuzzy*, R^*_{Gi}. Esta expansão gera a sub-consulta expandida final $qexp^T$ transposta. A Eq. 4.6 mostra como é realizada a expansão.

$$qexp^T = \bigvee_{i=1}^{K} \bigvee_{j=1}^{K} \max \begin{cases} qent_{ij}^T \\ w_r \left(R^*_{rj} \circ qent_{ij}^T \right) \end{cases} \quad \text{onde} \quad r \in \{S, G\} \tag{4.6}$$

No processo de expansão intra-domínios a partição transposta $qent_{ij}^T$, $1 \leq i, j \leq K$ é expandida para considerar as associações de especialização e generalização *fuzzy* entre conceitos de seu domínio, dado por D_j . Isto é feito através da composição entre as relações *fuzzy* correspondentes às associações de especilização (S) e generalização (G), entre os conceitos do domínio Dj, e a partição expandida $qent_{ij}^T$. Nesta fase, a partição da consulta $qent_{ij}$ é transposta para permitir a composição *fuzzy* com as relações de especilização (S) e generalização (G) *fuzzy*. Nesta expansão os conceitos mais gerais e mais específicos, correspondentes aos conceitos presentes na partição, são acrescentados à mesma. O modelo permite associar um valor $w_r \in [0, 1]$, $r \in \{S, G\}$ que define um peso para o tipo de associação. Desta forma a expansão pode ser ajustada para considerar mais a influência

de um tipo de associação em relação a outro. A expansão final é dada pelo valor máximo entre os valores associados aos conceitos presentes em $qent_{ij}^{T}$, resultante da expansão entre-domínios, e os novos valores nas partições $qent_{ij(r)}^{T}$ dados pelos tipos de associação $r \in \{S, G\}$.

A Fig. 4.4 representa, esquematicamente, o resultado do processo de expansão intra-domínios para a sub-consulta expandida $qent$. De acordo com a Fig. 4.4, a partição $qent_{12}^{T}$, resultante do processo de expansão entre domínios, é novamente expandida para considerar as relações de especialização *fuzzy*, $qent_{12(S)}^{T}$, e generalização *fuzzy*, $qent_{12(G)}^{T}$, entre os conceitos do domínio D_2. A partição final expandida $qexp_{12}^{T} = \max\left(qent_{12}^{T},\ qent_{12(S)}^{T},\ qent_{12(G)}^{T}\right)$ é dada pelo valor máximo entre os valores presentes na partição $qent_{12}^{T}$, resultante da primeira fase, e os valores presentes nas partições $qent_{12(S)}^{T}$ e $qent_{12(G)}^{T}$ resultantes da expansão considerando a associação de especialização *fuzzy* e a associação de generalização *fuzzy*, respectivamente, para o domínio D_2.

A Fig. 4.6 ilustra a segunda fase da expansão da consulta considerando as bases de conhecimento formada pelos domínios D_1 e D_2. Na ontologia que representa o domínio D_1 tem-se que o conceito c_{11} é mais geral que o conceito c_{14} e o conceito c_{15} é mais específico que o conceito c_{14}. Na ontologia que representa o domínio D_2 tem-se que o conceito c_{21} é mais geral que o conceito c_{22} e os conceitos c_{23} e c_{24} são mais específicos que o conceito c_{22}. Nesta fase da expansão, os conceitos mais específicos e mais gerais, que os conceitos presentes em $qent$, são adicionados à consulta. Depois da segunda fase da expansão tem-se o valor de $qexp = (\ c_{11}\ or\ c_{14}\ or\ c_{15})\ or\ (c_{21}\ or\ c_{22}\ or\ c_{23}\ or\ c_{24})$. Considere $w_G = 0.3$, $w_G = 0.7$ e os valores de associação entre conceitos iguais a 1.0. A representação da consulta na forma particionada é dada por $qent = ([0.3\ 0\ 0\ 1\ 0.7] \vee [0.21\ 0.7\ 0.49\ 0.49\ 0\ 0])$. Na forma particionada, os pesos com que os conceitos são adicionados à consulta são armazenados na consulta expandida.

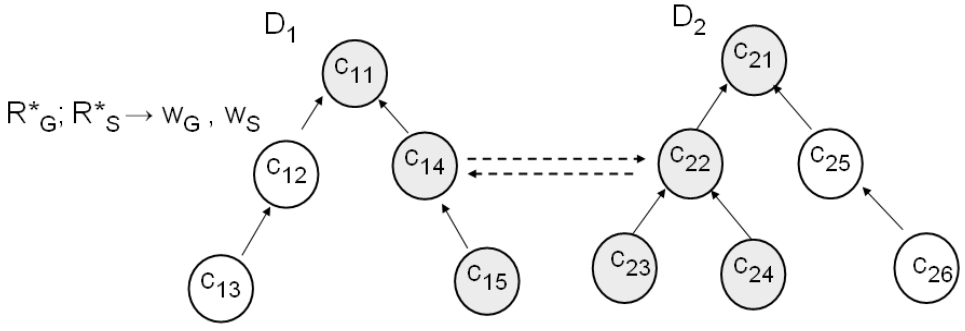

Fig. 4.6: Segunda fase da expansão da sub-consulta inicial em dois domínios considerando as relações de generalização *fuzzy* (G) e especialização (S).

4.2.5 Função de Relevância

A relevância dos documentos da coleção é dada pelo produto entre as representações dos documentos e as partições da consulta *fuzzy* $qexp^T$. Esta relevância é calculada pelo produto das relações U_j com cada partição $qexp_{ij}^T$ resultando no conjunto de documentos recuperados V_h como mostrado na Eq. 4.7.

Cada relação U_j associa os documentos da coleção aos conceitos do domínio D_j, onde $1 \leq j \leq K$. Por outro lado a partição $qexp_{ij}^T$ representa a expansão dos conceitos da partição q_i para o domínio D_j, onde $1 \leq i,\ j \leq K$. Neste caso $qexp_{ij}^T$ é um vetor constituído dos conceitos do domínio D_j. O produto da relação U_j com a partição $qexp_{ij}^T$, dado por $\left(U_j\ qexp_{ij}^T\right)$, pondera as associações dos documentos aos conceitos (expressos na relação U_j) pela força do relacionamento entre os conceitos (expressos na partição $qexp_{ij}^T$). Se o conceito c_1, da sub-consulta inicial, estiver associado a um conceito c_2, de outro domínio, por um grau com valor $\in [0,\ 1]$ significa que um documento associado ao conceito c_2 vai estar associado ao conceito c_1, expresso na sub-consulta inicial, ponderado pelo mesmo grau.

O conjunto $V_h\,(v_l)$ representa todos os documentos da coleção e seu valor $v_l \in [0, 1]$ indica o quanto o documento d_l, $1 \leq l \leq |DOC|$ é similar à sub-consulta inicial do usuário.

$$V_h = \bigvee_{i=1}^{K} \bigvee_{j=1}^{K} \left(U_j\ qexp_{ij}^T\right) \tag{4.7}$$

Conforme mostrado na Eq. 4.4, o conjunto final de documentos para a consulta do usuário, V, é dado pela interseção dos conjuntos de documentos resultantes para as sub-consultas, V_h. Ao usuário é interessante visualizar apenas um subconjunto de V que represente os documentos mais relevantes. Desta forma define-se um valor limite *t* que estabelece um valor mínimo para v_l para que o documento correspondente seja retornado para o usuário. Em termos de operações *fuzzy* isto significa aplicar um α-cut no conjunto V, onde $\alpha = t$, resultando o conjunto $(V)_t$. Os documentos em $(V)_t$ cujo $v_l \neq 0$ serão apresentados ao usuário na ordem decrescente de relevância.

4.3 Execução do Modelo

Nesta seção são apresentados os algoritmos que ilustram a execução do modelo *fuzzy* de múltiplas ontologias relacionadas para recuperação de informação em uma coleção de documentos. 0 primeiro algoritmo monta a base de conhecimento.

Algoritmo: Monta-Base-Conhecimento

Input: Conjuntos de conceitos dos domínios D_i
Ontologias relacionadas representando os domínios D_i

Output: Fecho transitivo ponderado das relações de especialização *fuzzy*, R^*_{Si}
Fecho transitivo ponderado das relações de generalização *fuzzy*, R^*_{Gi}
Relações positivas *fuzzy* entre os domínios, R^P_{ij}

Data: Pesos we_S, we_G

foreach *Domínio* D_i **do**
  Extrair, das ontologias relacionadas, os valores das relações R^S_i e R^G_i;
  Calcular os fechos transitivos ponderados R^*_{Si} e R^*_{Gi};
  foreach *Domínio* D_j **do**
    if $D_i \neq D_j$ **then**
      Extrair, das ontologias relacionadas, os valores da relação R^P_{ij};
    end
  end
end

O segundo algoritmo indexa os documentos da coleção.

Algoritmo: Indexa-Documentos

Input: Conjuntos de conceitos dos domínios D_j
Conjunto de documentos da coleção DOC

Output: Relações U_j

foreach *Documento* d_l *no conjunto* DOC **do**
  foreach *Domínio* D_j **do**
    foreach *Conceito* c_{jy} *no conjunto* D_j **do**
      Calcular a frequência normalizada, f_{ly}, no documento d_l, para o conceito c_{jy};
      Calcular a frequência inversa do documento, idf_y, para o conceito c_{jy};
      Calcular o peso $tf - idf$ dado por $u_{ly} = f_{ly}\, idf_y$, no documento d_l, para o conceito c_{jy};
      Fazer $U_j\,(d_l,\ c_{jy}) = u_{ly}$;
    end
  end
end

O terceiro algoritmo executa o modelo de recuperação proposto na tese.

Algoritmo: Recupera-Informação

Input: Consulta do usuário q

Output: Conjunto de documentos recuperados V

Data: Conjuntos de conceitos dos domínios D_i

Fecho transitivo ponderado das relações de especialização *fuzzy*, R^*_{Si}

Fecho transitivo ponderado das relações de generalização *fuzzy*, R^*_{Gi}

Relações positivas *fuzzy* entre os domínios, R^P_{ij}

Documentos indexados pelas relações U_j

Pesos w_P, w_G, w_S

Limite t

Gerar forma conjuntiva normal $q = \bigwedge q_h$;

foreach *Sub-consulta* q_h **do**

  Gerar partições q_i nos domínios D_i;

  foreach *Partição* q_i **do**

    foreach *Domínio* D_j **do**

      if $i = j$ **then**

        O valor da expansão é igual à partição: $qent_{ij} = q_i$;

      else

        Expandir a partição q_i no domínio D_j: $qent_{ij} = w_P \left(q_i \circ R^P_{ij}\right)$;

      end

    end

  end

  foreach *Expansão* $qent_{ij}$ **do**

    foreach *Domínio* D_j **do**

      Expandir $qent_{ij}$ no domínio D_j: $qexp^T_{ij}$ =
      $\max\left(qent^T_{ij}, w_S\left(R^*_{Sj} \circ qent^T_{ij}\right), w_G\left(R^*_{Gj} \circ qent^T_{ij}\right)\right)$;

    end

  end

  Calcular o conjunto de documentos para a partição q_i: $Vq_i = \max\left(U_j\, qexp^T_{ij}\right)$;

  Compor os conjuntos de documentos das partições q_i para a sub-consulta $V_h = \max Vq_i$;

end

Calcular o conjunto V de documentos para a consulta q do usuário: $V = \min V_h$;

Ordenar o conjunto de documentos V em ordem descrescente;

Apresentar, para o usuário, os documentos em V com relevância maior que o limite t;

4.4 Exemplo de Uso do Modelo

Para ilustrar a idéia do modelo *fuzzy* de múltiplas ontologias relacionadas foi elaborado um exemplo composto de dois domínios geográficos. Estes domínios são dados pelas estruturas que representam um subconjunto da divisão territorial do Brasil (D_1) e Climas do Brasil (D_2), a partir do mapa de clima da Fig. A.1, no Apêndice A. Os valores entre os conceitos, dentro das ontologias e entre as ontologias, são calculados em função da distribuição espacial das entidades, representadas pelos conceitos, no mapa. O Apêndice A mostra como estes valores são calculados. Os documentos do conjunto DOC foram indexados pelos conceitos das ontologias conforme ilustra a Fig. 4.7.

$$
\begin{aligned}
D_1 &= \{c_{11} : \text{Brasil},\ c_{12} : \text{Norte},\ c_{13} : \text{Pará},\ c_{14} : \text{Nordeste},\ c_{15} : \text{Maranhão}\} \\
D_2 &= \{c_{21} : \text{Clima}, c_{22} : \text{Tropical},\ c_{23} : \text{Am},\ c_{24} : \text{Aw},\ c_{25} : \text{Semi-árido},\ c_{26} : \text{BSh}\} \\
DOC &= \left\{ \begin{array}{l} d_1 : \text{Doc1},\ d_2 : \text{Doc2},\ d_3 : \text{Doc3},\ d_4 : \text{Doc4},\ d_5 : \text{Doc5}, d_6 : \text{Doc6}, \\ d_7 : \text{Doc7},\ d_8 : \text{Doc8},\ d_9 : \text{Doc9},\ d_{10} : \text{Doc10},\ d_{11} : \text{Doc11} \end{array} \right\}
\end{aligned}
$$

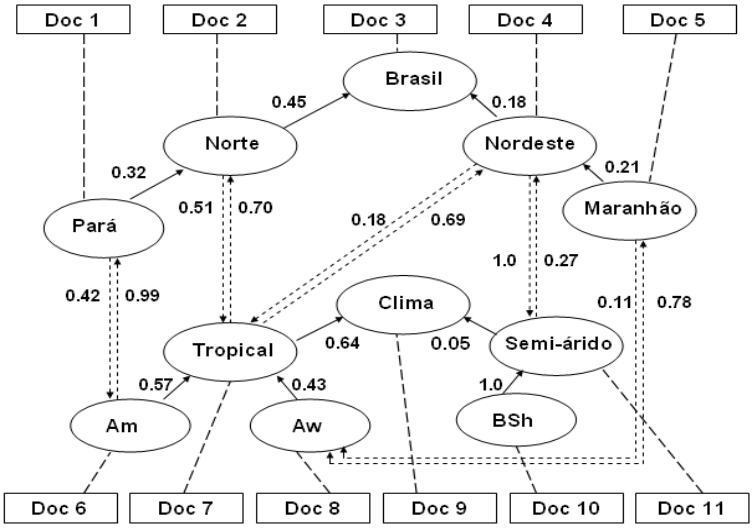

Fig. 4.7: Exemplo de uso do modelo *fuzzy* utilizando múltiplas ontologias relacionadas.

A partir do conhecimento expresso na Fig. 4.7, as relações de especialização e generalização *fuzzy* entre os conceitos dos domínios são extraídas. A seguir o fecho transitivo ponderado é calculado para estas relações utilizando os valores de peso $we_S = 0.8$ e $we_G = 0.2$. As relações de especialização e generalização *fuzzy* entre os conceitos do domínio D_1, R_1^S e R_1^G, e seus respectivos fechos transitivos ponderados, R_{S1}^* e R_{G1}^*, são dados por:

$$R_1^S = \begin{array}{c} \\ c_{11} \\ c_{12} \\ c_{13} \\ c_{14} \\ c_{15} \end{array} \begin{array}{c} \begin{array}{ccccc} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.45 & 0 & 0 & 0 & 0 \\ 0 & 0.32 & 0 & 0 & 0 \\ 0.18 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.21 & 0 \end{pmatrix} \end{array} \quad R_{S1}^* = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.45 & 0 & 0 & 0 & 0 \\ 0.256 & 0.32 & 0 & 0 & 0 \\ 0.18 & 0 & 0 & 0 & 0 \\ 0.144 & 0 & 0 & 0.21 & 0 \end{pmatrix}$$

A relação R_1^S indica, por exemplo, que o conceito c_{13}, associado a Pará, especializa o conceito c_{12}, associado a Norte, com o grau no valor de 0.32. Este valor é retirado das ontologias na Fig. 4.7. No cálculo do fecho transitivo ponderado, as associações implícitas entre os conceitos das ontologias aparecem. Pela Fig. 4.7 pode-se observar que, na ontologia que representa o domínio de divisão territorial, existe uma associação implícita entre o conceito c_{13}, associado a Pará e o conceito c_{11}, associado a Brasil. Esta associação torna-se explícita na relação R_{S1}^* onde o grau de associação entre os conceitos é calculado, pelo fecho transitivo ponderado, e seu valor é 0.256. Isto significa que o conceito c_{13}, associado a Pará, especializa o conceito conceito c_{11}, associado a Brasil, com o grau no valor de 0.256.

$$R_1^G = \begin{array}{c} \\ c_{11} \\ c_{12} \\ c_{13} \\ c_{14} \\ c_{15} \end{array} \begin{array}{c} \begin{array}{ccccc} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} \end{array} \\ \begin{pmatrix} 0 & 0.45 & 0 & 0.18 & 0 \\ 0 & 0 & 0.32 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.21 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \quad R_{G1}^* = \begin{pmatrix} 0 & 0.45 & 0.064 & 0.18 & 0.036 \\ 0 & 0 & 0.32 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.21 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

A relação R_1^G indica, por exemplo, que o conceito c_{12}, associado a Norte, generaliza o conceito c_{13}, associado a Pará, com o grau no valor de 0.32. Este valor é retirado das ontologias na Fig. 4.7. Deve-se lembrar que a relação de generalização é o inverso da relação de especialização. No cálculo do fecho transitivo ponderado as associações implícitas entre os conceitos das ontologias aparecem. Pela Fig. 4.7 pode-se observar que, na ontologia que representa o domínio de divisão territorial, existe uma associação implícita entre o conceito c_{11}, associado a Brasil, e o conceito c_{13}, associado a Pará. Esta associação torna-se explícita na relação R_{G1}^* onde o grau de associação entre os conceitos é calculado pelo fecho transitivo ponderado e seu valor é 0.064. Isto significa que o conceito c_{11}, associado a Brasil, generaliza o conceito c_{13}, associado a Pará, com o grau no valor de 0.064.

No cálculo do fecho transitivo ponderado considera-se o valor do peso $we_S = 0.8$ e o valor do

peso $we_G = 0.2$. Desta forma, a relação de especialização (S) é mais privilegiada que a relação de generalização (G). Isto pode ser visto pelos valores calculados pelo fecho transitivo ponderado. Ele atribui um valor maior para relação de especialização, $R^*_{S1}(c_{13}, c_{11}) = 0.256$, e um valor menor para a relação de generalização, $R^*_{G1}(c_{11}, c_{13}) = 0.064$.

As relações de especialização e generalização *fuzzy* entre os conceitos do domínio D_2, R^S_2 e R^G_2, e seus respectivos fechos transitivos ponderados, R^*_{S2} e R^*_{G2}, são dados por:

$$R^S_2 = \begin{array}{c} \\ c_{21} \\ c_{22} \\ c_{23} \\ c_{24} \\ c_{25} \\ c_{26} \end{array} \begin{array}{c} \begin{array}{cccccc} c_{21} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0.64 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.57 & 0 & 0 & 0 & 0 \\ 0 & 0.43 & 0 & 0 & 0 & 0 \\ 0.05 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 & 0 \end{pmatrix} \end{array} \quad R^*_{S2} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0.64 & 0 & 0 & 0 & 0 & 0 \\ 0.456 & 0.57 & 0 & 0 & 0 & 0 \\ 0.344 & 0.43 & 0 & 0 & 0 & 0 \\ 0.05 & 0 & 0 & 0 & 0 & 0 \\ 0.04 & 0 & 0 & 0 & 1.0 & 0 \end{pmatrix}$$

$$R^G_2 = \begin{array}{c} \\ c_{21} \\ c_{22} \\ c_{23} \\ c_{24} \\ c_{25} \\ c_{26} \end{array} \begin{array}{c} \begin{array}{cccccc} c_{21} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \end{array} \\ \begin{pmatrix} 0 & 0.64 & 0 & 0 & 0.05 & 0 \\ 0 & 0 & 0.57 & 0.43 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \quad R^*_{G2} = \begin{pmatrix} 0 & 0.64 & 0.114 & 0.086 & 0.05 & 0.01 \\ 0 & 0 & 0.57 & 0.43 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

As relações de associação positiva *fuzzy* entre os conceitos dos domínios, R^P_{12} e R^P_{21}, são dadas por:

$$R^P_{12} = \begin{array}{c} \\ c_{11} \\ c_{12} \\ c_{13} \\ c_{14} \\ c_{15} \end{array} \begin{array}{c} \begin{array}{cccccc} c_{21} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.51 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.42 & 0 & 0 & 0 \\ 0 & 0.18 & 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0.11 & 0 & 0 \end{pmatrix} \end{array} \quad R^P_{21} = \begin{array}{c} \\ c_{21} \\ c_{22} \\ c_{23} \\ c_{24} \\ c_{25} \\ c_{26} \end{array} \begin{array}{c} \begin{array}{ccccc} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.70 & 0 & 0.69 & 0 \\ 0 & 0 & 0.99 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.78 \\ 0 & 0 & 0 & 0.27 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Os valores das associações positivas, R^P_{12} e R^P_{21}, entre os conceitos são calculados em função da distribuição espacial das entidades, representadas por estes conceitos, no mapa da Fig. A.1. O

Apêndice A mostra como estes valores são calculados.

Os relacionamentos entre o conjunto de documentos DOC e os conceitos dos domínios D_1 e D_2, são dados, respectivamente, pelas relações U_1 e U_2 a seguir. O valor 1.0 foi atribuído na associação dos documentos aos conceitos para ilustrar melhor a influência do processo de expansão da consulta na recuperação de informação. Desta forma pode-se verificar como os pesos atribuídos aos conceitos, quando da sua adição à consulta inicial, influenciam a recuperação dos documentos.

$$U_1 = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \\ d_9 \\ d_{10} \\ d_{11} \end{array} \begin{array}{c} \begin{array}{ccccc} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} \end{array} \\ \begin{pmatrix} 0 & 0 & 1.0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 & 0 \\ 1.0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \qquad U_2 = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \\ d_9 \\ d_{10} \\ d_{11} \end{array} \begin{array}{c} \begin{array}{cccccc} c_{21} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.0 & 0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.0 & 0 & 0 \\ 1.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.0 \\ 0 & 0 & 0 & 0 & 1.0 & 0 \end{pmatrix} \end{array}$$

Considere a consulta do usuário q_{user} = Norte e Tropical. A consulta q_{user} já se encontra na forma conjuntiva normal, ou seja, $q_{user} \equiv q_h$, onde $h = 1$, significando que a consulta do usuário é composta por apenas uma sub-consulta.

Pela verificação da indexação dos documentos, na Fig. 4.7, observa-se que não existe um documento que seja simultaneamente associado aos conceitos Norte e Tropical. Desta forma, na execução de um sistema de recuperação de informação, sem uma base de conhecimento associada, nenhum documento seria recuperado para esta consulta. Por um outro lado, pelo conhecimento existente no mapa da Fig. A.1, o conjunto de documentos $R = \{\text{Doc2}, \text{Doc1}, \text{Doc6}\}$ é apontado como sendo relevante para a consulta q_{user}. O documento Doc2 é considerado relevante pois está associado ao conceito Norte, que faz parte da consulta, e ao mesmo tempo a região Norte do Brasil é, em sua maioria, associada ao clima Tropical. O documento Doc1 é considerado relevante pois está associado ao conceito Pará, que faz parte da região Norte, e ao mesmo tempo possui seu território quase que exclusivamente associado ao clima Tropical (Am da classificação Köppen). O documento Doc6 é considerado relevante pois está associado ao conceito Am que é do tipo Tropical e ao mesmo tempo constitui o clima predominante na região Norte. Neste exemplo é verificado como o modelo proposto se comporta na recuperação dos documentos relevantes para esta consulta.

Para o processo de recuperação de informação do modelo foi assumido o valor do limite $t = 0.2$ e dos pesos de ponderação da importância das relações *fuzzy*: $w_P = 0.7$, $w_S = 0.7$, $w_G = 0.3$. A expansão da consulta, seguindo o esquema apresentado na Fig. 4.4, é dada por:

(1) Particionar os conceitos da consulta, de acordo com os seus domínios, originando $q_1 = [0\ 1\ 0\ 0\ 0]$ e $q_2 = [0\ 1\ 0\ 0\ 0\ 0]$. Neste caso, $q_{user} = ([0\ 1\ 0\ 0\ 0] \wedge [0\ 1\ 0\ 0\ 0\ 0])$

(2) Calcular a expansão da consulta entre os domínios originando:
$qent = ((qent_{11} \vee qent_{12}) \wedge (qent_{21} \vee qent_{22}))$.

Seguindo a Eq. 4.5 tem-se que:

$$qent_{11} = q_1 = [0\ 1\ 0\ 0\ 0]$$

$$qent_{12} = 0.7\left(q_1 \circ R_{12}^P\right) = [0\ 0.357\ 0\ 0\ 0\ 0]$$

$$qent_{21} = 0.7\left(q_2 \circ R_{21}^P\right) = [0\ 0.49\ 0\ 0.483\ 0]$$

$$qent_{22} = q_2 = [0\ 1\ 0\ 0\ 0\ 0]$$

O resultado da expansão entre domínios é:

$$qent = (([0\ 1\ 0\ 0\ 0] \vee [0\ 0.357\ 0\ 0\ 0\ 0]) \wedge ([0\ 0.49\ 0\ 0.483\ 0] \vee [0\ 1\ 0\ 0\ 0\ 0]))$$

(3) Calcular a expansão da consulta intra-domínios originando:
$qexp^T = \left(\left(qexp_{11}^T \vee qexp_{12}^T\right) \wedge \left(qexp_{21}^T \vee qexp_{22}^T\right)\right)$.

Seguindo a Eq. 4.6 tem-se que:

$$\begin{aligned} qexp_{11}^T &= \max\left(qent_{11}^T, qent_{11(S)}^T, qent_{11(G)}^T\right) \\ &= \max\left(qent_{11}^T, 0.7\left(R_{S1}^* \circ qent_{11}^T\right), 0.3\left(R_{G1}^* \circ qent_{11}^T\right)\right) \\ &= \max\left(\begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0.224\\0\\0\end{bmatrix}, \begin{bmatrix}0.135\\0\\0\\0\\0\end{bmatrix}\right) = \begin{bmatrix}0.135\\1\\0.224\\0\\0\end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
qexp_{12}^{T} &= \max\left(qent_{12}^{T},\ qent_{12(S)}^{T},\ qent_{12(G)}^{T}\right) \\
&= \max\left(qent_{12}^{T},\ 0.7\left(R_{S2}^{*} \circ qent_{12}^{T}\right),\ 0.3\left(R_{G2}^{*} \circ qent_{12}^{T}\right)\right) \\
&= \max\left(\begin{bmatrix} 0 \\ 0.357 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0.249 \\ 0.249 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.107 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 0.107 \\ 0.357 \\ 0.249 \\ 0.249 \\ 0 \\ 0 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
qexp_{21}^{T} &= \max\left(qent_{21}^{T},\ qent_{21(S)}^{T},\ qent_{21(G)}^{T}\right) \\
&= \max\left(qent_{21}^{T},\ 0.7\left(R_{S1}^{*} \circ qent_{21}^{T}\right),\ 0.3\left(R_{G1}^{*} \circ qent_{21}^{T}\right)\right) \\
&= \max\left(\begin{bmatrix} 0 \\ 0.49 \\ 0 \\ 0.483 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0.224 \\ 0 \\ 0.147 \end{bmatrix}, \begin{bmatrix} 0.135 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 0.135 \\ 0.49 \\ 0.224 \\ 0.483 \\ 0.147 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
qexp_{22}^{T} &= \max\left(qent_{22}^{T}, qent_{22(S)}^{T}, qent_{22(G)}^{T}\right) \\
&= \max\left(qent_{22}^{T}, 0.7\left(R_{S2}^{*} \circ qent_{22}^{T}\right), 0.3\left(R_{G2}^{*} \circ qent_{22}^{T}\right)\right) \\
&= \max\left(\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0.399 \\ 0.301 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.192 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 0.192 \\ 1 \\ 0.399 \\ 0.301 \\ 0 \\ 0 \end{bmatrix}
\end{aligned}
$$

O resultado da expansão intra-domínios é:

$$
qexp^{T} = \left(\left(\begin{bmatrix} 0.135 \\ 1 \\ 0.224 \\ 0 \\ 0 \end{bmatrix} \vee \begin{bmatrix} 0.107 \\ 0.357 \\ 0.249 \\ 0.249 \\ 0 \\ 0 \end{bmatrix}\right) \wedge \left(\begin{bmatrix} 0.135 \\ 0.49 \\ 0.224 \\ 0.483 \\ 0.147 \end{bmatrix} \vee \begin{bmatrix} 0.192 \\ 1 \\ 0.399 \\ 0.301 \\ 0 \\ 0 \end{bmatrix}\right)\right)
$$

No resultado da expansão intra-domínios o fator de expansão correspondente ao conceito Norte,

com os conceitos e seus pesos adicionados à consulta, é dado por:

$$qexp^{T}_{Norte} = \left(\begin{bmatrix} \text{Brasil} : 0.135 \\ \text{Norte} : 1 \\ \text{Pará} : 0.224 \\ \text{Nordeste} : 0 \\ \text{Maranhão} : 0 \end{bmatrix} \vee \begin{bmatrix} \text{Clima} : 0.107 \\ \text{Tropical} : 0.357 \\ \text{Am} : 0.249 \\ \text{Aw} : 0.249 \\ \text{Semi-árido} : 0 \\ \text{BSh} : 0 \end{bmatrix} \right)$$

O fator de expansão correspondente ao conceito Tropical, com os conceitos e seus pesos adicionados à consulta, é dado por:

$$qexp^{T}_{Tropical} = \left(\begin{bmatrix} \text{Brasil} : 0.135 \\ \text{Norte} : 0.49 \\ \text{Pará} : 0.224 \\ \text{Nordeste} : 0.483 \\ \text{Maranhão} : 0.147 \end{bmatrix} \vee \begin{bmatrix} \text{Clima} : 0.192 \\ \text{Tropical} : 1 \\ \text{Am} : 0.399 \\ \text{Aw} : 0.301 \\ \text{Semi-árido} : 0 \\ \text{BSh} : 0 \end{bmatrix} \right)$$

Em ambos os casos os conceitos que possuem o valor 0 associado não são efetivamente adicionados à consulta. Deve-se notar que o processo de expansão preserva o valor 1 atribuído, pelo usuário, aos conceitos presentes na consulta inicial. Os conceitos presentes na consulta inicial sempre terão valor maior que os conceitos adicionados em função do processo de expansão.

(4) Verificar os documentos mais similares à consulta expandida:

$$V = \Big(\Big(U_1\, qexp^{T}_{11} \ \vee\ U_2\, qexp^{T}_{12} \Big) \ \wedge\ \Big(U_1\, qexp^{T}_{21} \ \vee\ U_2\, qexp^{T}_{22} \Big) \Big).$$

Aplicando a Eq. 4.7 tem-se:

$$V = \left(\left(\begin{bmatrix} 0.224 \\ 1.0 \\ 0.135 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \vee \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.249 \\ 0.357 \\ 0.249 \\ 0.107 \\ 0 \\ 0 \end{bmatrix}\right) \wedge \left(\begin{bmatrix} 0.224 \\ 0.49 \\ 0.135 \\ 0.483 \\ 0.147 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \vee \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.399 \\ 1 \\ 0.301 \\ 0.192 \\ 0 \\ 0 \end{bmatrix}\right)\right)$$

O resultado final de $(V)_t$, aplicando-se o limite $t = 0.2$ é dado por:

$$(V)_{0.2} = \left(\begin{bmatrix} 0.224 \\ 1.0 \\ 0.135 \\ 0 \\ 0 \\ 0.249 \\ 0.357 \\ 0.249 \\ 0.107 \\ 0 \\ 0 \end{bmatrix} \wedge \begin{bmatrix} 0.224 \\ 0.49 \\ 0.135 \\ 0.483 \\ 0.147 \\ 0.399 \\ 1.0 \\ 0.301 \\ 0.192 \\ 0 \\ 0 \end{bmatrix}\right)_{0.2} = \left(\begin{bmatrix} 0.224 \\ 0.49 \\ 0.135 \\ 0 \\ 0 \\ 0.249 \\ 0.357 \\ 0.249 \\ 0.107 \\ 0 \\ 0 \end{bmatrix}\right)_{0.2} = \begin{bmatrix} \text{Doc1} : 0.224 \\ \text{Doc2} : 0.49 \\ \text{Doc3} : 0 \\ \text{Doc4} : 0 \\ \text{Doc5} : 0 \\ \text{Doc6} : 0.249 \\ \text{Doc7} : 0.357 \\ \text{Doc8} : 0.249 \\ \text{Doc9} : 0 \\ \text{Doc10} : 0 \\ \text{Doc11} : 0 \end{bmatrix}$$

Após aplicar o valor limite $t = 0.2$ e fazer a ordenação pelos valores de relevância associados aos documentos o conjunto resposta $A = \{\text{Doc2}, \text{Doc7}, \text{Doc6}, \text{Doc8}, \text{Doc1}\}$ de documentos é apresentado para o usuário. Considerando que o conjunto de documentos relevantes para a consulta inicial q_{user} = Norte e Tropical é dado por $R = \{\text{Doc2}, \text{Doc1}, \text{Doc6}\}$ pode-se verificar que o modelo proposto recuperou todos estes documentos e apresentou-os ao usuário. Os documentos Doc8 e Doc7 também foram recuperados mas não fazem parte do conjunto de documentos relevantes. O documento Doc8 está associado ao conceito Aw e o documento Doc7 está associado ao conceito Tropical. Intuitivamente eles poderiam ser relevantes para a consulta do usuário mas neste exemplo este não foi o caso. Considerando o conjunto R de documentos relevantes para a consulta e o conjunto resposta A de documentos retornados para o usuário, a curva com as medidas de precisão *versus* cobertura para

a consulta q_{user} = Norte e Tropical, para o valor de $w_P = 0.7$, é mostrada na Fig. 4.8.

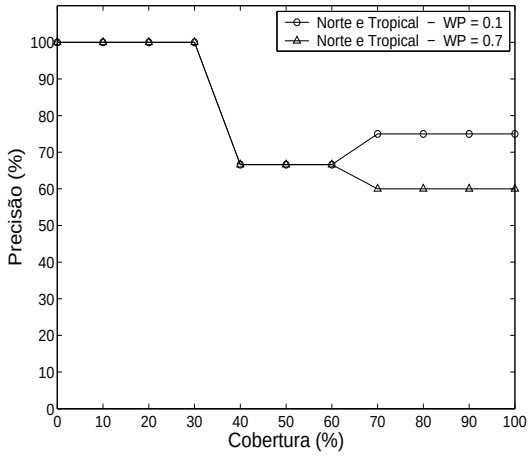

Fig. 4.8: Curva de precisão *versus* cobertura para a consulta inicial do exemplo de uso.

Nos experimentos realizados no modelo verificou-se que os melhores valores das medidas de precisão, para uma mesma medida de cobertura, são obtidos quando o valor de $w_P = 0.1$. No exemplo desenvolvido anteriormente, ao considerar o valor de $w_P = 0.1$ obtém-se o seguinte conjunto de documentos resposta ordenados por seu valor de relevância $A = \{\text{Doc2}, \text{Doc7}, \text{Doc1}, \text{Doc6}, \text{Doc8}\}$. A curva de precisão *versus* cobertura para a consulta inicial q_{user} = Norte e Tropical é mostrada na Fig. 4.8. Deve-se observar que houve uma melhora em alguns valores de precisão ao considerar o valor de $w_P = 0.1$. Em ambos os casos o modelo conseguiu recuperar os documentos considerados relevantes para a consulta inicial. Se o modelo não tivesse utilizado a base de conhecimento composta das múltiplas ontologias relacionadas nenhum documento seria recuperado pois não há um documento que possua, simultaneamente, os conceitos Norte e Tropical.

Neste mesmo exemplo, ao considerar apenas as relações de especialização e generalização *fuzzy* em cada ontologia, sem levar em conta a relação de associação positiva *fuzzy*, nenhum documento é recuperado. Neste caso o valor de $w_P = 0.0$. O resultado final da expansão é dado por:

$$qexp^T = \left(\left(\left(\begin{bmatrix} 0.135 \\ 1 \\ 0.224 \\ 0 \\ 0 \end{bmatrix} \vee \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) \wedge \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \vee \begin{bmatrix} 0.192 \\ 1 \\ 0.399 \\ 0.301 \\ 0 \\ 0 \end{bmatrix}\right)\right)\right)$$

O fator de expansão correspondente ao conceito Norte, com os conceitos e seus pesos adicionados à consulta, é dado por:

$$qexp^{T}_{Norte} = \left(\begin{bmatrix} \text{Brasil} : 0.135 \\ \text{Norte} : 1 \\ \text{Pará} : 0.224 \\ \text{Nordeste} : 0 \\ \text{Maranhão} : 0 \end{bmatrix} \vee \begin{bmatrix} \text{Clima} : 0 \\ \text{Tropical} : 0 \\ \text{Am} : 0 \\ \text{Aw} : 0 \\ \text{Semi-árido} : 0 \\ \text{BSh} : 0 \end{bmatrix} \right)$$

O fator de expansão correspondente ao conceito Tropical, com os conceitos e seus pesos adicionados à consulta, é dado por:

$$qexp^{T}_{Tropical} = \left(\begin{bmatrix} \text{Brasil} : 0 \\ \text{Norte} : 0 \\ \text{Pará} : 0 \\ \text{Nordeste} : 0 \\ \text{Maranhão} : 0 \end{bmatrix} \vee \begin{bmatrix} \text{Clima} : 0.192 \\ \text{Tropical} : 1 \\ \text{Am} : 0.399 \\ \text{Aw} : 0.301 \\ \text{Semi-árido} : 0 \\ \text{BSh} : 0 \end{bmatrix} \right)$$

Para esta consulta os documentos mais similares à mesma são:

$$V = \left(\left(\begin{bmatrix} 0.224 \\ 1.0 \\ 0.135 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \vee \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right) \wedge \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \vee \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.399 \\ 1 \\ 0.301 \\ 0.192 \\ 0 \\ 0 \end{bmatrix} \right) \right)$$

O resultado final de (V) é dado por:

$$(V) = \left(\begin{bmatrix} 0.224 \\ 1.0 \\ 0.135 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \wedge \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.399 \\ 1.0 \\ 0.301 \\ 0.192 \\ 0 \\ 0 \end{bmatrix} \right) = \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right)$$

Este resultado ilustra que a exploração da associação positiva *fuzzy*, entre as ontologias, possibilita a recuperação de documentos mais relevantes à consulta.

4.5 Resumo do Capítulo

Este capítulo apresentou o modelo *fuzzy* para recuperação de informação utilizando múltiplas ontologias relacionadas. Cada ontologia representa os conceitos de um domínio de conhecimento sendo que elas podem estar relacionadas para denotar os relacionamentos existentes entre os conceitos dos domínios distintos. O modelo propõe uma forma de organização de conhecimento onde as ontologias e seus relacionamentos são representados por relações *fuzzy*. Baseado nesta forma de organização de conhecimento um novo processo de expansão da consulta do usuário também é desenvolvido.

Para ilustrar a utilidade do modelo foi apresentado um exemplo de uso do mesmo. Os resultados mostram que o uso da base de conhecimento, composta por múltiplas ontologias relacionadas, e o método de expansão da consulta desenvolvido possibilitaram recuperar mais documentos relevantes à consulta do usuário. Para a consulta considerada no exemplo, a recuperação realizada com o uso apenas das palavras-chaves da consulta não retornaria nenhum documento. Se a expansão da consulta considerasse cada ontologia de maneira separada e explorasse apenas as associações de especialização *fuzzy* e a generalização *fuzzy* de cada ontologia também não haveria o retorno de documentos relevantes. A organização do conhecimento como ontologias relacionadas e a consideração deste conhecimento na expansão da consulta, pela utilização da associação positiva *fuzzy* juntamente com as associações de especialização e generalização *fuzzy*, é que possibilitou a recuperação de documentos relevantes à consulta do usuário.

Capítulo 5

Resultados Experimentais

A avaliação experimental foi realizada para testar o desempenho do modelo proposto, avaliar a tendência do seu comportamento e compará-lo com um modelo com enfoque similar, ou seja, o modelo de recuperação de informação baseado na rede de conceitos *fuzzy*, apresentado no Cap. 3. Ambos os modelos consideram a coleção de documentos indexada pelos conceitos e a existência da associação de especialização *fuzzy*, associação de generalização *fuzzy* e a associação positiva *fuzzy* entre os conceitos para realizar a expansão.

Além da avaliação da recuperação de informação entre o modelo proposto e o modelo de rede de conceitos *fuzzy* também foi avaliado o método de expansão da consulta proposto considerando os documentos indexados pela máquina de busca Lucene do Projeto Apache. Uma vez que a consulta inicial é expandida ela é traduzida para a representação de consulta própria do Apache Lucene. O objetivo desta avaliação é verificar o comportamento do método de expansão de consulta considerando uma máquina de busca comercial. Os resultados obtidos com a máquina de busca do Apache Lucene também são comparados com os resultados obtidos pelos modelos avaliados.

A avaliação experimental foi realizada considerando uma amostra de 129 documentos selecionados da coleção de documentos do domínio da Agrometeorologia no Brasil, mantida pela Embrapa. A amostra constituída por 129 documentos inclui todos os conceitos das ontologias construídas e acredita-se que seja o mínimo necessário para provar o potencial da forma de organização de conhecimento e do método de expansão de consulta propostos nesta tese. Além disto este tamanho de amostra viabiliza a associação de documentos relevantes à cada consulta por um especialista de domínio.

Para o domínio de agrometeorologia foram levantados os conceitos de dois domínios de conhecimento distintos e relacionados que estão associados à agrometeorologia: domínio de divisão territorial e domínio de clima. Estes conceitos foram utilizados para montar as bases de conhecimento dos modelos estudados. Para o modelo *fuzzy* baseado em múltiplas ontologias relacionadas foi desenvolvida

uma ontologia *lightweight* constituída pelos conceitos que modelam o domínio referente à divisão territorial do Brasil e uma ontologia *lightweight* constituída pelos conceitos que modelam o domínio de clima que atua sobre o território do Brasil. Ambas as ontologias foram construídas manualmente. Para o modelo de rede de conceitos *fuzzy* o mesmo conjunto de conceitos foi utilizado para construir a base de conhecimento automaticamente.

O conjunto de consultas contém 83 consultas e é composto por consultas contendo apenas um conceito de cada ontologia assim como consultas contendo combinações de dois conceitos de ambas as ontologias nos diferentes níveis. Os conceitos são conectados com operadores AND ou OR. O Apêndice A descreve a construção das ontologias, a coleção de documentos e a elaboração do conjunto de consultas. As medidas cobertura e precisão, descritas na seção 2.4, são consideradas para avaliar o desempenho de todos os modelos .

Para controlar a influência de cada tipo de associação *fuzzy* os modelos atribuem pesos para cada associação permitindo um balanceamento entre elas. A atribuição destes pesos reflete no resultado final no que diz respeito às medidas de cobertura e precisão. Considerando várias combinações dos pesos, após inúmeros testes, os modelos mostraram um padrão de comportamento para as medidas de cobertura e precisão.

Este capítulo discute o desempenho dos modelos considerando as combinações de pesos. Inicialmente são apresentados o modelo ontológico relacional *fuzzy*, o modelo de rede de conceitos *fuzzy* e o Apache Lucene. O modelo ontológico relacional *fuzzy* e o modelo de rede de conceitos *fuzzy* forneceram uma base inicial para o desenvolvimento do modelo de múltiplas ontologias relacionadas proposto nesta tese. Dentre os três modelos, o modelo de rede de conceitos *fuzzy* e o Apache Lucene foram utilizados para comparação com o modelo proposto. Após a apresentação dos modelos são apresentados os casos de teste construídos para validação e os resultados obtidos pelo modelo de múltiplas ontologias relacionadas e os modelos implementados para comparação.

5.1 Modelo Ontológico Relacional *Fuzzy*

O modelo FROM - *Fuzzy Relational Ontological Model* [100, 101, 102] representa a estrutura conceitual através de uma ontologia de dois níveis compostos por categorias e palavras-chaves que representam os conceitos do domínio. As categorias denotam os conceitos mais gerais e as palavras-chaves denotam os conceitos mais específicos. As palavras-chaves e as categorias estão associados por relações *fuzzy* que determinam o grau de associação entre elas. Não existem associações entre duas palavras-chaves ou duas categorias. Os documentos estão associados tanto às categorias quanto às palavras chaves de forma independente sendo que o grau de associação entre documentos e palavras-chaves ou categorias também é dado por relações *fuzzy*. Baseado nas relações *fuzzy* expres-

sas na ontologia o sistema faz a expansão dos conceitos presentes na consulta do usuário. A consulta expandida é então aplicada à base de documentos.

Na ontologia relacional o conjunto de palavras-chaves é dado por K, o conjunto de categorias é dado por C e a ontologia *fuzzy* é representada pela relação *fuzzy* $R_0 : K \times C$ onde $R_0(k_i, c_j) = r_{ij} \in [0,\ 1]$ é o grau de relevância entre a palavra-chave $k_i \in K$ com respeito à categoria $c_j \in C$.

Os documentos estão representados pelo conjunto D. O conjunto de documentos está associado às palavras-chaves pela relação *fuzzy* $D_k : D \times K$. Dado um documento $d_w \in D$, então $D_k(d_w, k_i) \in [0,\ 1]$ expressa a relevância de palavra-chave k_i como um descritor do conteúdo do documento d_w. De forma semelhante a relação *fuzzy* $D_c : D \times C$ representa os relacionamentos entre os documentos e as categorias. Assim, $D_c(d_w,\ c_j) \in [0,\ 1]$ expressa a relevância da categoria c_j em descrever o conteúdo do documento d_w.

As consultas dos usuários podem ser compostas apenas por palavras-chaves ou categorias ou ambas. Uma vez que as consultas são aplicadas às representações dos documentos e considerando que os documentos possuem representações independentes relacionando-os às categorias e às palavras-chaves então as consultas são divididas em uma representação considerando apenas as categorias e outra considerando apenas as palavras-chaves. O sistema FROM considera que as consultas são transformadas na forma canônica de um produto de somas, ou seja, um conjunto de sub-consultas conectadas pelo operador lógico AND sendo cada sub-consulta composta por termos conectados pelo operador lógico OR. Por exemplo, dados os conjuntos $C = \{c_1,\ c_2\}$ e $K = \{k_1,\ k_2,\ k_3\}$, um consulta válida seria $q = (k_1 \vee k_2 \vee c_2) \wedge (k_3 \vee c_1)$. O resultado da consulta q é a interseção dos resultados de cada de suas sub-consultas. Desta forma o trabalho se concentra em demonstrar o tratamento para uma sub-consulta. Ao fazer a consulta o usuário atribui valor (1) para indicar a presença ou (0) para indicar a ausência de um conceito nos documentos a serem recuperados. A seguir a consulta é dividida em uma representação relacionada às palavras chaves e em uma representação relacionada às categorias. Considerando os conjuntos de categorias, C, e palavras-chaves, K, citados anteriormente, a sub-consulta $q = (k_1 \vee k_2 \vee c_2)$ seria dividida nas representações $q_k = [\,1\,1\,0\,]$ e $q_c = [\,0\,1\,]$.

A partir da representação relacionada às palavras-chaves q_k e da relação *fuzzy* R_0 é possível derivar um conjunto *fuzzy* das categorias que estão implicitamente associadas às palavras-chaves da consulta. A composição max-min dada por $q_k \circ R_0$ gera um conjunto *fuzzy* de dimensão m onde $m = |C|$ é o número de categorias. Cada elemento deste conjunto indica o grau de cada categoria para a consulta em função das relações existentes entre palavras-chaves e categorias na ontologia relacional *fuzzy*. No conjunto *fuzzy* das categorias inferidas aquelas que possuírem um grau abaixo de um valor limite dado por t serão descartadas por não contribuírem para o resultado final na recuperação de documentos. Neste caso t representa o α-cut para o conjunto de categorias *fuzzy*. A representação expandida da

consulta, pelas categorias, é dada por $f_c = q_c \cup (q_k \circ R_0)_t$

Da mesma forma a composição max-min dada por $R_0 \circ q_c^{'}$, onde $q_c^{'}$ indica a transposta de q_c, gera um conjunto *fuzzy* de dimensão n onde $n = |K|$ é o número de palavras-chaves. Cada elemento deste conjunto indica o grau de relevância de cada palavra-chave para a consulta em função das relações existentes entre categorias e palavras-chaves na ontologia relacional *fuzzy*. Considerando o valor limite de t que representa o α-cut para o conjunto de palavras-chaves *fuzzy*, a representação expandida da consulta, pelas palavras-chaves, é dada por $f_k = q_k^{'} \cup \left(R_0 \circ q_c^{'}\right)_t$

As consultas expandidas f_k e f_c são aplicadas às representações de documentos dadas por D_k e D_c, respectivamente, para calcular a relevância dos mesmos em relação às consultas. A composição $V_k = D_k \circ f_k$ gera a relevância de todos os documentos $d \in D$ baseado nas palavras-chaves. A composição $V_c = D_c \circ f_c^{'}$ representa a relevância dos documentos baseado nas categorias. Finalmente os conjuntos *fuzzy* dados por V_k e V_c são combinados para gerar a relevância final dos documentos. Uma vez que os elementos de cada sub-consulta são conectados pelo operador lógico OR então a relevância final é dada por $V = V_k \cup V_c$.

Uma redução no número de documentos recuperados é obtida através da definição de um outro limite r, para realizar um α-cut no conjunto *fuzzy* V dado por $(V)_r = (V_k \cup V_c)_r$. Os documentos cujo valor dado em $(V)_r$ forem diferentes de zero, ou seja, $(V)_r \neq 0$ são apresentados ao usuário em ordem descrescente do valor de relevância.

5.2 Modelo de Rede de Conceitos *Fuzzy*

O modelo de rede de conceitos *fuzzy* é apresentado com mais detalhes nesta seção pois constitui o modelo *fuzzy* implementado para ser comparado ao modelo de recuperação de informação proposto neste trabalho. No sistema de rede de conceitos *fuzzy* [19, 56, 57] a estrutura conceitual consiste em uma rede de multi-relacionamentos *fuzzy* onde os conceitos estão representados pelos nós e os múltiplos relacionamentos entre os conceitos são dados pelas ligações entre os nós. Existem três tipos de relacionamentos na rede: associação positiva *fuzzy*, associação de generalização *fuzzy* e associação de especialização *fuzzy*.

Os documentos estão associados aos conceitos da rede através de uma relação *fuzzy* que indica a força da associação entre cada documento e o conceitos da rede. Através de inferências baseadas na rede de conceitos de multi-relacionamentos *fuzzy* o sistema de recuperação de informação pode recuperar documentos contendo conceitos que não foram diretamente especificados pelo usuário mas que podem estar relacionados com a sua consulta. Para realizar as inferências, os relacionamentos entre os conceitos são representados por relações que representam os seus graus de relacionamento *fuzzy*. As inferências são realizadas através do cálculo do fecho transitivo das relações. Ao calcular o

fecho transitivo os relacionamentos *fuzzy* implícitos entre os conceitos são obtidos. Uma vez que os relacionamentos implícitos entre os conceitos são calculados as associações dos documentos também são ajustadas para refletir a influência das inferências dos conceitos nestas associações. A partir deste ajuste uma consulta do usuário pode recuperar documentos que não tinham sido inicialmente associados aos conceitos.

5.2.1 Construção da Rede de Conceitos *Fuzzy*

No sistema de rede de conceitos *fuzzy* a estrutura conceitual consiste em uma rede de multi-relacionamentos *fuzzy* onde os conceitos estão representados pelos nós e os múltiplos relacionamentos entre os conceitos são dados pelas ligações entre os nós. A rede de conceitos é representada pelo conjunto de relações $V_r : C \times C \rightarrow [0,1]$ onde C representa o conjunto de conceitos da estrutura conceitual e $r \in \{P, G, S\}$ são os diversos tipos de relacionamentos conforme descrito a seguir:

- Associação Positiva *Fuzzy* $(V_P : C \times C \rightarrow [0,1])$: relaciona conceitos que possuem significados similares em alguns contextos. Exemplo: pessoa $\leftrightarrow$ indivíduo.
- Associação de Generalização *Fuzzy* $(V_G : C \times C \rightarrow [0,1])$: um conceito é uma generalização de outro conceito se ele consitir daquele conceito $(\text{máquina} \rightarrow \text{parafuso})$ ou se ele incluir aquele conceito no sentido partitivo $(\text{veículo} \rightarrow \text{carro})$.
- Associação de Especialização *Fuzzy* $(V_S : C \times C \rightarrow [0,1])$: é o inverso da relação de generalização.

Nas relações $V_r,\ r \in \{P, G, S\}$ valores iguais a zero indicam a ausência do relacionamento, valores iguais a 1 indicam uma relação forte e valores intermediários indicam os graus da relação situados entre inexistente (0) e fortemente relacionados (1).

A rede de conceitos é construída automaticamente baseada na co-ocorrência sintática. Cada documento na coleção é relacionado a um ou mais conceitos do conjunto de conceitos C. O conteúdo dos documentos é dado pelas palavras existentes nos mesmos. Se um documento d_i é associado a um conceito c_j então palavras contidas no documento d_i também estão contidas no conceito c_j. Desta forma cada conceito engloba um conjunto de palavras derivadas de um conjunto de documentos que contém estas palavras. Comparando os conjuntos de palavras contidas em cada conceito é possível estabelecer as relações *fuzzy* entre os conceitos segundo o seguinte algoritmo:

1. Extrair as palavras dos documentos.
2. Calcular os pesos das palavras nos documentos.

3. Calcular os pesos das palavras nos conceitos.

4. Calcular os pesos das ligações entre conceitos e documentos.

5. Calcular os tipos de associações e seus graus de associação entre os conceitos.

Extrair as palavras dos documentos

As palavras presentes nos documentos são extraídas pelo uso de um parser e algumas informações associadas às palavras são registradas para utilização nos próximos passos: documentos onde as palavras ocorrem e o número de vezes em que cada palavra ocorre em cada documento. As palavras que não agregam informação sobre o conteúdo do documento são colocadas em uma lista de *stopwords* para serem descartadas pelo parser. Desta forma o espaço de palavras é reduzido.

Calcular os pesos das palavras nos documentos

A equação para calcular o peso de uma palavra em um documento é baseado método $tf \times idf$ normalizado. Se uma palavra t aparece mais freqüentemente do que outras no documento d_i então o peso da palavra t no documento d_i é proporcional à sua ocorrência no documento d_i. Entretanto se a palavra t também aparece freqüentemente em outros documentos na coleção então a importância da palavra t no documento é reduzida. Assim o peso da palavra t no documento d_i é inversamente proporcional à sua freqüência nos documentos da coleção. O peso $w\_word\_document\,(t,\ d_i)$ é dado pela Eq. 5.1 onde tf_{it} é a frequencia da palavra t no documento d_i, df_t é o número de documentos contendo a palavra t, L é o número de palavras no documento d_i e N é o numero de documentos na coleção. Quanto maior o valor de $w\_word\_document\,(t,\ d_i)$ maior é a importância da palavra t no documento d_i. Pela Eq. 5.1 pode-se notar que $0 \leq w\_word\_document\,(t,\ d_i) \leq 1$ sendo um valor normalizado.

$$w\_word\_document\,(t,\ d_i) = \frac{\left(0,5 + 0.5\frac{tf_{it}}{\max_{K=1,2,\ldots,L} tf_{ik}}\right)\log\frac{N}{df_t}}{\max_{j=1,2,\ldots,L}\left\{\left(0,5 + 0.5\frac{tf_{it}}{\max_{K=1,2,\ldots,L} tf_{ik}}\right)\log\frac{N}{df_j}\right\}} \tag{5.1}$$

Calcular os pesos das palavras nos conceitos

O peso $w\_word\_concept\,(t,\ c)$ da palavra t para o conceito c é calculado pela Eq. 5.2 onde m é o número de documentos que contém a palavra t e pertencem ao conceito c. Pela Eq. 5.2 pode-se notar que quanto maior o peso da palavra t em documentos que contém o conceito c maior o peso da palavra t com relação ao conceito c. Se $w\_word\_concept\,(t,\ c) > 0$ diz-se que a palavra t está

contida no conceito c.

$$w\_word\_concept\,(t,\ c)\ =\ \frac{\sum_{i=1}^{m}\ w\_word\_document\,(t,\ d_i)}{m} \tag{5.2}$$

Calcular os pesos das ligações entre conceitos e documentos

Se a maioria das palavras contidas no documento d_i possuem um peso alto com relação ao conceito c então o peso do documento d_i com relação ao conceito c deve ser alto. O peso do documento d_i com relação ao conceito c, $w\_document\_concept\,(d_i,\ c)$, é dado pela Eq. 5.3. O número de palavras no documento d_i é dado por k.

$$w\_document\_concept\,(d_i,\ c)\ =\ \frac{\sum_{j=1}^{k}\ w\_word\_concept\,(t_j,\ c)}{k} \tag{5.3}$$

Calcular os tipos de associações e seus graus de associação entre os conceitos

Uma vez que cada conceito c contém um conjunto particular de palavras do conjunto de palavras W então é possível utilizar uma função de mapeamento M que representa cada conceito pelo seu conjunto *fuzzy* no conjunto W. A função de mapeamento é dada pela Eq. 5.4 onde w_{ij} é o peso da palavra t_j no conceito c_i e h é o número de palavras no conjunto de palavras W. Se $w_{ij} > 0$ então a palavra t_j está contida no conceito c_i. Seja $|M(c_i)| = \sum_{j=1,2,\cdots,h} w_{ij}$ então $|M(c_i)|$ é a cardinalidade do conjunto *fuzzy* $M(c_i)$ do conceito c_i no conjunto de palavras W.

$$M\,(c_i)\ =\ w_{i1}/t_1\ +\ w_{i2}/t_2\ +\ \ldots\ +\ w_{ih}/t_h \tag{5.4}$$

Sejam c_i e c_j dois conceitos arbitrários no conjunto de conceitos C. Para estabelecer o tipo de associação entre os conceitos as condições a seguir devem ser observadas.

1. Se os conceitos c_i e c_j contém palavras diferentes então eles não estão relacionados.

2. Se o conceito c_i e o conceito c_j contém quase as mesmas palavras mas os pesos das palavras no conceito c_i são maiores que o seu peso no conceito c_j então o conceito c_i domina o conceito c_j e deve ser mais geral que o conceito c_j.

3. Se o conceito c_i e o conceito c_j contém quase as mesmas palavras mas os pesos das palavras no conceito c_i são menores que o seu peso no conceito c_j então o conceito c_i é dominado pelo conceito c_j e deve ser mais específico que o conceito c_j.

4. Se a maioria das palavras contidas no conceito c_j também estão contidas no conceito c_i mas muitas palavras contidas no conceito c_i não estão contidas no conceito c_j, então o conceito c_i

representa mais aspectos que o conceito c_j e deve ser mais geral que o conceito c_j.

5. Se a maioria das palavras contidas no conceito c_i também estão contidas no conceito c_j mas muitas palavras contidas no conceito c_j não estão contidas no conceito c_i, então o conceito c_i representa menos aspectos que o conceito c_j e deve ser mais específico que o conceito c_j.

6. Se o conceito c_i e o conceito c_j contém quase as mesmas palavras e os pesos das palavras são similares em ambos os conceitos então estes dois conceitos devem ser similares entre si e possuem um relação de associação positiva *fuzzy*.

Dadas estas condições é possível definir as relações de associação *fuzzy* e os graus de associação entre os conceitos comparando os seus conjuntos *fuzzy* correspondentes no conjunto de palavras W. O grau em que o conceito c_j é mais geral que o conceito c_i é dado por $G(c_i, c_j)$ e é igual ao grau em que $M(c_i)$ está contido em $M(c_j)$. Um método para calcular $G(c_i, c_j)$ é através da medida de *subsethood* dada pela Eq. 5.5 onde w_{ki} é o peso da palavra t_k no conceito c_i, w_{kj} é o peso da palavra t_k no conceito c_j, $WC(c_i)$ é o número de palavras contidas no conceito c_i, $WC(c_j)$ é o número de palavras contidas no conceito c_j e h é o número de palavras no conjunto de palavras W.

$$G\left(c_i, c_j\right) = \begin{cases} \left(\frac{|M(c_i) \bigcap M(c_j)|}{|M(c_i)|}\right)^{\frac{WC(c_i)}{\max\left(WC(c_i), WC\left(c_j\right)\right)}} & , \quad if\ M\left(c_j\right) \neq \oslash \\ 1 & , \quad if\ M\left(c_j\right) = \oslash \end{cases} \tag{5.5}$$

Na Eq. 5.5 o cálculo do valor de $G\left(c_i, c_j\right)$ quando $M\left(c_j\right) \neq \oslash$ é dado pela Eq. 5.6.

$$\left(\frac{|M(c_i) \bigcap M(c_j)|}{|M(c_i)|}\right)^{\frac{WC(c_i)}{\max\left(WC(c_i), WC\left(c_j\right)\right)}} = \left(\frac{\sum_{k=1}^{h} \min\left(w_{ki}, w_{kj}\right)}{\sum_{k=1}^{h} w_{ki}}\right)^{\frac{WC(c_i)}{\max\left(WC(c_i), WC\left(c_j\right)\right)}} \tag{5.6}$$

O grau do conceito c_j contido no conceito c_i (isto é, o conceito c_j é mais específico que o conceito c_i) é dado por $S(c_i, c_j)$. Uma vez que a relação de associação de especialização *fuzzy* é o inverso da relação de generalização *fuzzy* então seu valor é dado pela Eq. 5.7.

$$S(c_i, c_j) \ = \ G(c_j, c_i) \tag{5.7}$$

O grau de associação positiva *fuzzy* entre o conceito c_i e o conceito c_j, dado por $P(c_i, c_j)$, é calculado pela Eq. 5.8. Neste caso se ambos os valores de $G(c_i, c_j)$ e $S(c_i, c_j)$ são altos então o valor de $P(c_i, c_j)$ também é alto. Isto significa que se o conceito c_i e o conceito c_j contém a maioria das mesmas palavras com um peso similar então o grau de associação positiva *fuzzy* entre o conceito c_i e o conceito c_j é alto.

$$P(c_i, c_j) \ = \min(G(c_i, c_j),\ S(c_i, c_j)) \tag{5.8}$$

5.2.2 Associação dos Documentos aos Conceitos

O conjunto de documentos da coleção D está associado ao conjunto de conceitos da estrutura C por uma relação *fuzzy* $U : D \times C \rightarrow [0, 1]$ que indica o grau de associação entre o conceito e o documento. Cada documento possui um peso associado ao conceito significando o quanto o conceito é importante no documento. Valores iguais a 0 indicam a ausência de associação entre o documento e o conceito, valores iguais a 1 indicam uma relação forte de associação entre o documento e o conceito e valores intermediários indicam os graus de associação situados entre ausente (0) e fortemente associados (1).

Por meio dos relacionamentos existentes na rede de conceitos o sistema infere novos relacionamentos explorando as relações implícitas entre os conceitos mesmo que elas não tenham sido especificadas inicialmente pelo especialista do domínio. Desta forma um documento que antes não era associado a um determinado conceito pode vir a ter um grau de associação maior que 0 em função das inferências realizadas na rede. Deve-se observar que a associação dos documentos a novos conceitos passa a existir em função das inferências realizadas na rede de conceitos. A inferência na rede de conceitos é realizada através do cálculo dos fechos transitivos para cada uma das relações $V_r,\ r \in \{P, G, S\}$ resultando o fecho transitivo V_r^* de V_r.

Utilizando a relação $U : D \times C$ e o fecho transitivo de cada relacionamento, V_r^*, tem-se as relações de documentos expandidas U_r^*, $U_r^* = U \otimes V_r^*$, que refletem as relações inferidas entre os documentos e os conceitos do domínio considerando os tipos de relacionamentos $r \in \{P, G, S\}$. O cálculo de U_r^* é realizado como na Eq. 5.9 onde u_{ij} é um elemento de $U, 1 \leq i \leq m$, e $1 \leq j \leq n$ sendo m o número de documentos e n o número de conceitos. v_{ij} é um elemento de $V_r^*, 1 \leq i,\ j \leq m$; $\bigvee$ é o operador $\max$ e "$\cdot$" é o produto aritmético. Estas relações expandidas dos documentos vão constituir a base para as medidas de similaridade entre as consultas do usuário e os documentos.

$$U_r^* = U \otimes V_r^* = \begin{bmatrix} \bigvee_{i=1,\ldots,n}(u_{1i} \cdot v_{i1}) & \bigvee_{i=1,\ldots,n}(u_{1i} \cdot v_{i2}) & \cdots & \bigvee_{i=1,\ldots,n}(u_{1i} \cdot v_{in}) \\ \bigvee_{i=1,\ldots,n}(u_{2i} \cdot v_{i1}) & \bigvee_{i=1,\ldots,n}(u_{2i} \cdot v_{i2}) & \cdots & \bigvee_{i=1,\ldots,n}(u_{2i} \cdot v_{in}) \\ \vdots & \vdots & \vdots & \vdots \\ \bigvee_{i=1,\ldots,n}(u_{mi} \cdot v_{i1}) & \bigvee_{i=1,\ldots,n}(u_{mi} \cdot v_{i2}) & \cdots & \bigvee_{i=1,\ldots,n}(u_{mi} \cdot v_{in}) \end{bmatrix} \tag{5.9}$$

5.2.3 Especificação da Consulta

Uma consulta do usuário é expressa pelo vetor descritor $\bar{q} = \langle x_1, x_2, \cdots, x_n \rangle$ onde $x_i \in [0, 1]$ indica o grau desejado de associação entre o documento com relação ao conceito $c_i \in C$, $1 \leq i \leq n$ e n é o número de conceitos. Se $x_i = {}'-'$ indica que o grau de associação dos documentos com relação ao conceito c_i deve ser desconsiderado.

Na consulta o usuário especifica os conceitos que ele deseja que estejam representados nos documentos e associa um valor no intervalo [0,1] a cada conceito. Este peso vai significar o quanto o conceito é importante nos documentos a serem recuperados. Se o valor do peso é zero significa que o conceito não deve estar presente no documento. Caso a presença ou ausência de um conceito seja irrelevante o usuário coloca um "-" no lugar do mesmo. Quanto maior o valor do peso relativo a um conceito mais documentos que também tenham um alto grau de associação ao conceito devem ser recuperados. Quando a consulta é submetida ao sistema ele verifica a similaridade entre os pesos dos conceitos na consulta e os pesos dos mesmos conceitos nos documentos.

O vetor descritor de um documento é dado por $\bar{dr}_i = \langle s_{i1}, s_{i2}, \cdots, s_{in} \rangle$ onde $s_{ij} \in [0,1]$, $1 \leq i \leq m$, $1 \leq j \leq n$, m é o número de documentos, n é o número de conceitos e $r \in \{P, G, S\}$. $\bar{dr}_i$ representa a $i - ésima$ linha da relação U_r^*.

O grau de satisfação DS com que um documento d_i satisfaz à consulta Q pela relação *fuzzy* r é dado pela Eq. 5.10, onde $T(x,\ y) = 1 - |x - y|$, $T(x,\ y) \in [0,1]$, $DS_r\,(d_i) \in [0,1]$, $1 \leq i \leq m$ e K é o número de conceitos considerados na consulta do usuário.

$$DS_r\,(d_i) = \frac{\sum_{q(\bar{j}) \neq '' - '' \ and \ j=1,\,2,\,\cdots,\,n} T(s_{ij},\ x_j)}{K} \tag{5.10}$$

Quando maior o valor de $DS_r\,(d_i)$ maior é o grau com que um documento d_i satisfaz à consulta do usuário pela relação r, $r \in \{P, G, S\}$. Os graus de satisfação com os quais um documento satisfaz à consulta pelos diferentes tipos de relacionamentos r são agregados para obter o grau de satisfação geral do documento com relação à consulta. O usuário atribui um peso de importância w_r para cada grau de satisfação $DS_r\,(d_i)$ com o qual o documento d_i satisfaz à consulta do usuário pelo relacionamento $r \in \{P, G, S\}$, onde $0 \leq w_r \leq 1$. Quanto maior o valor de w_r maior é a importância do relacionamento dado por r. Uma vez estabelecidos os valores dos pesos w_r o grau de satisfação geral com que um documento d_i satisfaz à consulta do usuário é calculado pela Eq. 5.11.

$$DS(d_i) = w_P\ DS_P(d_i) + w_G\ DS_G(d_i) + w_S\ DS_S(d_i) \tag{5.11}$$

Uma vez que o grau de satisfação geral de cada documento é calculado os documentos são apresentados para o usuário em ordem decrescente do valor do seu grau de satisfação geral.

5.3 Máquina de Busca Apache Lucene

Esta seção apresenta a máquina de busca Apache Lucene de forma concisa pois ela foi utilizada para que o método de expansão de consulta, proposto na tese, pudesse ser testado em uma máquina de busca comercial.

A máquina de busca Apache Lucene consiste em uma biblioteca na linguagem Java [5], disponível como software livre, que permite prover uma aplicação com capacidade de indexação e busca em texto [46, 50]. Estas funcionalidades de busca são implementadas pelo uso de uma API (*Application Program Interface*) provida pelo Projeto Apache. As principais classes providas para criar uma aplicação com a máquina de busca Lucene são:

- *Document*: classe que representa um documento no Lucene. O *Document* é o objeto utilizado para indexar um documento. Quando a busca é realizada o documento retornado também é representado pelo objeto *Document*.

- *Field*: classe que representa as seções do documento como título, texto, palavras-chaves e outras.

- *Analyzer*: classe que vai extrair as palavras (*tokens*) dos documentos para serem indexados. Ele considera uma lista de *stopwords* para não indexar palavras comuns que não carregam significado como artigos e pronomes.

- *IndexWriter*: classe utilizada para criar e manter os arquivos de índices dos documentos.

- *IndexSearcher*: classe utilizada para fazer a busca no arquivo de índices.

- *Query*: classe que armazena a representação de uma consulta a ser submetida ao Lucene.

- *QueryParser*: classe utilizada para criar um objeto do tipo *Query*, a partir da consulta inicial do usário, a ser submetida ao Lucene.

- *Hits*: classe que contém a lista de documentos retornados quando o objeto *Query* é submetido ao *IndexSearcher*.

O Lucene permite especificar vários tipos de expressão de busca [7]. Entre suas opções ele permite fazer um *boost* em um conceito de busca aumentando a relevância dos documentos indexados pelo conceito. Para fazer o *boost* em um conceito é necessário usar o símbolo (^) com o fator de *boost* (um número) no final do conceito a ser pesquisado. Quanto maior o fator de *boost* maior a relevância do conceito. No modelo proposto o fator de *boost* é um valor no intervalo $[0, 1]$ e representa o valor de associação dos conceitos calculados no processo de expansão para um conceito presente na consulta inicial . Na consulta $q_{user} =$ Norte e Tropical, do exemplo da seção 4.4, o conceito Norte é expandido como:

$$qexp^{T}_{Norte} = \left(\begin{bmatrix} \text{Brasil} : 0.135 \\ \text{Norte} : 1 \\ \text{Pará} : 0.224 \\ \text{Nordeste} : 0 \\ \text{Maranhão} : 0 \end{bmatrix} \vee \begin{bmatrix} \text{Clima} : 0.107 \\ \text{Tropical} : 0.357 \\ \text{Am} : 0.249 \\ \text{Aw} : 0.249 \\ \text{Semi-árido} : 0 \end{bmatrix} \right)$$

A expansão $qexp^{T}_{Norte}$ será traduzida, para a representação de consulta do Apache Lucene, como:

$(\text{Brasil\^{}}0.135 \text{ or Norte\^{}}1.0 \text{ or Pará\^{}}0.224)$ or

$(\text{Clima\^{}}0.107 \text{ or Tropical\^{}}0.357 \text{ or Am\^{}}0.249 \text{ or Aw\^{}}0.249)$.

No caso do Lucene, além da indexação considerar a freqüência com que um termo ocorre em um documento e o inverso da freqüência do termo no documento ela também considera o tamanho do texto onde o termo ocorreu. Desta forma se um termo t_1 aparece n vezes em um documento D_1 composto de X termos e n vezes em um documento D_2 composto por Y termos onde $|X| < |Y|$ então o peso de t_1 em D_1 é maior que o peso de t_1 em D_2. Este fator extra no cálculo dos pesos dos conceitos, durante a indexação, acarreta maior variação nas curvas de precisão *versus* cobertura para o Apache Lucene.

Um exemplo de como o modelo de indexação do Lucene influencia o valor da precisão é dado a seguir. Supor que a busca do usuário contenha o conceito “Amazonas”. Pelo conhecimento existente na ontologia *lightweight* referente à divisão territorial do Brasil (Apêndice A) verifica-se que o conceito “Brasil”, pelo uso da associação de generalização, está associado ao conceito “Amazonas”. Pelos experimentos apresentados neste capítulo conclui-se que os melhores resultados ocorrem quando os conceitos mais gerais são menos privilegiados e, desta forma, os valores dos pesos relacionados à associação de generalização são baixos. Ao utilizar valores baixos para os pesos de associação de generalização o conceito “Brasil” é associado ao conceito “Amazonas” com um valor de associação mais próximo de 0. Por um outro lado supor que um documento D_1 possua o seu conteúdo contendo apenas o termo “Brasil”. Como o tamanho do texto onde o termo ocorre influencia o valor final de indexação do termo ao documento então, neste caso, o valor de associação do termo “Brasil” ao documento D_1 vai ser alto. Ao se executar a busca do usuário na base, apesar do conceito “Brasil” estar associado ao conceito inicial “Amazonas” com um valor baixo, ao mesmo tempo o termo “Brasil” vai estar associado ao documento D_1 com um valor alto. Neste processo o valor final da similaridade do documento D_1 para a consulta inicial “Amazonas” terá um valor competitivo. Neste caso o documento D_1 será trazido no topo da lista de documentos recuperados mesmo que ele não seja um documento relevante para o conceito “Amazonas”. Desta maneira a forma como os documentos são indexados no Apache Lucene pode trazer documentos considerados não relevantes no topo da lista de documentos recuperados para uma dada consulta inicial.

5.4 Construção dos Casos de Teste

Nos modelos testados utiliza-se uma base de conhecimento composta por conceitos e relacionamentos entre os conceitos expressos por relações de especialização *fuzzy*, generalização *fuzzy* e associação positiva *fuzzy*. No caso do modelo *fuzzy* de múltiplas ontologias e do Apache Lucene o conhecimento existente na base é utilizado para realizar a expansão da consulta adicionando novos conceitos à mesma. No caso do modelo de rede de conceitos *fuzzy* o conhecimento é utilizado para a indexar os documentos a novos conceitos. Em todos os modelos é possível controlar a influência que cada tipo de relacionamento terá na recuperação final de documentos atribuindo-se pesos para cada um deles. Para o modelo *fuzzy* de múltiplas ontologias e o Apache Lucene o peso atribuído ao relacionamento vai influenciar o valor associado a cada conceito quando da sua adição na expansão da consulta. Para o modelo de rede de conceitos *fuzzy* o peso atribuído ao relacionamento vai influenciar o valor de relevância final do documento. Em todos os casos os pesos atribuídos aos tipos de relacionamento vão influenciar a recuperação dos documentos e a relevância atribuída a cada um deles para uma determinada consulta.

Os casos de testes foram construídos de forma a testar como as variações nos valores dos pesos atribuídos aos relacionamentos influenciam no resultado final da recuperação de informação. Desta forma procura-se estabelecer combinações que privilegiam mais um tipo de relacionamento para verificar o quanto este tipo de relacionamento influencia o desempenho final do processo de recuperação. O objetivo é verificar se a base de conhecimento formada por múltiplas ontologias relacionadas por associação positiva *fuzzy* representam um ganho no processo de recuperação de informação. Este ganho é medido pela melhoria nas medidas de precisão e cobertura. Os casos de teste construídos com as variações de valores para os tipos de relacionamentos são apresentados nesta seção.

5.4.1 Construção dos Casos de Teste para os Modelos Baseados em Múltiplas Ontologias Relacionadas

Para o modelo *fuzzy* de múltiplas ontologias e para o Apache Lucene considerou-se combinações dos pesos we_t, $t \in \{S,\ G\}$ e w_r, $r \in \{S,\ G,\ P\}$ na elaboração dos casos de teste. O peso we_t está relacionado ao cálculo do fecho transitivo ponderado entre os conceitos das relações de associação de especialização (S) e generalização (G) *fuzzy* apresentado pela Def. 4.2. Este peso penaliza a força da associação entre conceitos distantes nas ontologias. Isto significa que conceitos mais próximos nas taxonomias, que representam as ontologias, possuem um valor de associação mais alto. O peso w_r controla a influência de cada tipo de associação *fuzzy* no cálculo do valor final dos conceitos expandidos em função do conhecimento existente na base.

Nestes dois modelos, além das variações de pesos, também considerou-se dois tipos de ontologias:

fuzzy e *crisp*. O objetivo é verificar se existe alguma variação no desempenho final da recuperação de informação no caso de existirem valores que indicam o grau de associação entre os conceitos (ontologia *fuzzy*) ou se apenas a indicação da existência da associação entre os conceitos (ontologia *crisp*) já garante bons resultados. A construção das ontologias *fuzzy* é mostrada no Apêndice A. Para construir as ontologias *crisp* os valores da associação de generalização *fuzzy* e de especialização *fuzzy* são considerados no conjunto $\{0,\ 1\}$ denotando apenas a existência (1) ou ausência (0) da associação entre os conceitos sem caracterizar a força desta relação.

Na literatura [51] é discutido que o uso indiscriminado de estruturas conceituais como tesauros e ontologias para a expansão de consulta pode deteriorar a qualidade do conjunto resposta. Isto acontece pois a adição de novos termos à consulta original pode acarretar a recuperação de um número maior de documentos melhorando a cobertura e deteriorando a precisão. Neste processo a adição de conceitos mais gerais, em geral, causa um ruído maior no resultado do que a adição de conceitos mais específicos. Em função desta observação os valores dos pesos foram atribuídos de forma a construir situações de teste com casos que privilegiam a expansão considerando os conceitos mais específicos, casos que privilegiam conceitos mais gerais e casos onde ambos os conceitos são considerados com o mesmo peso possuindo o mesmo grau de importância.

Para os pesos we_t, $t \in \{S,\ G\}$ foram utilizados os valores do conjunto $\{0.2,\ 0.8,\ 1.0\}$, atribuídos de forma empírica, de forma a considerar três situações conforme descrito a seguir:

Situação 1: Os conceitos mais específicos são mais privilegiados que os conceitos mais gerais nas taxonomias: $we_G = 0.2$ e $we_S = 0.8$.

Situação 2: Os conceitos mais gerais são mais privilegiados que os conceitos mais específicos nas taxonomias: $we_G = 0.8$ e $we_S = 0.2$.

Situação 3: Não existe diferenciação entre os conceitos mais gerais e mais específicos nas taxonomias $we_G = 1.0$ e $we_S = 1.0$.

Para cada uma das três situações foi realizada uma série de combinações de valores para os pesos w_r, $r \in \{S,\ G,\ P\}$ referentes à influência dos tipos de associações *fuzzy* no processo de expansão. Para os pesos w_r, $r \in \{S,\ G\}$ foram utilizados os valores do conjunto $\{0.0,\ 0.3,\ 0.5,\ 0.7,\ 1.0\}$, atribuídos de forma empírica. O valor 0.0 indica que o tipo de relacionamento não terá influência no processo de expansão, o valor 0.5 indica que o tipo de relacionamento terá influência média e o valor 1.0 indica que o tipo de relacionamento terá influência completa. Os valores 0.3 e 0.7 são combinados de forma a privilegiar um tipo de relacionamento em detrimento do outro, ou seja, hora o relacionamento de especialização é mais privilegiado ($w_S = 0.7$) que o de generalização ($w_G = 0.3$) e hora o relacionamento de generalização é mais privilegiado ($w_G = 0.7$) que o de especialização ($w_S = 0.3$).

Para o peso w_P foram utilizados os valores do conjunto $\{0.0,\ 0.1,\ 0.5,\ 1.0\}$. O valor 0.0 indica que o tipo de relacionamento não terá influência no processo de expansão, o valor 0.1 indica que o tipo de relacionamento terá influência mínima, o valor 0.5 indica que o tipo de relacionamento terá influência média e o valor 1.0 indica que o tipo de relacionamento terá influência completa. O valor 0.1 foi considerado para o relacionamento de associação positiva pois observou-se, nos primeiros experimentos realizados, que os melhores valores para as medidas de precisão e cobertura para o modelo proposto eram obtidos quando o valor do peso $w_P = 0.1$. Os experimentos iniciais consideraram os valores no conjunto $\{0.0,\ 0.1,\ 0.3,\ 0.5,\ 0.7,\ 0.9,\ 1.0\}$ para o peso w_P.

A atribuição de valores aos pesos w_r, $r \in \{S,\ G,\ P\}$ gera os casos a seguir:

Caso 1: Apenas a associação positiva é considerada possuindo um valor nulo (0.0), mínimo (0.1), médio (0.5) e máximo (1.0). Para este caso tem-se os valores $w_G = 0.0$, $w_S = 0.0$ e $w_P \in \{0.0,\ 0.1,\ 0.5,\ 1.0\}$. Quando $w_P = 0.0$ não há a expansão da consulta e a consulta final corresponde à consulta inicial do usuário, ou seja, a consulta é feita apenas com as palavras-chaves.

Caso 2: Os conceitos gerados pela associação de especialização são mais privilegiados que os gerados pela associação de generalização com os valores: $w_G = 0.3$, $w_S = 0.7$. Para esta configuração considera-se os valores de associação positiva nulo, mínimo, médio e máximo para o peso w_P, ou seja, $w_P \in \{0.0,\ 0.1,\ 0.5,\ 1.0\}$. Quando $w_P = 0.0$ apenas os conceitos mais gerais e mais específicos dos conceitos presentes na consulta do usuário são gerados.

Caso 3: Os conceitos gerados pela associação de generalização são mais privilegiados que os gerados pela associação de especialização com os valores: $w_G = 0.7$, $w_S = 0.3$. Para esta configuração considera-se os valores de associação positiva nulo, mínimo, médio e máximo para o peso w_P, ou seja, $w_P \in \{0.0,\ 0.1,\ 0.5,\ 1.0\}$. Quando $w_P = 0.0$ apenas os conceitos mais gerais e mais específicos dos conceitos presentes na consulta do usuário são gerados.

Caso 4: Os conceitos gerados pela associação de generalização e os gerados pela associação de especialização são privilegiados igualmente com um valor médio (0.5), ou seja, $w_G = 0.5$, $w_S = 0.5$. Para esta configuração considera-se os valores de associação positiva nulo, mínimo, médio e máximo para o peso w_P, ou seja, $w_P \in \{0.0,\ 0.1,\ 0.5,\ 1.0\}$. Quando $w_P = 0.0$ apenas os conceitos mais gerais e mais específicos dos conceitos presentes na consulta do usuário são gerados.

Caso 5: Os conceitos gerados pela associação de generalização e os gerados pela associação de especialização são privilegiados igualmente com o valor máximo (1.0), ou seja, $w_G = 1.0$, $w_S = 1.0$. Para esta configuração considera-se os valores de associação positiva nulo, mínimo,

médio e máximo para o peso w_P, ou seja, $w_P \in \{0.0,\ 0.1,\ 0.5,\ 1.0\}$. Quando $w_P = 0.0$ apenas os conceitos mais gerais e mais específicos dos conceitos presentes na consulta do usuário são gerados.

A combinação entre os três tipos de situação e os cinco casos considerados, para cada situação, define um escopo de testes que abrange os experimentos executados para detectar o comportamento e o desempenho dos modelos em função dos tipos de relacionamentos existentes entre eles. O conjunto de combinações de testes final é mostrado na Tab. 5.1.

Combinações	**Situação 1** $we_G = 0.2$ $we_S = 0.8$	**Situação 2** $we_G = 0.8$ $we_S = 0.2$	**Situação 3** $we_G = 1.0$ $we_S = 1.0$
Caso 1	$w_G = 0.0$ $w_S = 0.0$ $w_P = 0.0, 0.1, 0.5, 1.0$	$w_G = 0.0$ $w_S = 0.0$ $w_P = 0.0, 0.1, 0.5, 1.0$	$w_G = 0.0$ $w_S = 0.0$ $w_P = 0.0, 0.1, 0.5, 1.0$
Caso 2	$w_G = 0.3$ $w_S = 0.7$ $w_P = 0.0, 0.1, 0.5, 1.0$	$w_G = 0.3$ $w_S = 0.7$ $w_P = 0.0, 0.1, 0.5, 1.0$	$w_G = 0.3$ $w_S = 0.7$ $w_P = 0.0, 0.1, 0.5, 1.0$
Caso 3	$w_G = 0.7$ $w_S = 0.3$ $w_P = 0.0, 0.1, 0.5, 1.0$	$w_G = 0.7$ $w_S = 0.3$ $w_P = 0.0, 0.1, 0.5, 1.0$	$w_G = 0.7$ $w_S = 0.3$ $w_P = 0.0, 0.1, 0.5, 1.0$
Caso 4	$w_G = 0.5$ $w_S = 0.5$ $w_P = 0.0, 0.1, 0.5, 1.0$	$w_G = 0.5$ $w_S = 0.5$ $w_P = 0.0, 0.1, 0.5, 1.0$	$w_G = 0.5$ $w_S = 0.5$ $w_P = 0.0, 0.1, 0.5, 1.0$
Caso 5	$w_G = 1.0$ $w_S = 1.0$ $w_P = 0.0, 0.1, 0.5, 1.0$	$w_G = 1.0$ $w_S = 1.0$ $w_P = 0.0, 0.1, 0.5, 1.0$	$w_G = 1.0$ $w_S = 1.0$ $w_P = 0.0, 0.1, 0.5, 1.0$

Tab. 5.1: Conjunto de combinações de pesos para os experimentos nos modelos baseados em múltiplas ontologias relacionadas.

5.4.2 Construção dos Casos de Teste para o Modelo de Rede de Conceitos *Fuzzy*

No modelo de rede de conceitos *fuzzy* a base de conhecimento é composta de uma rede de conceitos construída automaticamente a partir da coleção de documentos. Desta forma o modelo não utiliza os pesos we_t, $t \in \{S,\ G\}$ referente ao cálculo do fecho transitivo ponderado nas ontologias. O modelo considera apenas os pesos w_r, $r \in \{S,\ G,\ P\}$ que controla a influência de cada tipo de associação *fuzzy* no cálculo do valor final dos conceitos expandidos. Os casos de teste deste modelo

consideram os mesmos valores para os pesos w_r, discutidos na seção 5.4.1, para possibilitar a comparação de seus resultados com os modelos que utilizam as ontologias relacionadas. O conjunto de combinações de testes final é mostrado na Tab. 5.2.

Caso 1	**Caso 2**	**Caso 3**	**Caso 4**	**Caso 5**
$w_G = 0.0$ $w_S = 0.0$ $w_P = 0.0, 0.1,$ $0.5, 1.0$	$w_G = 0.3$ $w_S = 0.7$ $w_P = 0.0, 0.1,$ $0.5, 1.0$	$w_G = 0.7$ $w_S = 0.3$ $w_P = 0.0, 0.1,$ $0.5, 1.0$	$w_G = 0.5$ $w_S = 0.5$ $w_P = 0.0, 0.1,$ $0.5, 1.0$	$w_G = 1.0$ $w_S = 1.0$ $w_P = 0.0, 0.1,$ $0.5, 1.0$

Tab. 5.2: Conjunto de combinações de pesos a serem utilizados para os experimentos no modelo de rede de conceitos *fuzzy*.

5.5 Apresentação de Resultados

Todas as combinações de teste, apresentadas na seção 5.4, foram executadas para os modelos coletando-se as medidas de cobertura e precisão para cada uma das combinações. As medidas de cobertura e precisão foram agregadas e foram gerados três tipos de gráficos. O primeiro gráfico ilustra a média das medidas de precisão agregadas por cada um dos parâmetros. O segundo é um gráfico de precisão *versus* cobertura, conforme descrito na seção 2.4.1, construído para cada um dos modelos. O terceiro gráfico mostra as curvas da diferença pelos maiores valores obtidas em cada um dos modelos para possibilitar uma comparação entre eles. Na análise dos resultados é dada uma atenção maior para a influência que a associação positiva possui na obtenção de melhores resultados. Isto ocorre pois este tipo de associação é utilizada para relacionar as ontologias na base de conhecimento do modelo proposto e a sua influência na melhoria dos resultados obtidos constitui um fator que se deseja verificar nos experimentos realizados.

Nos testes realizados o valor adotado para o limite t é 0.001. O limite t estabelece um valor mínimo para o valor de relevância do documento para que ele seja retornado no conjunto resposta de uma consulta. As medidas de precisão e cobertura não dependem do tamanho total do conjunto de documentos retornados mas sim do posicionamento destes documentos na lista ordenada do conjunto resposta. O valor do limite $t = 0.001$ possibilitou o retorno de todos os documentos relevantes na resposta e a verificação da precisão obtida para o valor de cobertura igual a 100%.

Além dos três gráficos gerados também é utilizada a ferramenta Treemap [59], de visualização de estruturas hierárquicas de dados, onde todos os resultados gerados pelos testes são visualizados de forma organizada permitindo a exploração de padrões e excessões nos resultados. A visualização pela ferramenta Treemap permite confirmar as conclusões obtidas nos experimentos.

5.5.1 Gráfico da Média das Medidas de Precisão

Os gráficos das médias dos valores de precisão são mostrados para cada valor dos parâmetros. O objetivo deste gráfico é ilustrar para qual valor de parâmetro cada um dos modelos obteve a melhor precisão. Cada gráfico mostra os valores obtidos pelos modelos segundo a legenda:

- MO Fuzzy: modelo *fuzzy* de múltiplas ontologias considerando as ontologias *fuzzy*.
- Luc Fuzzy: Apache Lucene considerando as ontologias *fuzzy*.
- MO Crisp: modelo *fuzzy* de múltiplas ontologias considerando as ontologias *crisp*.
- Luc Crisp: Apache Lucene considerando as ontologias *crisp*.
- Rede Fuzzy: modelo de rede de conceitos *fuzzy*.

A Fig. 5.1 ilustra o gráfico de médias dos valores de precisão considerando as combinações dos pesos we_G e we_S. Os valores para este gráfico foram obtidos fixando-se os valores dos pesos $we_G\_we_S \in \{0.2\_0.8,\ 0.8\_0.2,\ 1.0\_1.0\}$ e calculando a média para todos os valores de precisão obtidos pela variação dos parâmetros w_G, w_S e w_P, considerando todos os valores de cobertura.

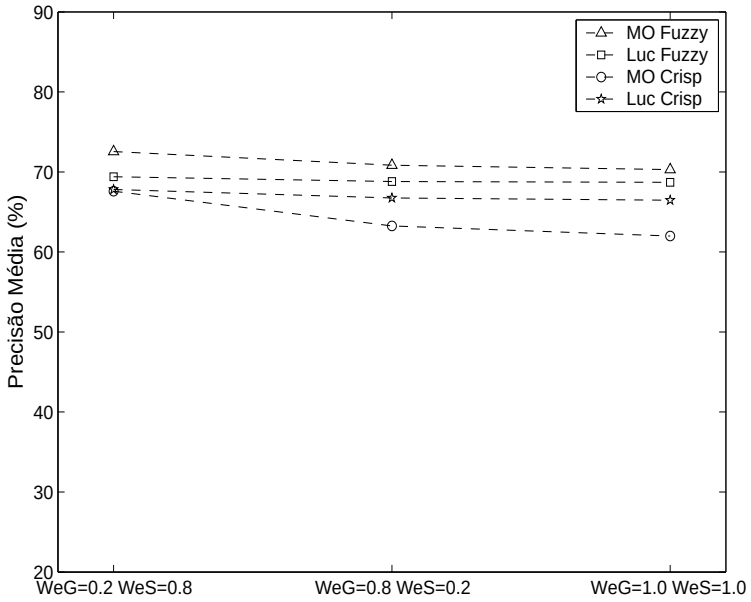

Fig. 5.1: Precisão média considerando os pesos we_G e we_S.

A Eq. 5.12 mostra o cálculo da média da precisão para a Fig. 5.1.

$$MédiaPrecisão_{we_G\_we_S} = \frac{\sum Precisão_{w_G\_w_S,w_P,cobertura}}{|w_G\_w_S|\ |w_P|\ |cobertura|} \tag{5.12}$$

Nesta equação os parâmetros assumem os seguintes valores:

- $w_G\_w_S \in \{0.0\_0.0,\ 0.3\_0.7,\ 0.7\_0.3,\ 0.5\_0.5,\ 1.0\_1.0\}$
- $w_P \in \{0.1,\ 0.5,\ 1.0\}$. Os valores para $w_P = 0.0$ não foram considerados neste cálculo pois deseja-se medir o quanto a associação positiva *fuzzy* influencia na média final.
- $cobertura \in \{0,\ 10,\ 20,\ 30,\ 40,\ 50,\ 60,\ 70,\ 80,\ 90,\ 100\}$

Como os pesos we_G e we_S aplicam-se apenas à base de conhecimento formada pelas ontologias então não existem os valores correspondentes para o modelo de rede de conceitos *fuzzy*. Observando o gráfico nota-se que todos os modelos obtiveram as melhores médias de precisão correspondentes aos valores de $we_G = 0.2$ e $we_S = 0.8$ indicando que quando se privilegia os conceitos mais específicos nas taxonomias, que definem as ontologias, melhores resultados de precisão são obtidos.

A Fig. 5.2 ilustra o gráfico de médias dos valores de precisão considerando as combinações dos pesos w_G e w_S. Os valores para este gráfico foram obtidos fixando-se os valores dos pesos $w_G\_w_S \in \{0.0\_0.0,\ 0.3\_0.7,\ 0.5\_0.5,\ 0.7\_0.3,\ 1.0\_1.0\}$ e calculando a média para todos os valores de precisão obtidos pela variação dos parâmetros we_G, we_S e w_P considerando todos os valores de cobertura.

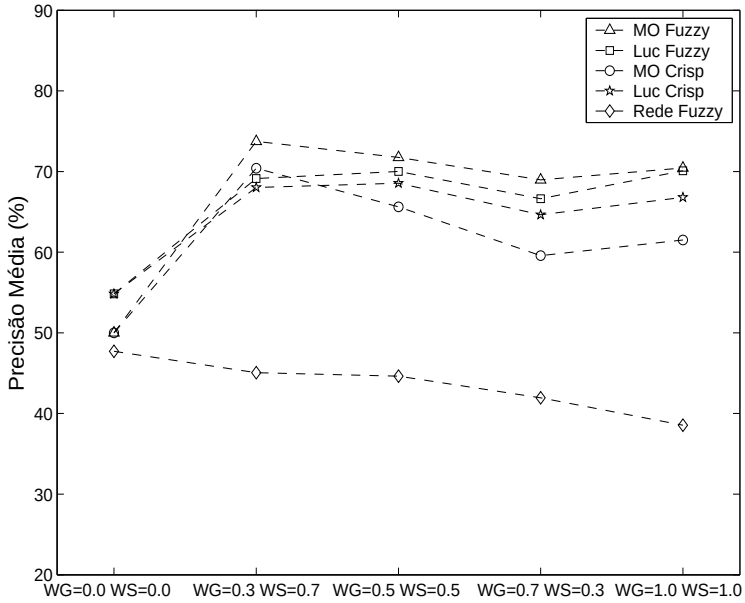

Fig. 5.2: Precisão média considerando os pesos w_G e w_S.

A Eq. 5.13 mostra o cálculo da média da precisão para a Fig. 5.2.

$$MédiaPrecisão_{w_G\_w_S} = \frac{\sum Precisão_{we_G\_we_S,w_P,cobertura}}{|we_G\_we_S|\ |w_P|\ |cobertura|} \tag{5.13}$$

Nesta equação os parâmetros assumem os seguintes valores:

- $we_G\_we_S \in \{0.2\_0.8,\ 0.8\_0.2,\ 1.0\_1.0\}$
- $w_P \in \{0.1,\ 0.5,\ 1.0\}$. Os valores para $w_P = 0.0$ não foram considerados neste cálculo pois deseja-se medir o quanto a associação positiva *fuzzy* influencia na média final.
- $cobertura \in \{0,\ 10,\ 20,\ 30,\ 40,\ 50,\ 60,\ 70,\ 80,\ 90,\ 100\}$

Considerando os resultados no gráfico de médias dos valores de precisão para as combinações dos pesos w_G e w_S os modelos apresentaram resultados diferenciados. O modelo *fuzzy* de múltiplas ontologias, considerando tanto as ontologias *crisp* quanto *fuzzy*, obteve os melhores resultados de precisão média para os valores de $w_G = 0.3$ e $w_S = 0.7$. Ao privilegiar os conceitos mais específicos na expansão melhores resultados de precisão são obtidos.

O Apache Lucene, utilizando ontologias *crisp*, obteve melhores resultados de precisão média para os valores de $w_G = 0.5$ e $w_S = 0.5$. Ao utilizar ontologias *fuzzy* os melhores valores foram obtidos para os valores de $w_G = 1.0$ e $w_S = 1.0$.

O modelo de rede de conceitos *fuzzy* obteve os melhores resultados para os valores de $w_G = 0.0$ e $w_S = 0.0$. O modelo de rede de conceitos *fuzzy* calcula a relação de associação de especialização *fuzzy* e de generalização *fuzzy* baseada na co-ocorrência de palavras. No caso deste modelo os melhores valores de precisão são obtidos quando os conceitos mais específicos e mais gerais não são considerados. O modelo de rede de conceitos *fuzzy* obtém os melhores resultados quando apenas os conceitos relacionados pela associação positiva são considerados. Este fato poderá ser observado na descrição do gráfico de precisão média considerando o peso w_P.

Pelos resultados mostrados no gráfico da Fig. 5.2 observa-se que, nos modelos que utilizam as ontologias como base de conhecimento, o valor mais baixo de precisão foi obtido para os valores de $w_G = 0.0$ e $w_S = 0.0$, ao contrário do que ocorreu no modelo de rede de conceitos *fuzzy*. Quando as taxonomias possuem um significado semântico bem estabelecido o fato de considerar conceitos mais gerais e mais específicos na expansão da consulta melhora a precisão dos resultados.

A Fig. 5.3 ilustra o gráfico de médias dos valores de precisão considerando as combinações do peso w_P. Para este gráfico fixou-se os valores do peso $w_P \in \{0.0,\ 0.1,\ 0.5,\ 1.0\}$ e calculou-se a média para todos os valores de precisão obtidos pela variação dos parâmetros we_G e we_S, w_G e w_S considerando todos os valores de cobertura. Notar que, neste gráfico, o valor de $w_P = 0.0$ é considerado para comparar o quanto a associação positiva contribui para a média final do valor de precisão.

A Eq. 5.14 mostra o cálculo da média da precisão para a Fig. 5.3.

$$MédiaPrecisão_{w_P} = \frac{\sum Precisão_{we_G\_we_S,\ w_G\_w_S,\ cobertura}}{|we_G\_we_S|\ |w_G\_w_S|\ |cobertura|} \tag{5.14}$$

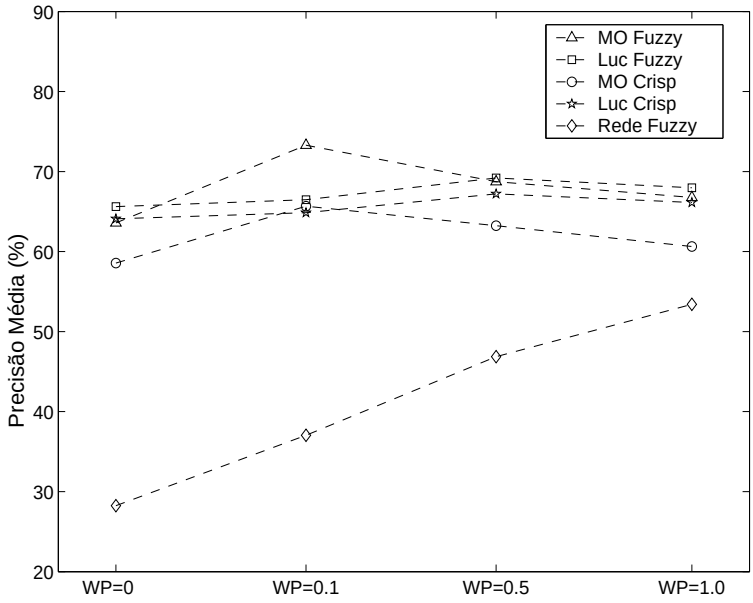

Fig. 5.3: Precisão média considerando os pesos w_P.

Nesta equação os parâmetros assumem os seguintes valores:

- $we_G\_we_S \in \{0.2\_0.8,\ 0.8\_0.2,\ 1.0\_1.0\}$
- $w_G\_w_S \in \{0.0\_0.0,\ 0.3\_0.7,\ 0.7\_0.3,\ 0.5\_0.5,\ 1.0\_1.0\}$.
- $cobertura \in \{0,\ 10,\ 20,\ 30,\ 40,\ 50,\ 60,\ 70,\ 80,\ 90,\ 100\}$

Considerando o gráfico das médias dos valores de precisão para as combinações do peso w_P os modelos apresentaram resultados diferenciados. O modelo *fuzzy* de múltiplas ontologias, considerando tanto as ontologias *crisp* quanto as ontologias *fuzzy*, obteve os melhores resultados de precisão média para os valores de $w_P = 0.1$. Quando o valor do peso da associação positiva *fuzzy* $w_P > 0.0$ o modelo expande a consulta com conceitos de outros domínios com um peso maior que zero. Esta expansão indica que existe uma co-ocorrência semântica entre os conceitos dos domínios distintos.

O Apache Lucene, considerando tanto as ontologias *crisp* quanto as ontologias *fuzzy*, obteve os melhores resultados de precisão média para os valores de $w_P = 0.5$.

O modelo de rede de conceitos *fuzzy* obteve os melhores resultados para os valores de $w_P = 1.0$. O modelo de rede de conceitos *fuzzy* calcula a relação de associação positiva *fuzzy* baseada na co-ocorrência de palavras. A rede de relacionamentos positivos, construída pelo modelo, é muito rica e a exploração deste tipo de relacionamento garante a melhora da média da precisão.

5.5.2 Gráfico da Precisão *versus* Cobertura

Nesta seção são apresentados os gráficos de precisão *versus* cobertura para cada um dos modelos. Este gráfico ilustra o desempenho médio de cada modelo considerando todas as combinações de teste executadas no modelo. Este gráfico mostra várias informações:

Curva 1: Para cada valor de cobertura é mostrada a curva que indica o valor médio de precisão obtida para as várias combinações onde $w_P > 0.0$. Além disto também é mostrado o maior e menor valor de precisão obtido. O objetivo desta curva é mostrar o desempenho do sistema ao se utilizar a associação positiva relacionando o conhecimento de domínios distintos. A Eq. 5.15 mostra o cálculo da média da precisão:

$$MédiaPrecisão_{cobertura} = \frac{\sum Precisão_{we_G\_we_S,\, w_G\_w_S,\, w_P}}{|we_G\_we_S|\ |w_G\_w_S|\ |w_P|} \tag{5.15}$$

Nesta equação os parâmetros assumem os seguintes valores:

- $we_G\_we_S \in \{0.2\_0.8,\ 0.8\_0.2,\ 1.0\_1.0\}$
- $w_G\_w_S \in \{0.0\_0.0,\ 0.3\_0.7,\ 0.7\_0.3,\ 0.5\_0.5,\ 1.0\_1.0\}$.
- $w_P \in \{0.1,\ 0.5,\ 1.0\}$.
- $cobertura \in \{0,\ 10,\ 20,\ 30,\ 40,\ 50,\ 60,\ 70,\ 80,\ 90,\ 100\}$

Curva 2: Indica a diferença pelos maiores valores, ou seja, nas várias combinações de parâmetros qual foi a combinação que gerou a curva que mais se aproxima dos maiores valores de precisão mostrados no item 1.

Curva 3: Esta curva é erada utilizando a informação do primeiro tipo de gráfico, discutido na seção 5.5.1, considerando os valores dos pesos que geraram o maior valor de precisão médio.

Curva 4: Mostra o desempenho do modelo quando apenas as associações de generalização e especialização *fuzzy*, dentro das estruturas conceituais, são consideradas, ou seja, $w_P = 0.0$. O objetivo é estabelecer uma comparação com as curvas 1, 2 e 3 e verificar o quanto a associação positiva *fuzzy* influencia o desempenho final do modelo. A Eq. 5.16 mostra o cálculo da média da precisão:

$$MédiaPrecisão_{cobertura} = \frac{\sum Precisão_{we_G\_we_S,\, w_G\_w_S,\, w_P}}{|we_G\_we_S|\ |w_G\_w_S|\ |w_P|} \tag{5.16}$$

Nesta equação os parâmetros assumem os seguintes valores:

- $we_G\_we_S \in \{0.2\_0.8,\ 0.8\_0.2,\ 1.0\_1.0\}$

- $w_G\_w_S \in \{$ 0.3_0.7, 0.7_0.3, 0.5_0.5, 1.0_1.0$\}$.
- $w_P \in \{0.0\}$.
- $cobertura \in \{0,\ 10,\ 20,\ 30,\ 40,\ 50,\ 60,\ 70,\ 80,\ 90,\ 100\}$

Curva 5: Indica o desempenho do modelo utilizando somente as palavras-chave da consulta inicial. A busca por palavras-chave ocorre quando os pesos assumem os valores $w_G = 0.0$, $w_S = 0.0$ e $w_P = 0.0$ mostrando que as relações nas ontologias não são consideradas e a consulta não é expandida.

Modelo *Fuzzy* de Múltiplas Ontologias

A Fig. 5.4 ilustra o gráfico de precisão *versus* cobertura para o modelo *fuzzy* de múltiplas ontologias considerando as ontologias *fuzzy*. Pode-se observar que a curva que indica diferença pelos maiores valores e a curva que indica a média dos parâmetros coincidem com o maior valor de precisão para cada valor de cobertura. Os valores de parâmetros que geram o melhor resultado são indicados na própria figura. Estes valores indicam que ao se privilegiar os conceitos mais específicos do que os conceitos mais gerais, tanto nas taxonomias quanto no processo de expansão, melhores valores de precisão são obtidos. Também pode-se observar que qualquer combinação de valores de parâmetros gera um resultado de precisão melhor do que o gerado utilizando apenas as palavras-chave. A curva onde a associação positiva *fuzzy* não é considerada, $w_P = 0.0$, apresenta um valor de precisão alto para valores de cobertura baixos mas a precisão decresce rapidamente à medida que o valor de cobertura aumenta. Isto indica que a associação positiva é importante na manutenção do valor de precisão alto à medida que o valor de cobertura aumenta.

A Fig. 5.5 ilustra o gráfico de precisão *versus* cobertura para o modelo *fuzzy* de múltiplas ontologias considerando as ontologias *crisp*. Neste gráfico pode-se notar que a curva que indica diferença pelos maiores valores e a curva que indica a média dos parâmetros coincidem e a maioria dos valores de precisão é dado pelo maior valor para cada valor de cobertura. A curva gerada utilizando somente as palavras-chaves apresenta os mesmos valores de precisão que os indicados na Fig. 5.4. Uma vez que o conhecimento nas ontologias não foi utilizado na expansão então a curva é a mesma para os gráficos que representam ambos os tipos de ontologia: *fuzzy* e *crisp*. A curva onde a associação positiva *fuzzy* não é considerada, $w_P = 0.0$, apresenta valores de precisão mais próximos da média para valores de cobertura baixos e a precisão decresce para valores baixos à medida que o valor de cobertura aumenta.

Comparando os resultados das Fig.s 5.4 e 5.5 pode-se notar que os valores de precisão obtidos com a ontologia *fuzzy* são um pouco melhores que os valores obtidos com a ontologia *crisp*. Tomando por base que a curva gerada pelas palavras-chaves é a mesma para ambos os modelos observa-se que

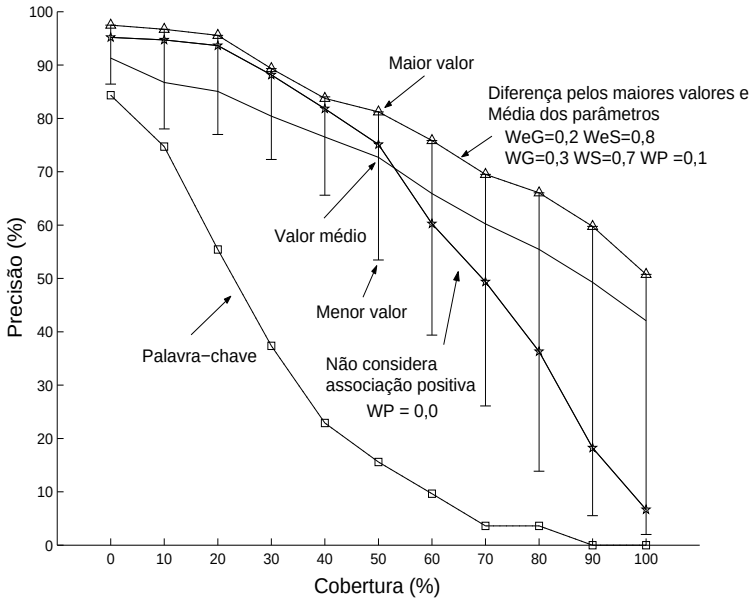

Fig. 5.4: Gráfico de precisão *versus* cobertura para o modelo *fuzzy* de múltiplas ontologias considerando as ontologias *fuzzy*.

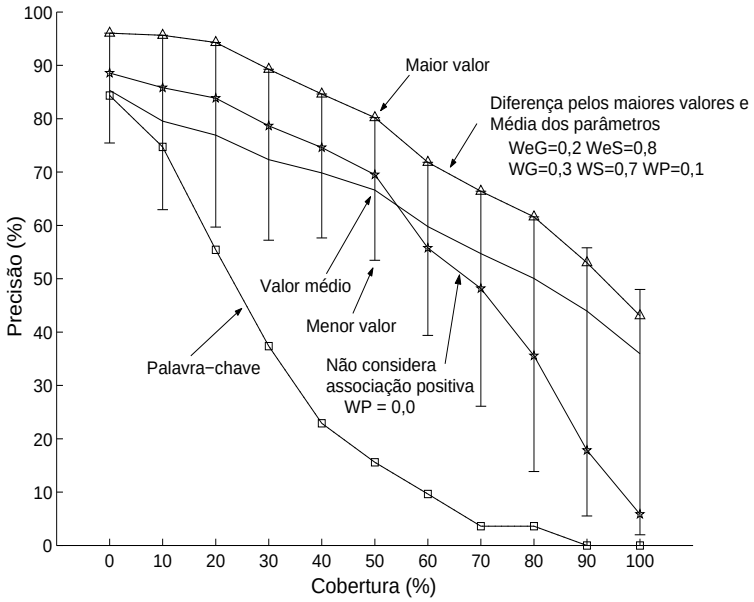

Fig. 5.5: Gráfico de precisão *versus* cobertura para o modelo *fuzzy* de múltiplas ontologias considerando as ontologias *crisp*.

no gráfico da Fig. 5.4, que considera as ontologias *fuzzy*, a curva que representa a média dos valores de precisão fica com os valores acima dos obtidos com as palavras-chave para todos os valores de cobertura. No gráfico da Fig. 5.5, que considera as ontologias *crisp*, os valores de precisão média para valores de cobertura mais baixos ficam mais próximos dos valores obtidos com as palavras-chave. Este resultado indica que o uso de pesos para definir o grau de força das relações de associação de especialização e generalização nas ontologias *fuzzy* resultam em uma melhoria do resultado.

Modelo de Rede de Conceitos *Fuzzy*

A Fig. 5.6 ilustra o gráfico de precisão *versus* cobertura para o modelo rede de conceitos *Fuzzy*. Neste gráfico deve-se observar que a curva que indica a diferença pelos maiores valores coincide com os maiores valores de precisão obtidos para cada valor de cobertura. Estes valores são obtidos quando o valor de associação positiva *fuzzy* $w_P = 1.0$ e os valores de associação de especialização e generalização *fuzzy* $w_G = 0.0$ e $w_S = 0.0$ respectivamente. Isto indica que, para os testes realizados no domínio considerado, a associação positiva *fuzzy* colabora sozinha para obtenção do melhor resultado. Na literatura não foram encontrados outros experimentos realizados com o modelo de rede de conceitos *fuzzy* que pudessem ser comparados com os resultados obtidos no experimento realizado nesta tese. Uma curva contendo o melhor resultado para os parâmetros com valores diferentes de zero foi traçada. Apesar dos valores de precisão obtidos serem melhores que o resultado médio eles estão abaixo daqueles obtidos considerando apenas a associação positiva *fuzzy*. Quando as associações positivas *fuzzy* não são consideradas os valores da precisão são abaixo da média. Isto indica que, no domínio considerado, as associações de especialização e generalização *fuzzy* contribuem pouco para a obtenção dos melhores resultados. Na forma como o modelo é construído não é possível elaborar uma consulta onde os pesos são $w_G = 0.0$, $w_S = 0.0$ e $w_P = 0.0$ simulando apenas o uso de palavras-chave. Pela Eq. 5.11 pode-se ver que com estes valores de peso o grau de satisfação dos documentos é sempre zero.

O modelo de rede de conceitos *fuzzy* monta a base de conhecimento, de forma automática, a partir da co-ocorrência de conceitos e palavras nos textos dos documentos. A forma como a rede de conceitos é construída, baseada na medida de *subsethood* da Eq. 5.5, faz com que o modelo estabeleça um número maior de associações positivas, entre os conceitos, do que o número de associações de generalização e especialização. Desta forma, quando a associação positiva é considerada com um peso maior, acarreta na recuperação de documentos mais relevantes à consulta do que quando apenas as associações de especialização e generalização são consideradas.

Um outro ponto a ser levantado é que, como as relações entre os conceitos são estabelecidas em função da co-ocorrência de palavras nos documentos, pode ocorrer de nem sempre a relação estabelecida possuir a semântica correta. Por exemplo, para a coleção de documentos do experimento, o

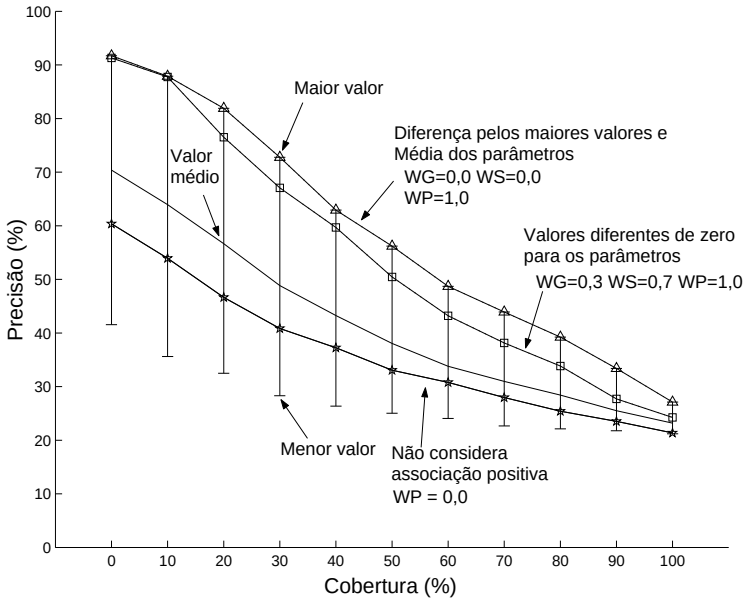

Fig. 5.6: Gráfico de precisão *versus* cobertura para o modelo rede de conceitos *fuzzy*.

modelo estabeleceu que o conceito “Santa Catarina” é mais específico que o conceito “Região Centro-Oeste” com um grau de 0.36. Este valor é encontrado pois os conceitos “Santa Catarina” e “Região Centro-Oeste” co-ocorrem em alguns documentos da coleção. Apesar da co-ocorrência ser captada pelo modelo, o significado da relação de especialização entre os conceitos não possui a semântica correta para este domínio de ontologia territorial. Devido a este tipo de equívoco, provocado pelo cálculo automático dos tipos de associação, atribui-se o fato do modelo não ser muito eficiente quando apenas as associações de especialização e generalização são consideradas.

Apache Lucene

A Fig. 5.7 ilustra o gráfico de precisão *versus* cobertura para o Apache Lucene considerando as ontologias *fuzzy*. Para valores menores de cobertura a busca apenas por palavras-chave apresenta um valor melhor de precisão embora este valor decresça rapidamente. A curva que indica diferença pelos maiores valores e a curva que indica a média dos parâmetros não coincidem mas seguem o mesmo padrão mantendo-se próximas dos maiores valores. A curva onde a associação positiva *fuzzy* não é considerada apresenta um valor de precisão acima dos maiores valores para valores de cobertura baixos. Isto indica que para valores de cobertura baixos o fato de se considerar apenas as associações de especialização e generalização, dentro das ontologias, induz uma precisão melhor do que ao se considerar a associação positiva *fuzzy*. Mas, da mesma forma como acontece no modelo *fuzzy* de múltiplas ontologias, à medida que o valor de cobertura aumenta o valor da precisão descresce rapidamente.

A Fig. 5.8 ilustra o gráfico de precisão *versus* cobertura para o Apache Lucene considerando as

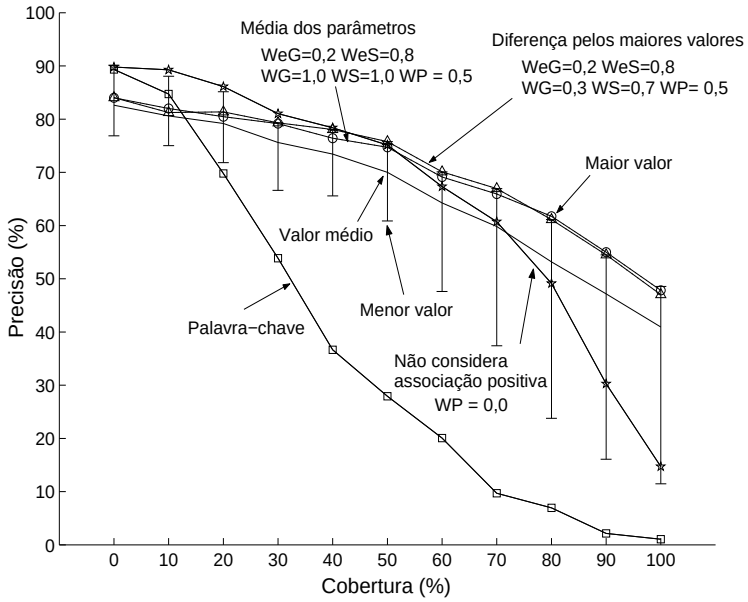

Fig. 5.7: Gráfico de precisão *versus* cobertura para o Apache Lucene considerando as ontologias *fuzzy*.

ontologias *crisp*. As curvas apresentam o mesmo padrão de comportamento que as curvas associadas às ontologias *fuzzy*. Para a curva que representa a diferença pelos maiores valores o valor dos parâmetros é igual aos utilizados com as ontologias *fuzzy*. Para a curva que representa a média dos parâmetros os valores de w_G e w_S são diferentes para os dois tipos de ontologia. A curva onde a associação positiva *fuzzy* não é considerada apresenta um valor de precisão alto para valores de cobertura baixos mas, da mesma forma que ao se utilizar as ontologias *fuzzy*, à medida que o valor de cobertura aumenta o valor da precisão descresce rapidamente.

5.5.3 Gráfico com os Maiores Valores

A Fig. 5.9 mostra as curvas que contém os maiores valores de precisão, correspondente às curvas 2 nos gráficos de precisão *versus* cobertura (seção 5.5.2), considerando cada um dos modelos.

Pela observação do gráfico na Fig. 5.9 pode-se notar que o modelo *fuzzy* de múltiplas ontologias, proposto nesta tese, apresentou melhor desempenho que os demais modelos. Considerando as ontologias *fuzzy* o valor de precisão para as taxas de cobertura mais baixas é acima de 95% e este valor mantém-se acima de 50% à medida que o valor da cobertura aumenta. Considerando as ontologias *crisp* o valor de precisão para as taxas de cobertura mais baixas é acima de 95% e este valor mantém-se acima de 43% à medida que o valor da cobertura aumenta.

O modelo Apache Lucene, em função de seu mecanismo de indexação, exibe valores de precisão mais baixos. Considerando as ontologias *fuzzy*, o valor de precisão para as taxas de cobertura mais

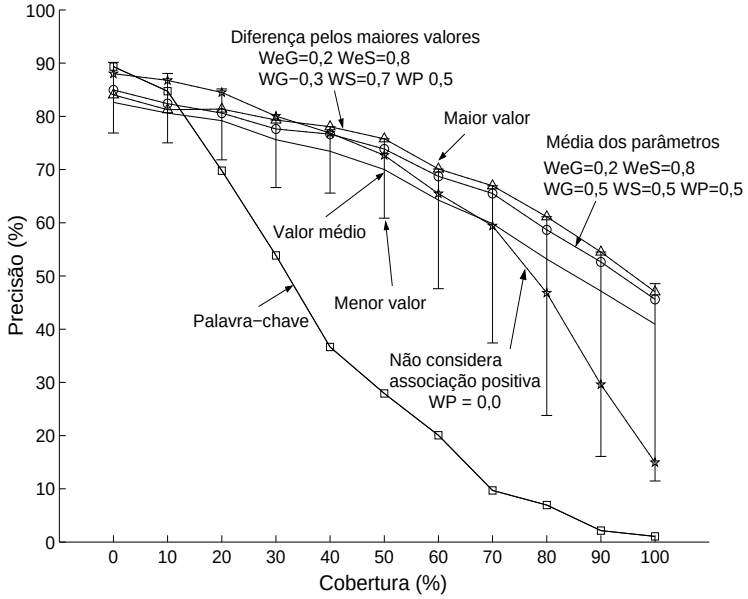

Fig. 5.8: Gráfico de precisão *versus* cobertura para o Apache Lucene considerando as ontologias *crisp*.

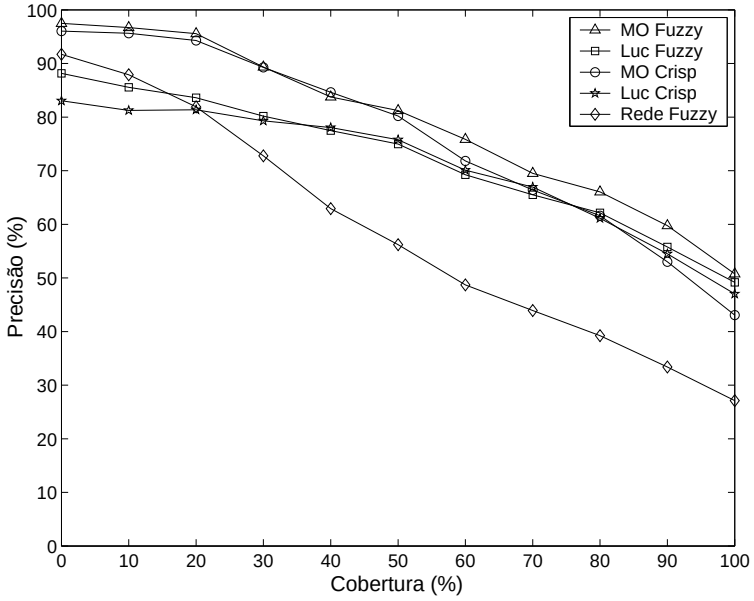

Fig. 5.9: Gráfico de precisão *versus* cobertura comparando o desempenho dos modelos.

baixas é acima de 85% e este valor mantém-se acima de 49% à medida que o valor da cobertura aumenta. Considerando as ontologias *crisp*, o valor de precisão para as taxas de cobertura mais baixas é acima de 80% e este valor mantém-se acima de 47% à medida que o valor da cobertura aumenta.

Ao utilizar as ontologias *fuzzy*, tanto o modelo *fuzzy* de múltiplas ontologias quanto o Apache Lucene obtiveram um desempenho melhor do que ao utilizar as ontologias *crisp*.

O modelo de rede de conceitos *fuzzy* exibe valores de precisão acima de 90% para valores baixos da cobertura e mantém este valor acima de 27% à medida que o valor da cobertura aumenta. O modelo rede de conceitos *fuzzy*, comparado com os demais, apresenta valores menores de precisão à medida que a cobertura aumenta. O fato da sua base de conhecimento ser gerada automaticamente significa que ela não possui a semântica existente em uma base gerada pelo conhecimento humano o que acarreta esta diferença nos resultados. Ao mesmo tempo pode-se concluir que o modelo apresenta um bom desempenho considerando-se que sua base é gerada automaticamente.

5.5.4 Visualização dos Resultados Utilizando *Treemap*

A visualização de informações visa auxiliar o processo de análise e compreensão de um conjunto de dados usando representações gráficas manipuláveis destes dados. As técnicas de visualização de informações procuram representar graficamente dados de um determinado domínio de aplicação de modo que a representação visual gerada explore a capacidade de percepção humana para, a partir das relações espaciais exibidas, interpretar e compreender as informações apresentadas [65].

Entre a família de ferramentas analíticas de visualização de dados os *treemaps* estão se sobressaindo em organizações que requerem monitoramento diário de atividades complexas que envolvem milhares de dados. Os *treemaps* são baseados no enfoque *space-filling* [62, 113] para mostrar hierarquias onde o espaço da tela é dividido em regiões e cada região é dividida novamente ilustrando os vários níveis de hierarquia aninhados. Os dados nestas regiões são diferenciados em função de seus atributos podendo ser utilizadas características visuais de cor e tamanho para ilustrar suas características.

A visualização de dados foi utilizada para permitir a visualização gráfica de todas as combinações de teste realizadas para os modelos de recuperação de informação. A informação mínima representada é o valor da precisão, para um valor de cobertura, considerando uma combinação específica para os valores dos parâmetros.

A Fig. 5.10 mostra uma visualização da tela da ferramenta *Treemap*. A ferramenta é dividida em três grandes áreas. À esquerda a ferramenta mostra o mapa hierárquico em árvore que é navegável e interativo. No canto superior direito a ferramenta ilustra informações detalhadas de um item selecionado no mapa. Neste caso o item selecionado corresponde àquele em que as bordas que limi-

tam a visualização dos quadrados estão na cor azul. No canto inferior direito a ferramenta apresenta seus controles de interação. Todos os controles do *Treemap* estão distribuídos em quatro abas: *Main*, *Legend*, *Filters* e *Hierarchy*.

- Na aba *Main* os usuários podem selecionar um dos três tipos de algoritmos para o desenho do mapa: quadrado, fatia e corta ou em tiras. Também podem ser definidas as opções de fonte e borda.
- A aba *Legend* (mostrada na figura) permite que os usuários atribuam os mapeamentos entre os atributos dos dados aos atributos visuais do mapa. É possível escolher o atributo que vai nomear os dados e a possibilidade de diferenciá-los por tamanho ou por cor. O usuário também pode escolher o número de níveis hierárquicos a serem mostrados no mapa, a forma como os valores dos dados serão agregados nos níveis superiores e o tipo de escala para ajudar a visualizar dados distorcidos. Nesta aba é possível fixar os intervalos (bins) de valores para o atributo escolhido para definir a visualização por cor deste atributo.
- Na aba *Filters* o usuário pode filtrar os dados usando *widgets* que permitem que ele selecione tanto o tipo de informação como intervalos de valores que ele deseja ver disponibilizados no mapa. Desta forma ele pode selecionar apenas um subconjunto de dados para visualização.
- Na aba *Hierarchy* novos grupamentos hierárquicos podem ser montados selecionando-se dentre os atributos quais formarão o topo da árvore e quais formarão as folhas.

Para apresentação dos dados dos testes selecionou-se o algoritmo de quadrados por permitir mais clareza na visualização. A hierarquia é constituída por seis níveis descritos a seguir:

Nível 1: Descreve os três tipos de modelo: Múltiplas Ontologias, Lucene ou Rede de Conceitos.

Nível 2: Descreve a base de conhecimento: Ontologia *Fuzzy*, Ontologia *Crisp* ou Base Automática gerada pelo modelo de Rede de Conceitos.

Nível 3: Define as combinações para os pesos we_G e we_S dadas pelos rótulos: $(we_G\,0.2\ we_S\,0.8)$ ou $(we_G\,0.8\ we_S\,0.2)$ ou $(we_G\,1.0\ we_S\,1.0)$. Para os valores de $w_G = 0.0$ e $w_S = 0.0$ o rótulo de we_G e we_S é 'não importa'. Neste caso o valor we_G e we_S não influencia o resultado final. No modelo Rede de Conceitos o rótulo de we_G e we_S também é 'não importa' pois estes pesos não são utilizados no modelo.

Nível 4: Define as combinações para os pesos w_G e w_S com os rótulos: $(w_G 0.3 w_S 0.7)$ ou $(w_G 0.7 w_S 0.3)$ ou $(w_G\,0.5\,w_S\,0.5)$ ou $(w_G\,1.0\,w_S\,1.0)$. Os valores de $(w_G\,0.0\,w_S\,0.0)$ são associados ao rótulo de $we_G\ we_S$ 'não importa'.

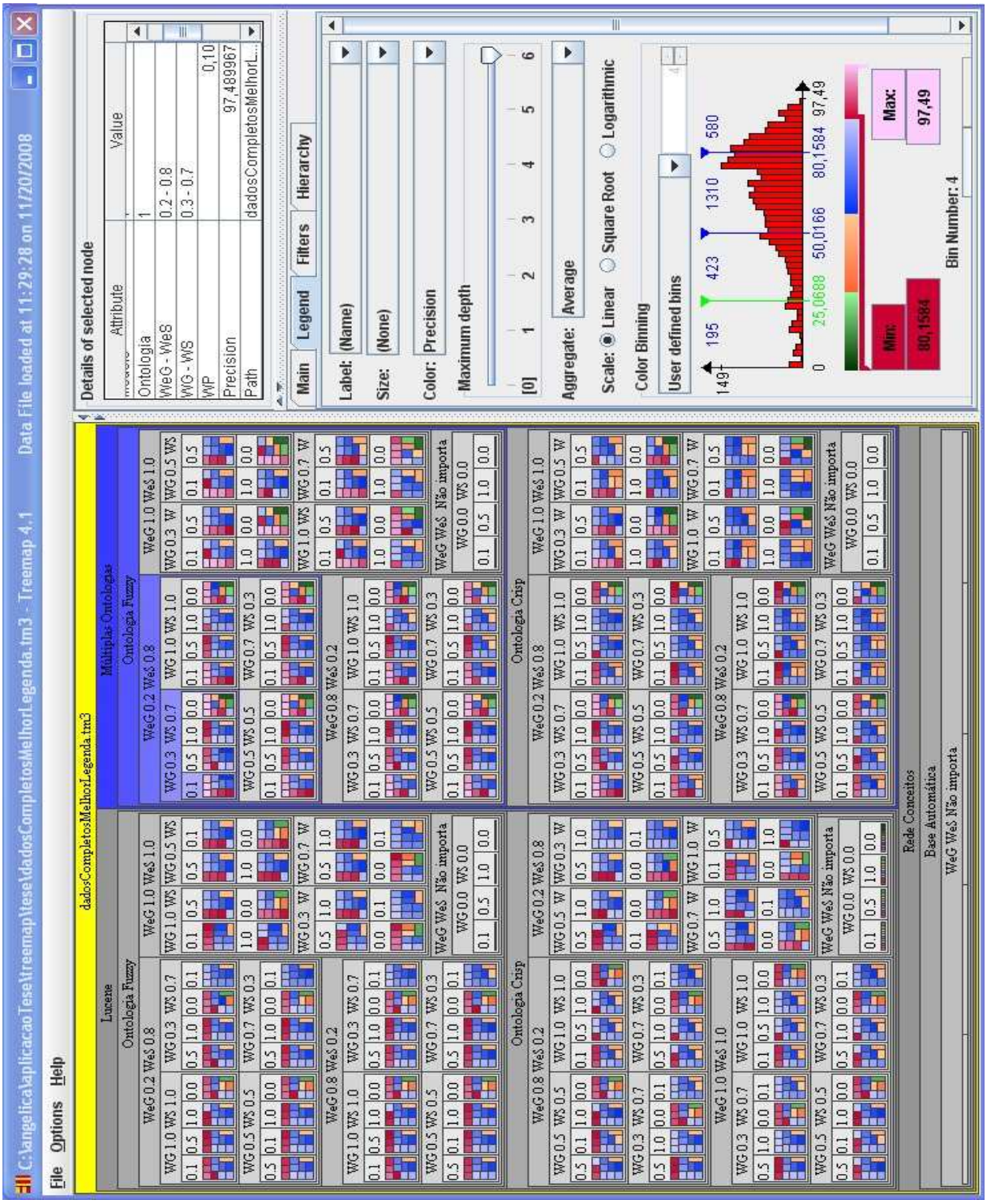

Fig. 5.10: Ferramenta Treemap.

Nível 5: Define as combinações para os pesos w_P dadas pelos rótulos: 0.0, 0.1, 0.5, 1.0.

Nível 6: Define os valores de precisão/cobertura com os rótulos: (Precisão 0, Precisão 10, Precisão 20, Precisão 30, Precisão 40, Precisão 50, Precisão 60, Precisão 70, Precisão 80, Precisão 90, Precisão 100). O rótulo indica o valor da precisão para a taxa de cobertura correspondente, ou seja, Precisão 10 vai mostrar o valor da precisão para a cobertura 10%. Os rótulos destes valores não aparecem na visualização. Os valores são representados pelos pequenos quadrados coloridos na tela. Se o mouse for posicionado em um dos quadrados o rótulo fica visível em uma janela de pop-up juntamente com o valor da precisão correspondente.

Para a visualização das medidas de precisão, ocorridas nas combinações de teste, escolheu-se a visualização por cor associada aos valores de precisão. Para tanto foram definidos quatro intervalos para os valores de precisão conforme a Tab. 5.3. Dentro de cada intervalo o tom mais claro está relacionado com os valores mais altos do intervalo e o tom mais escuro está relacionado com os valores mais baixos do intervalo. Os tons de cores dos intervalos podem ser observados na aba *Lengend* da Fig. 5.10. Por exemplo os tons de rosa claro são associados a valores próximos a $97,49$ e os tons de rosa escuro a valores próximos a 80. Para agregação da precisão nos níveis superiores escolheu-se a agregação pela média dos valores.

Cor	**Intervalo de precisão**
Rosa	≈ 80 até $97,49$ (maior valor)
Azul	≈ 50 até ≈ 80
Laranja	≈ 25 até ≈ 50
Verde	0 até ≈ 25

Tab. 5.3: Associação de cores a intervalos de precisão.

A Fig. 5.11 mostra a visão completa do mapa gerado pela ferramenta Treemap correspondente aos valores de configuração de visualização selecionados. O mapa mostra os níveis com a maior média de precisão ordenados da esquerda para a direita e de cima para baixo. Quanto maior o valor de precisão médio maior é o tamanho do quadrado nas hierarquias. Observando o mapa algumas conclusões podem ser tiradas.

O modelo Lucene obteve a maior média geral de precisão, seguido pelo modelo de múltiplas ontologias e depois pelo modelo de Rede de Conceitos.

Tanto no modelo Lucene quanto no modelo *fuzzy* de múltiplas ontologias a base de conhecimento composta de ontologias *fuzzy* ocasionou uma média de precisão maior que a base composta pelas ontologias *crisp*.

Fig. 5.11: Visualização dos dados dos modelos de recuperação de informação pela ferramenta Treemap.

Para valores baixos de cobertura, o modelo *fuzzy* de múltiplas ontologias exibiu melhores taxas de precisão que os demais modelos tanto ao utilizar ontologias *fuzzy* quanto ontologias *crisp*. Isto pode ser observado pela predominância da cor rosa claro nos valores de precisão ligados ao modelo. O Lucene exibiu melhores valores de precisão para valores de cobertura mais altos. Isto pode ser observado pelos quadrados laranja e verde claro quando comparado com os quadrados verdes escuros do modelo *fuzzy* de múltiplas ontologias.

No modelo *fuzzy* de múltiplas ontologias e no Lucene os valores de precisão média decaíram quando as associações de generalização e especialização não foram utilizadas, ou seja, $w_G 0.0, w_S 0.0$. Isto pode ser visto pois os quadrados correspondentes estão posicionados no canto direito inferior de cada modelo. O contrário aconteceu no modelo de rede de conceitos *fuzzy* onde os melhores valores de precisão são obtidos. Isto pode ser visto pois os quadrados correspondentes estão posicionados no canto esquerdo do modelo.

Quando apenas as palavras-chaves das consultas são consideradas os valores médios de precisão são os mais baixos alcançados pelos modelos. Isto pode ser visto pois os quadrados que correspondem à recuperação por palavras-chaves estão localizados no canto direito inferior de cada um dos modelos. Estes quadrados correspondem aos valores de pesos dados por $w_G = 0.0$, $w_S = 0.0$ e $w_P = 0.0$. No caso do modelo de rede de conceitos *fuzzy*, o modelo não recupera nenhum documento como mostrado na seção 5.5.2. Isto pode ser visto pela cor verde escura presente em todos os valores de precisão correspondentes.

Para os três modelos quando os valores de associação positiva não são considerados os valores das taxas de precisão, para valores de cobertura alto, caem rapidamente. Isto pode ser observado pela presença das cores verdes nos quadrados que representam curvas com rótulo 0.0 representando o valor $w_P = 0.0$.

O modelo rede de conceitos *fuzzy* apresentou a menor média da precisão. Isto pode ser observado pois o quadrado representando o modelo é menor que o dos outros modelos e se encontra localizado no canto inferior direito da tela. O melhor resultado ocorre quando $w_G\ 0.0,\ w_S\ 0.0$. Observar que neste caso qualquer valor de $w_P > 0.0$ gera o melhor resultado do modelo.

5.6 Resumo do Capítulo

Neste capítulo foram apresentados os resultados observados nos experimentos realizados nos três modelos de recuperação de informação estudados: o modelo *fuzzy* de múltiplas ontologias relacionadas proposto nesta tese, o modelo de rede de conceitos *fuzzy* e o Apache Lucene. Todos os modelos utilizam uma base de conhecimento para inferir novos conceitos a serem utilizados para recuperação dos documentos. A base de conhecimento possui três tipos de associações entre os conceitos:

associação de especialização *fuzzy*, associação de generalização *fuzzy* e associação positiva *fuzzy*.

No modelo *fuzzy* de múltiplas ontologias relacionadas a associação de especialização *fuzzy* e a associação de generalização *fuzzy* são utilizadas para construir múltiplas ontologias *lightweight* onde estas relações são utilizadas para definir a taxonomia das ontologias. A associação positiva *fuzzy* é utilizada para relacionar os conceitos das múltiplas ontologias. Este conjunto de ontologias relacionadas formam a base de conhecimento. O conhecimento existente na base é utilizado para fazer a expansão da consulta inicial do usuário visando recuperar mais documentos que sejam semanticamente relevantes à consulta. O método de expansão de consulta proposto também é testado na máquina de busca Lucene do projeto Apache.

No modelo de rede de conceitos os três tipos de associações são utilizados para construir uma rede de conceitos de multi-relacionamentos *fuzzy* automaticamente. Esta rede de conceitos compõe a base de conhecimento do modelo.

Para controlar a influência dos tipos de associação são associados pesos a cada um deles. A combinação destes pesos gerou uma série de combinações de testes para avaliar o desempenho dos modelos. Os dados coletados nos testes são organizados em três tipos de gráficos: gráfico da média das medidas de precisão, gráfico da precisão *versus* cobertura e o gráfico com os maiores valores para permitir uma comparação entre os modelos. Além dos gráficos foi gerada uma visualização com todas as combinações de teste utilizando a ferramenta Treemap.

A partir dos gráficos de média das medidas de precisão, pode-se tirar algumas conclusões em função dos valores associados aos parâmetros utilizados. Para os parâmetros we_G e we_S os melhores valores de precisão são obtidos quando $we_G = 0.2$ e $we_S = 0.8$ indicando que quando se privilegia os conceitos mais específicos, nas taxonomias que definem as ontologias, melhores resultados de precisão são obtidos.

Para os parâmetros w_G e w_S conclui-se que, nos modelos baseados nas ontologias relacionadas, o fato de considerar os conceitos mais específicos e mais gerais na expansão, isto é $w_G > 0.0$ e $w_S > 0.0$, ocorre uma melhoria no resultado de precisão final. No modelo de rede de conceitos *fuzzy* ao se considerar os conceitos mais específicos e mais gerais, na rede de conceitos, os valores de precisão apresentaram resultados abaixo da média. Este comportamento ocorre porque nas ontologias relacionadas as taxonomias possuem uma semântica mais definida do que no modelo de rede de conceitos *fuzzy* onde as relações de especialização e generalização são calculadas baseada na co-ocorrência de palavras.

Para o parâmetro w_P, valores de $w_P > 0$ melhoram o resultado de precisão final para todos os modelos indicando que a associação positiva é importante para melhorar o desempenho dos modelos.

A partir dos gráficos de precisão *versus* cobertura observa-se que todos os modelos melhoraram as medidas de precisão ao se utilizar uma base de conhecimento quando comparado com a busca

por palavras-chaves apenas. Ao se utilizar uma base de conhecimento que modela conceitos mais específicos e conceitos mais gerais o fato de privilegiar os conceitos mais específicos melhora o desempenho dos modelos. Ao se utilizar a associação positiva *fuzzy* a precisão de todos os modelos melhorou à medida que a taxa de cobertura aumentou. Isto indica que mais documentos relevantes estão localizados mais próximos do topo da lista de documentos retornados.

Considerando o gráfico com os maiores valores pode-se verificar que o modelo *fuzzy* de múltiplas ontologias apresentou melhor desempenho que os outros modelos ao se utilizar o conhecimento nas ontologias relacionadas tanto *fuzzy* como *crisp*. O modelo *fuzzy* de múltiplas ontologias e o Apache Lucene tiveram o mesmo padrão de comportamento ao se utilizar o conhecimento nas ontologias relacionadas mas os resultados do modelo *fuzzy* de múltiplas ontologias foram melhores. O modelo de rede de conceitos *fuzzy* apresentou resultados inferiores aos demais modelos.

A partir da visualização das combinações de teste pela ferramenta Treemap observa-se que a base de conhecimento composta de ontologias *fuzzy* ocasionou uma média de precisão maior que a base composta pelas ontologias *crisp* tanto para o modelo *fuzzy* de múltiplas ontologias como para o Apache Lucene. O modelo *fuzzy* de múltiplas ontologias exibiu taxas de precisão mais altas que os demais modelos. O uso da associação positiva contribui para manter o nível de precisão alto à medida que o valor de cobertura aumenta. O uso das associações de especialização e generalização nas ontologias contribuem para melhorar o valor da precisão. Já no modelo de rede de conceitos *fuzzy* o uso das associações de especialização e generalização acarreta o efeito contrário, ou seja, ocorre um decréscimo na medida de precisão quando o valor de cobertura aumenta. Para o modelo *fuzzy* de múltiplas ontologias e para o Apache Lucene o uso da base de conhecimento garantiu resultados melhores quando comparado com a consulta utilizando apenas as palavras-chaves.

Os experimentos apresentados neste capítulo mostram que o modelo *fuzzy* de múltiplas ontologias relacionadas obteve o melhor desempenho quando comparado com os outros modelos testados. O uso da base de conhecimento composta por múltiplas ontologias relacionadas e o método de expansão da consulta desenvolvido possibilitaram ao modelo de recuperação de informação proposto atingir seus bons resultados. A exploração da relação entre as ontologias, por meio da consideração da associação positiva *fuzzy* no processo de expansão da consulta, mostrou que há uma melhora na precisão das consultas fazendo com que mais documentos relevantes apareçam no topo da lista de documentos retornados.

Capítulo 6

Conclusões

Com a crescente popularidade da WWW mais pessoas têm acesso à informação e o volume desta informação vem crescendo ao longo do tempo. A área de recuperação de informação ganhou um novo desafio que é o de recuperar os documentos pelo significado da informação neles contida. Esta preocupação culminou no que está sendo chamado de Web Semântica que pretende recuperar a informação existente na WWW pelo seu conteúdo semântico. A informação poderá ser recuperada por meio de agentes inteligentes que são softwares que percorrem as páginas da WWW em busca de dados anotados semanticamente. Com o progresso da Web Semântica, a codificação de bases de conhecimento como ontologias têm aumentado. Aplicações de recuperação de informação estão empregando esta organização do conhecimento para melhorar a qualidade dos resultados retornando documentos semanticamente relacionados e mais relevantes à consulta inicial do usuário.

Uma coleção de documentos pode tratar de assuntos relacionados a vários domínios distintos. Neste caso, cada domínio pode ser representado por uma ontologia distinta. No caso do conhecimento dos domínios estarem relacionados pode-se estabelecer relacionamentos entre os conceitos das diversas ontologias. A recuperação de informação semântica, baseada no conhecimento organizado por ontologias distintas e relacionadas, foi apontada como uma área a ser explorada na literatura. Em geral, uma coleção de documentos utiliza apenas uma ontologia para organizar o conhecimento.

Neste contexto, a teoria de conjuntos *fuzzy* tem sido empregada para lidar com a imprecisão e a incerteza presente no conhecimento e no processo de recuperação de informação. Particularmente, ontologias *fuzzy* tem sido construídas para modelar a incerteza presente no conhecimento do domínio.

Esta tese apresentou um modelo *fuzzy* para melhorar o processo de recuperação de informação possibilitando a recuperação de documentos pelo seu conteúdo semântico baseada no conhecimento existente em uma base de conhecimento. Ao contrário de outros enfoques que consideram uma base de conhecimento composta por apenas uma ontologia, o modelo proposto explora a organização de conhecimento expressa em múltiplas ontologias *lightweight* independentes nas quais os relaciona-

mentos são expressos por relações *fuzzy*. Esta organização do conhecimento é usada para representar domínios cujos conceitos podem ser relacionados. Em alguns contextos os conceitos destes domínios podem ser relacionados por relações causais, espaciais ou de similaridade. O conhecimento expresso como ontologias *lightweight* relacionadas é utilizado em um novo método de expansão da consulta.

A avaliação experimental foi realizada utilizando uma amostra de 129 documentos extraídos de uma coleção de documentos no domínio da agrometeorologia no Brasil disponibilizada pela Embrapa. Para o experimento foram construídas duas ontologias *lightweight* nas versões *fuzzy* e *crisp*. Uma ontologia modela o domínio referente à divisão territorial do Brasil e a outra modela o domínio de climas que ocorrem no Brasil. Os experimentos foram realizados utilizando o modelo *fuzzy* de múltiplas ontologias relacionadas, proposto nesta tese, e o modelo de rede de conceitos *fuzzy* para fazer uma comparação. Além disto o método de expansão da consulta foi utilizado na máquina de busca do Apache Lucene.

Todos os modelos avaliados melhoraram as medidas de precisão ao utilizar uma base de conhecimento, quando comparado com a busca por palavras-chaves apenas. Ao utilizar uma base de conhecimento, que modela conceitos mais específicos e conceitos mais gerais, o fato de privilegiar os conceitos mais específicos melhora o desempenho dos modelos. Ao utilizar a associação positiva *fuzzy*, a precisão de todos os modelos melhorou à medida em que a taxa de cobertura aumentou. O modelo *fuzzy* de múltiplas ontologias e o Apache Lucene exibiram o mesmo padrão de comportamento. O modelo de rede de conceitos *fuzzy* apresentou resultados inferiores aos demais modelos.

O modelo *fuzzy* de múltiplas ontologias apresentou melhor desempenho que os outros modelos ao utilizar o conhecimento nas ontologias relacionadas, tanto *fuzzy* como *crisp*. O uso da base de conhecimento, composta por múltiplas ontologias relacionadas, e o método de expansão da consulta desenvolvido possibilitaram ao modelo de recuperação de informação proposto atingir seus bons resultados. A exploração da relação entre as ontologias, por meio da consideração da associação positiva *fuzzy* no processo de expansão da consulta, mostrou que há uma melhora na precisão das consultas fazendo com que mais documentos relevantes apareçam no topo da lista de documentos retornados.

Este capítulo apresenta as principais contribuições, os problemas em aberto e os trabalhos futuros relativos a esta tese.

6.1 Principais Contribuições

Esta tese possui três contribuições que constituem o fator de originalidade da mesma. A primeira contribuição é a proposta para organização do conhecimento formada por ontologias distintas e relacionadas. Cada ontologia representa um domínio de conhecimento. A construção de uma ontologia

para cada domínio de conhecimento facilita a manutenção e o reuso. Uma mesma ontologia pode ser reusada em diversas aplicações. Cada ontologia pode ser desenvolvida e evoluída de forma independente por especialistas do domínio. No caso de haver alterações nas ontologias os grupos de especialistas podem trabalhar em paralelo para posteriormente verificar se há necessidade de relacionar estas estruturas trabalhando apenas nos relacionamentos entre elas. No modelo proposto cada ontologia possui associações de especialização *fuzzy* e generalização *fuzzy* entre seus conceitos e as ontologias podem estar relacionadas entre si através de associação positiva *fuzzy*.

Para calcular os relacionamentos implícitos, entre os conceitos de uma ontologia, emprega-se o algoritmo de cálculo de fecho transitivo ponderado. O fecho transitivo ponderado permite que os conceitos mais próximos dentro da taxonomia tenham um valor de associação maior. Assim o modelo permite expressar os conceitos mais semanticamente relacionados, dentro das ontologias, de forma diferenciada. Os tipos de relações de associação são armazenados de forma independente através de relações matemáticas. Desta forma pode-se representar tanto ontologias *fuzzy* (relações matemáticas *fuzzy*) quanto ontologias *crisp* (relações matemáticas clássicas) tornando o modelo flexível para modelar diferentes tipos de ontologias.

A segunda contribuição é o método de expansão de consulta utilizando a base de conhecimento. O método explora os tipos de relações entre os conceitos das ontologias para selecionar, automaticamente, os novos conceitos a serem adicionados à consulta. O método associa um peso para cada tipo de associação permitindo ajustar a influência do tipo de associação no processo de expansão da consulta. O ajuste dos pesos proporciona melhoras nas taxas de precisão e cobertura. Quando o valor dos pesos das associações assume o valor 0 o sistema realiza a busca apenas pelas palavras-chaves da consulta inicial do usuário. A base de conhecimento e o método de expansão de consulta são independentes do modelo de recuperação de informação e podem ser utilizados em outras aplicações de recuperação de informação. Como um exemplo, nesta tese utilizou-se estas representações para recuperar informação utilizando a máquina de busca Apache Lucene.

A terceira contribuição é um modelo *fuzzy* de recuperação de informação que indexa os documentos e utiliza a base de conhecimento e o método de expansão de consulta propostos para fazer a recuperação semântica de documentos. O modelo *fuzzy* de recuperação de informação permitiu testar o desempenho do uso da base de conhecimento e do método de expansão da consulta. Os resultados obtidos com o modelo foram melhores do que dos outros modelos utilizados para comparação.

Uma contribuição secundária da tese é o conjunto das classes que implementam os três sistemas relativos aos modelos de recuperação de informação testados. O primeiro é o sistema que implementa o modelo *fuzzy* de recuperação de informação utilizando múltiplas ontologias relacionadas, proposto nesta tese. O segundo é um sistema que implementa o modelo de rede de conceitos *fuzzy* e o terceiro é um sistema que implementa uma aplicação de recuperação de informação utilizando a máquina de

busca do Apache Lucene. Todos os sistemas foram implementados utilizando a linguagem Java e o modelo de implementação MVC (*Model View Controller*) [10, 27, 104, 107]. As classes utilizadas para implementar estes modelos estão disponíveis na página da tese[1] .

6.2 Problemas em Aberto

Nesta tese foi desenvolvido e testado o modelo *fuzzy* utilizando múltiplas ontologias relacionadas. Os resultados obtidos na avaliação experimental foram satisfatórios e encorajam a evolução deste trabalho. Alguns problemas ficaram em aberto pois não foram aqui tratados.

A polissemia é a propriedade que uma mesma palavra tem de apresentar vários significados. No exemplo utilizado na avaliação experimental, o conceito "Am" pode significar uma classificação do clima Köppen como também indica o estado "Amazonas". Da mesma forma que o conceito "Af" é um tipo de classificação de clima Köppen mas também é a sigla para "Área foliar". A polissemia não foi considerada no tratamento do conhecimento na base e nem no modelo de recuperação de informação.

No que se refere às ontologias muitas delas podem possuir uma interseção no conhecimento que elas representam havendo uma sobreposição entre elas. Muitas vezes esta sobreposição permite realizar o mapeamento, alinhamento ou junção de ontologias. Neste trabalho foi considerado que os domínios representados por cada ontologia não se sobrepõem, isto é, não existe interseção de conhecimento nas ontologias.

Este trabalho considerou ontologias *lightweight* constituídas por uma taxonomia onde suas relações são expressas por associações de generalização *fuzzy* e especialização *fuzzy*. Além disto as ontologias podem ser relacionadas por associações positivas *fuzzy*. Todos estes tipos de associações são expressas por meio de relações matemáticas *fuzzy*. As ontologias também podem ser definidas por meio de linguagens de especificação lógica como OWL e KIF. Este tipo de linguagem não foi considerado na especificação das ontologias desenvolvidas. Para especificar as ontologias em uma linguagem baseada em lógica deve-se verificar sua capacidade em expressar o grau *fuzzy* nas relações entre os conceitos. Para a linguagem OWL, por exemplo, tem havido propostas de extensão para tratar aspectos da teoria *fuzzy* [44, 121].

A relação positiva entre as ontologias define uma relação de similaridade entre os conceitos. Na aplicação desenvolvida a associação positiva indica a relação espacial existente entre os conceitos que representam as entidades geográficas. Esta relação não possui uma especificação formal bem estabelecida. Este ponto deve ser revisto para o caso de se querer utilizar o modelo de organização de conhecimento proposto para outro tipo de aplicação que não a recuperação de informação.

[1] `http://www.dca.fee.unicamp.br/~ricarte/MORFuzz/`

Durante a avaliação experimental observou-se que a associação positiva estabelecida entre conceitos internos das ontologias, ou seja, conceitos que não estão nas folhas das taxonomias, embora intuitivamente signifique uma relação correta, pode acarretar em perda de semântica em função de inconsistências geradas quando se considera a forma de expansão de consulta proposta. Por exemplo, a associação do conceito “Região Nordeste” ao conceito “Semi-árido” possui um significado correto. Ao se fazer a expansão de uma consulta contendo o conceito “Semi-árido” o método vai considerar inicialmente o conceito “Região Nordeste” utilizando a associação positiva. Ao considerar as associações de generalização e especialização, em cada um dos domínios, o conceito “Maranhão” será automaticamente adicionado à consulta por ser um conceito mais específico do conceito “Região Nordeste”. Mas, ao observar o mapa da distribuição climática no território do Brasil, na Fig. A.1, pode-se verificar que o estado Maranhão não possui o clima semi-árido. Uma solução a ser investigada é considerar apenas as relações estabelecidas entre os conceitos nos nós folhas das ontologias e verificar uma forma de inferência que permita derivar as relações implícitas entre os conceitos mais internos nas taxonomias.

6.3 Trabalhos Futuros

Os trabalhos futuros possuem duas linhas de direção. A primeira inclui melhorias e evolução no modelo de recuperação de informação proposto e a segunda inclui possibilidades de uso do modelo em aplicações práticas.

Com relação a melhorias e evolução do modelo proposto, a representação da base de conhecimento por ontologias múltiplas e relacionadas, onde tanto as ontologias quanto as relações de associação positiva entre seus conceitos são representadas de forma independente, possibilita o reuso de ontologias. Neste caso apenas as associações positivas entre as ontologias devem ser estabelecidas. O estabelecimento das relações positivas entre os conceitos das ontologias de forma automática ou semi-automática constitui um problema a ser estudado. Nesta direção existe um trabalho que desenvolveu um método interativo para relacionar conceitos de uma coleção de documentos [16]. Técnicas de aprendizado de máquina tais como algoritmos genéticos e agrupamentos são utilizadas para extrair e relacionar conceitos de um corpo de documentos associados a uma área de conhecimento. A expectativa é que esta estratégia semi-automática vai ajudar a construir as associações positivas *fuzzy* entre os conceitos das ontologias. O próximo passo é a realização de testes, utilizando a estratégia proposta, para verificar sua eficiência em estabelecer as associações positivas entre conceitos de ontologias distintas. Uma outra linha de investigação, para estabelecer as relações positivas entre as ontologias, é o estudo do problema de mapeamento de ontologias [63] que propõe formas de alinhar ontologias [92] baseado em medidas de similaridade entre os seus conceitos.

Na base de conhecimento foram utilizadas ontologias *lightweight* que descrevem uma hierarquia de conceitos relacionados por relações de subsunção expressas por relações matemáticas. Um trabalho futuro é utilizar linguagens baseadas em lógica para definir as ontologias de uma maneira mais formal. O uso de linguagens baseadas em lógica permite a criação de ontologias *heavyweight*. Neste caso as ontologias podem expressar conhecimento que vai permitir realizar mais inferências e explorar melhor a semântica do domínio na recuperação de informação.

O conhecimento expresso em uma base de conhecimento pode mudar ao longo do tempo. A consideração das variações temporais do conhecimento também é um ponto a ser investigado. No domínio de conhecimento utilizado nesta tese esta variação está ligada a mudanças das características climáticas, e consequentemente do próprio clima, que podem ocorrer nas diferentes regiões geográficas ao longo do tempo.

Algumas questões levantadas em relação ao uso do modelo se referem a como será seu desempenho no caso do uso de ontologias constituídas por muitos conceitos ou no caso de haver muitas ontologias relacionadas. Estas questões demandam a necessidade de considerar bases de conhecimento maiores. Neste sentido o modelo proposto possui algumas características que permitem contornar estas questões.

Com relação ao uso de ontologias constituídas por muitos conceitos dois problemas podem aparecer. O primeiro é o processo de expansão da consulta ficar muito lento. O tempo de execução no processo de expansão pode ser resolvido fazendo um pré-processamento no sistema e expandindo todos os conceitos previamente. Estes conceitos expandidos podem ser armazenados em uma estrutura de dados. Assim quando uma nova consulta for tratada basta recuperar a expansão de cada conceito sem precisar processar a base de conhecimento. O segundo problema é a consulta expandida ficar com um número muito grande de novos conceitos. No caso de ontologias muito grandes pode haver vários níveis de conceitos mais gerais e mais específicos. Neste caso muitos conceitos novos poderão ser adicionados em função do número de níveis da taxonomia. Para o caso da taxonomia, que representa uma ontologia, expandir para vários níveis de conceitos mais gerais e mais específicos deve-se utilizar o peso que regula o cálculo do fecho transitivo ponderado nas ontologias. Este peso vai penalizar os conceitos que estiverem distantes na taxonomia e vai permitir limitar o número de conceitos a serem adicionados à consulta no processo de expansão. Conceitos cujo valor de associação estejam abaixo de um limite estabelecido não devem ser adicionados à consulta. A avaliação do desempenho do modelo considerando ontologias com um número grande de conceitos constitui um trabalho a ser realizado.

No caso de haver muitas ontologias relacionadas, uma proposta para limitar o número de conceitos a serem adicionados, devido à associação positiva *fuzzy*, é considerar apenas aqueles conceitos cujo grau da força de associação positiva esteja acima de um limite estabelecido. O estabelecimento de

valores limites para considerar ou não determinados conceitos, no processo de expansão da consulta, depende de uma análise prévia da base de conhecimento para se verificar o intervalo de valores nos graus de associação entre os conceitos nas ontologias. Nesta análise deve ser avaliada a escalabilidade deste limite. O uso do modelo considerando mais de duas ontologias, para avaliar seu desempenho, também é um trabalho futuro.

Um outro fator, que depende da análise prévia da base de conhecimento, é a atribuição de valores aos pesos $w_{eG},\ w_{eG},\ w_P,\ w_G,\ w_S$, utilizados no processo da expansão da consulta. A aplicação do modelo *fuzzy* de recuperação de informação, em outros domínios de conhecimento, é um trabalho futuro e vai ajudar a averiguar a estabilidade dos valores dos pesos, ou seja, vai ser possível verificar se os valores de pesos, utilizados na avaliação experimental desta tese, são eficazes em outros domínios de conhecimento ou se é necessário atribuir valores ajustados às características dos domínios.

Na motivação desta tese considerou-se o problema de recuperação de informação tratando-se recursos de informação que apresentam uma descrição textual para sua caracterização. Assim podem ser considerados documentos textuais em si e mesmo arquivos de imagem, som e vídeo desde que possuam informação textual associada que possa ser utilizada para caracterização e recuperação. Na validação do modelo, proposto na tese, considerou-se uma coleção de documentos textuais. Uma outra possibilidade é considerar bases compostas por outros tipos de recursos de informação, como imagens ou vídeos, com uma descrição textual associada, e verificar se o modelo *fuzzy* de recuperação de informação apresenta os bons resultados obtidos com documentos textuais.

Com relação a possibilidades de uso do modelo em aplicações práticas, a principal motivação para o desenvolvimento de um modelo de recuperação de informação nesta tese foi poder investigar o uso de bases de conhecimento no processo de recuperação de informação visando dotar a Embrapa com uma forma de recuperação de informação que explora o conhecimento de domínios. A disponibilização de informação produzida pela Embrapa sempre foi uma preocupação da empresa e, ao longo dos anos, ela, vem direcionando esforços nesta área. Alguns resultados são: desenvolvimento do AINFO, o aplicativo utilizado nas bibliotecas da empresa para cadastramento e recuperação das publicações do seu acervo [31]; as Bases de Dados da Pesquisa Agropecuária que agregam várias bases de dados que expressam o conhecimento gerado e adquirido pela Embrapa [32] e a Biblioteca Digital da Embrapa que apresenta os textos integrais, em meio eletrônico, dos trabalhos técnico-científicos gerados pela área de Pesquisa, Desenvolvimento e Inovação [34].

Além de disponibilizar sua informação em bases de documentos a Embrapa resolveu, através da Agência de Informação Embrapa [29, 36], doravante referida apenas como Agência, agregar maior conhecimento no processo de disponibilização e acesso à informação. A organização da informação na Agência é feita de modo hierárquico seguindo uma taxonomia, na forma de árvores do conhecimento, onde cada árvore trata de uma determinada tecnologia ou produto da Embrapa. Em cada

árvore tem-se o conhecimento organizado do nível mais genérico (nós internos da árvore) para os mais específicos (nós folhas da árvore). Na construção de uma árvore são selecionados um conjunto de termos ou conceitos pertencentes ao domínio de discurso de agricultores, técnicos de extensão rural e pesquisadores. Cada nó corresponde a um tópico do domínio de conhecimento sendo associado a um conceito e a uma descrição. Na Agência uma árvore do conhecimento pode ser vista como uma ontologia *lightweight* no sentido de que se trata de uma hierarquia construída com conceitos identificados no domínio de discurso dos atores envolvidos, ou seja, agricultores, pesquisadores e extensionistas rurais [117]. A Fig. 6.1 ilustra a árvore de conhecimento para o produto Feijão. A cada nó são associados documentos textuais sobre o tópico.

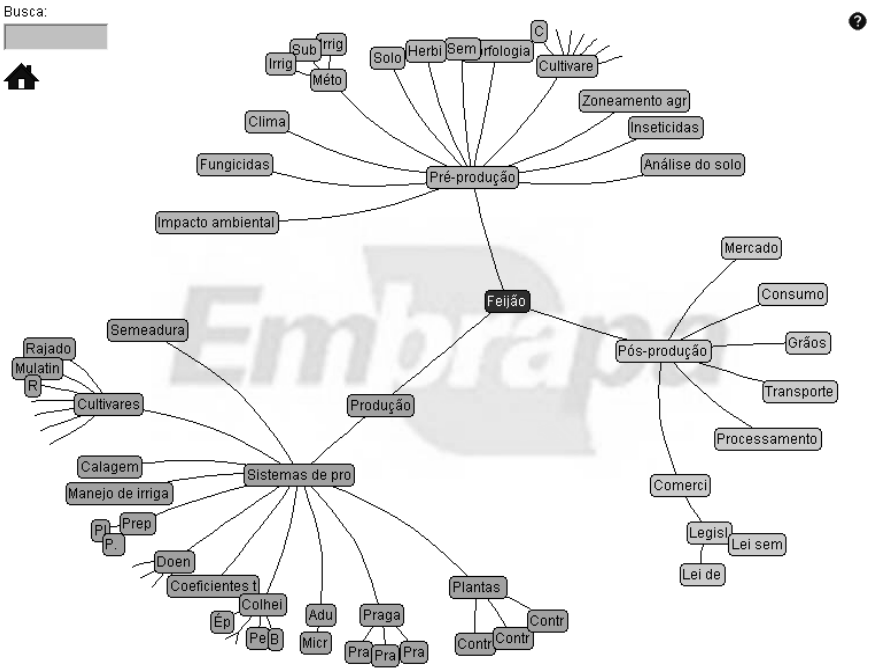

Fig. 6.1: Árvore do conhecimento do produto feijão [30].

A experiência na criação das árvores de conhecimento tem mostrado que existem informações que permeiam vários domínios e que por este motivo devem ser desenvolvidas como árvores separadas para serem referenciadas e reusadas por outras árvores. Por exemplo, na construção das árvores para produtos do tipo Feijão, Soja e Milho foi detectado que existem vários Processos Agrícolas referentes à plantio e manejo de solo que são comuns para estas culturas. Desta forma os processos agrícolas devem ser construídos como uma árvore de conhecimento independente para posteriormente ser relacionada às árvores referentes à Soja, Feijão e Milho. No caso das árvores associadas a produtos animais como Gado de corte, Suínos e Caprinos existem Processos Animais envolvendo aspectos de embalagem, armazenamento e distribuição da carne que também são comuns e que por isto deve ser

desenvolvida uma árvore de conhecimento específica para estes processos para ser relacionada às árvores de produtos animais. Assim, no futuro, existe a previsão de que a Agência seja composta por uma série de árvores de conhecimento sendo que várias delas estarão relacionadas entre si como mostra a Fig. 6.2.

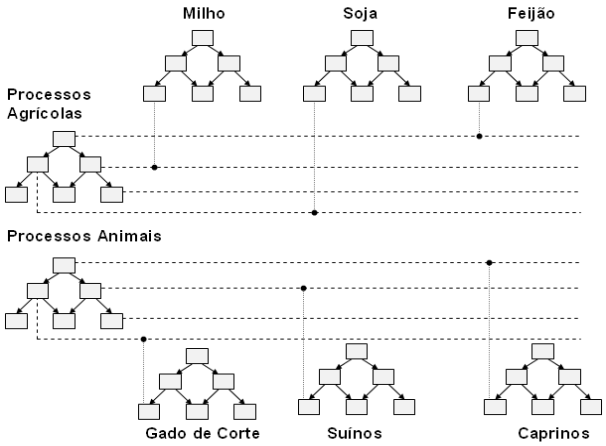

Fig. 6.2: Relacionamento entre as árvores de conhecimento da Agência de Informação Embrapa.

Este ambiente de disponibilização de informação, composto por várias árvores ou ontologias relacionadas associadas a documentos, constitui um campo para aplicação do método de expansão de consulta e recuperação de informação propostos nesta tese. Neste caso o modelo proposto será exercitado em outros domínios de conhecimento.

Um outro ponto a ser investigado é o uso do modelo de recuperação de informação proposto considerando um tesauro como a estrutura conceitual que compõe a base de conhecimento. Neste caso a associação de especialização vai armazenar a relação de termo mais específico do tesauro (NT), a associação de generalização vai armazenar a relação de termo mais geral do tesauro (BT) e a associação positiva vai armazenar a relação de termo relacionado do tesauro (RT). A idéia é investigar a recuperação de informação no acervo documental da Embrapa utilizando o modelo proposto na tese e um tesauro na área agrícola para compor a base de conhecimento. Um exemplo de tesauro agrícola é o Thesagro (Thesaurus Agrícola Nacional) [14]. A idéia é verificar se a forma de organização e representação de conhecimento pode se adaptar para armazenar as relações expressas no tesauro e se, neste caso, há melhoria no processo de recuperação de informação ao utilizar o processo de expansão de consulta proposto na tese.

Referências Bibliográficas

[1] Muhammad Abulaish and Lipika Dey. A fuzzy ontology generation framework for handling uncertainties and nonuniformity in domain knowledge description. In *ICCTA '07: Proceedings of the International Conference on Computing: Theory and Applications*, pages 287–293, Washington, DC, USA, 2007. IEEE Computer Society.

[2] Mikhail Ageev, Boris Dobrov, and Natalia Loukachevitch. Socio-political thesaurus in concept-based information retrieval. *Lecture Notes in Computer Science*, 4022:141–150, 2006.

[3] Eija Airio, Kalervo Järvelin, Pirkko Saatsi, Jaana Kekäläinen, and Sari Suomela. CIRI - an ontology-based query interface for text retrieval. In Eero Hyvönen, Tomi Kauppinen, Mirva Salminen, Kim Viljanen, and Pekka Ala-Siuru, editors, *Proceedings of the 11th Finnish Artificial Intelligence Conference STeP 2004, September 1-3, Vantaa, Finland*, volume 2 of *Conference Series – No 20*, pages 73–82. Finnish Artificial Intelligence Society, 2004.

[4] Grigoris Antoniou and Frank van Harmelen. *A Semantic Web Primer*. The MIT Press, 2004.

[5] Apache Project. Apache Download Mirrors . Página na internet, The Apache Software Foundation, Acesso em: Novembro 2008. `http://www.apache.org/dyn/closer.cgi/lucene/java/`.

[6] Apache Project. Apache Lucene Overview. Página na internet, The Apache Software Foundation, Acesso em: Agosto 2008. `http://lucene.apache.org/java/docs/index.html`.

[7] Apache Project. Query Parser Syntax. Página na internet, The Apache Software Foundation, Acesso em: Novembro 2008. `http://lucene.apache.org/java/2_3_2/queryparsersyntax.html`.

[8] Nathalie Aussenac-Gilles and Josiane Mothe. Ontologies as background knowledge to explore document collections. In *RIAO 2004: Recherche d'Information Assistée par Ordinateur.*

Coupling approaches, coupling media and coupling languages for information retrieval, pages 129–142, 2004.

[9] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[10] Hans Bergsten. *JavaServer Pages*. O'Reilly Media Inc., 2003.

[11] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, pages 34–40, Maio 2001.

[12] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information Processing and Management*, 43(4):866–886, 2007.

[13] Biblioteca Nacional Agricultura. BINAGRI. Página na internet, Ministério Agricultura, Pecuária e Abastecimento, Acesso em: janeiro 2009. `http://extranet.agricultura.gov.br/primeira_pagina/sistemas/BINAGRI.htm`.

[14] BINAGRI. Thesaurus Agrícola Nacional. Página na internet, Biblioteca Nacional de Agricultura, Acesso em: janeiro 2009. `http://www.agricultura.gov.br/portal/page?_pageid=33,959135&_dad=portal&_schema=PORTAL`.

[15] Gloria Bordogna and Gabriella Pasi. Modeling vagueness in information retrieval. *Lectures on information retrieval*, pages 207–241, 2001.

[16] Sérgio William Botero. Extração de relações semânticas via análise de correlação de termos em documentos. Master's thesis, Universidade Estadual de Campinas – Faculdade de Engenharia Elétrica e de Computação, Campinas, São Paulo, Brasil, Dezembro 2008.

[17] Davide Buscaldi, Paolo Rosso, and Piedachu Peris García. Inferring geographical ontologies from multiple resources for geographical information retrieval. In *GIR '06: Proceedings of the 3rd ACM workshop on Geographical information retrieval*, New York, NY, USA, 2006. ACM.

[18] B. Chandrasekaran, John R. Josephson, and V. Richard Benjamins. What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 14(1):20–26, 1999.

[19] Shyi-Ming Chen, Yih-Jen Horng, and Chia-Hoang Lee. Fuzzy information retrieval based on multi-relationship fuzzy concept networks. *Fuzzy Sets and Systems*, 140(1):183–205, 2003.

[20] Steven M. Cherry. Semantic web: weaving a web of ideas. *IEEE Spectrum*, 39(9):65–69, 2002.

[21] Y. H. Chuang and C. C. Kao. Computer expansion of boolean expressions. In *DAC '71: Proceedings of the 8th workshop on Design automation*, pages 378–383, New York, NY, USA, 1971. ACM Press.

[22] Fabio Crestani, Mounia Lalmas, Cornelis J. Van Rijsbergen, and Iain Campbell. Is this document relevant? probably: A survey of probabilistic models in information retrieval. *ACM Computing Surveys*, 30(4):528–552, 1998.

[23] Stefan Decker, Michael Erdmann, Dieter Fensel, and Rudi Studer. Ontobroker: Ontology based access to distributed and semi-structured information. In *Database Semantics: Semantic Issues in Multimedia Systems*, pages 351–369. Kluwer Academic Publisher, 1999.

[24] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[25] William Denton. How to Make a Faceted Classification and Put It On the Web. Página na internet, Miskatonic University Press, Acesso em: abril 2009. `http://www.miskatonic.org/library/facet-web-howto.html`.

[26] M. C. Díaz-Galiano, M. Á. García-Cumbreras, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. A. Ureña-López. Integrating mesh ontology to improve medical information retrieval. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 601–606, Berlin, Heidelberg, 2008. Springer-Verlag.

[27] Jon Eaves, Rupert Jones, and Warner Godfrey. *Apache Tomcat Bible*. Wiley Publishing Inc., 2002.

[28] Elsevier. Elsevier Bibliographic Databases. Página na internet, Elsevier, Acesso em: Abril 2006. `http://www1.elsevier.com/homepage/sah/spd/site/`.

[29] Embrapa. Agência de Informação Embrapa. Página na internet, Embrapa, Acesso em: Outubro 2008. `http://www.agencia.cnptia.embrapa.br/`.

[30] Embrapa. Agência de Informação Feijão. Página na internet, Embrapa, Acesso em: Outubro 2008. `http://www.agencia.cnptia.embrapa.br/Agencia4/AG01/Abertura.html`.

[31] Embrapa. AINFO . Página na internet, Empresa Brasileira de Pesquisa Agropecuária, Acesso em: Novembro 2008. `http://www.ainfo.cnptia.embrapa.br/index.html`.

[32] Embrapa. Bases de Dados da Pesquisa Agropecuária . Página na internet, Empresa Brasileira de Pesquisa Agropecuária, Acesso em: Novembro 2008. http://www.bdpa.cnptia.embrapa.br/.

[33] Embrapa. Bases Temáticas Embrapa. Página na internet, Embrapa, Acesso em: Junho 2008. http://www.bdpa.cnptia.embrapa.br/index.jsp?url=basesTematicas.jsp&baseDados=AGROMETEOROLOGIA.

[34] Embrapa. Biblioteca Digital Embrapa . Página na internet, Empresa Brasileira de Pesquisa Agropecuária, Acesso em: Novembro 2008. http://www.bibdigital.cnptia.embrapa.br/.

[35] Embrapa. Empresa Brasileira de Pesquisa Agropecuária. Página na internet, Embrapa, Acesso em: Junho 2008. http://www.embrapa.br/.

[36] Silvio R. M. Evangelista, Kleber X. S. de Souza, Marcia I. F. Souza, Sérgio A. B. da Cruz, Maria A. A. Leite, Adriana D. dos Santos, and Maria F. Moura. Gerenciador de conteúdos da agência embrapa de informação. In *International Symposium on Knowledge Management - ISKM*, pages 1–12, CD–ROM ISKM. Pontifícia Universidade Católica do Paraná, 2003.

[37] Wei-Dong Fang, Ling Zhang, Yan-Xuan Wang, and Shou-Bin Dong. Toward a semantic search engine based on ontologies. In *Fourth International Conference on Machine Learning and Cybernetics*, pages 1913–1918, Washington, DC, USA, 2005. IEEE Computer Society.

[38] FAO. Food and Agriculture Organization of the United Nations. Página na internet, Food and Agriculture Organization of the United Nations, Acesso em: Junho 2006. http://www.fao.org/.

[39] Tim Finin, James Mayfield, Anupam Joshi, R. Scott Cost, and Clay Fink. Information retrieval and the semantic web. In *HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4*, page 113.1, Washington, DC, USA, 2005. IEEE Computer Society.

[40] Food and Agriculture Organization of the United Nations. AGROVOC THESAURUS. Página na internet, Food and Agriculture Organization of the United Nations, Acesso em: Junho 2006. http://www.icpa.ro/AgroWeb/AIC/RACC/Agrovoc.htm.

[41] Gleb Frank. JTP: An Object-Oriented Modular Reasoning System. Página na internet, Stanford University, Acesso em: Abril 2009. http://www.ksl.stanford.edu/software/JTP/.

[42] Fred Freitas, Heiner Stuckenschmidt, and Natalya F. Noy. Ontology issues and applications: Guest editorial. *Journal of the Brazilian Computer Society*, pages 5–16, Novembro 2005.

[43] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–480, New York, NY, USA, 1988. ACM Press.

[44] Mingxia Gao and Chunnian Liu. Extending owl by fuzzy description logic. In *ICTAI '05: Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence*, pages 562–567, Washington, DC, USA, 2005. IEEE Computer Society.

[45] Michael R. Genesereth and Richard E. Fikes. Knowledge Interchange Format. Version 3.0. Reference Manual. Página na internet, Stanford University, Acesso em: Abril 2009. `http://logic.stanford.edu/kif/Hypertext/kif-manual.html`.

[46] Brian Goetz. The Lucene search engine: Powerful, flexible, and free . Página na internet, Java World, Acesso em: Novembro 2008. `http://www.javaworld.com/jw-09-2000/jw-0915-lucene.html?page=1`.

[47] Hagar Espanha Gomes. *Manual de Elaboração de Tesauros Monolíngües*. Ministério da Educação e Ministério da Ciência e Tecnologia, Brasília, Brasil, 1990.

[48] Asunción Gomez-Pérez, Mariano Fernández-Lopez, and Oscar Corcho. *Ontological Engineering*. Springer-Verlag, 2003.

[49] Andrew S. Gordon and Eric A. Domeshek. Deja vu: a knowledge-rich interface for retrieval in digital libraries. In *IUI '98: Proceedings of the 3rd international conference on Intelligent user interfaces*, pages 127–134, New York, NY, USA, 1998. ACM Press.

[50] Otis Gospodnetic. Introduction to Text Indexing with Apache Jakarta Lucene . Página na internet, O'Reilly, Acesso em: Novembro 2008. `http://www.onjava.com/pub/a/onjava/2003/01/15/lucene.html?page=1`.

[51] 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.

[52] Nicola Guarino. Understanding, building and using ontologies. *International Journal of Human-Computer Studies*, 46(2-3):293–310, 1997.

[53] Nicola Guarino. Formal ontology and information systems. In *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems*, pages 3– 15. IOS Press, 1998.

[54] Nicola Guarino, Claudio Masolo, and Guido Vetere. Ontoseek: Content-based access to the web. *IEEE Intelligent Systems*, 14(3):70–80, 1999.

[55] Frank van Harmelen, Peter F. Patel-Schneider, and Ian Horrocks. DAML+OIL (March 2001) ontology markup language. Página na internet, DARPA, Acesso em: Abril 2009. `http://www.daml.org/2001/03/reference.html`.

[56] Yih-Jen Horng, Shyi-Ming Chen, and Chia-Hoang Lee. Automatically constructing multi-relationship fuzzy concept networks in fuzzy information retrieval systems. In *The 10th IEEE International Conference on Fuzzy Systems*, pages 606–609. IEEE Computer Society, 2001.

[57] Yih-Jen Horng, Shyi-Ming Chen, and Chia-Hoang Lee. Automatically constructing multi-relationship fuzzy concept networks for document retrieval. *Applied Artificial Intelligence,*, 17(1):303–328, 2003.

[58] Ian Horrocks. Using an Expressive Description Logic: FaCT or Fiction? Página na internet, University of Manchester, Acesso em: Abril 2009. `http://www.comlab.ox.ac.uk/people/ian.horrocks/Publications/download/1998/kr98.pdf`.

[59] Human-Computer Interaction Lab. Treemap. Página na internet, University of Maryland, Acesso em: Novembro 2008. `http://www.cs.umd.edu/hcil/treemap/`.

[60] IBGE. IBGE Mapas Interativos. Página na internet, Instituto Brasileiro de Geografia e Estatística; Ministério do Planejamento, Orçamento e Gestão, Acesso em: Setembro 2006. `http://mapas.ibge.gov.br/clima/viewer.htm`.

[61] Theresa I. Jefferson and Thomas J. Nagy. A domain-driven approach to improving search effectiveness in traditional online catalogs. *Inf. Manage.*, 39(7):559–570, 2002.

[62] Brian Johnson and Ben Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *VIS '91: Proceedings of the 2nd conference on Visualization '91*, pages 284–291, Los Alamitos, CA, USA, 1991. IEEE Computer Society Press.

[63] Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping: the state of the art. *The Knowledge Engineering Review*, 18(1):1–31, 2003.

[64] Mukundan Karthik, Mariappan Marikkannan, and Arputharaj Kannan. An intelligent system for semantic information retrieval information from textual web documents. In *IWCF '08: Proceedings of the 2nd international workshop on Computational Forensics*, pages 135–146, Berlin, Heidelberg, 2008. Springer-Verlag.

[65] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[66] Latifur Khan, Dennis McLeod, and Eduard Hovy. Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal*, 13(1):71–85, 2004.

[67] Michael Kifer, Georg Lausen, and James Wu. Logical foundations of object-oriented and frame-based languages. *Journal of the ACM*, 42(4):741–843, 1995.

[68] Michel Klein. Combining and relating ontologies: an analysis of problems and solutions. In Asuncion Gomez-Perez, Michael Gruninger, Heiner Stuckenschmidt, and Michael Uschold, editors, *Workshop on Ontologies and Information Sharing, IJCAI'01*, Seattle, USA, august 2001.

[69] George J. Klir, Ute St. Clair, and Bo Yuan. *Fuzzy set theory : Foundations and Applications*. Prentice Hall, 1997.

[70] George J. Klir and Bo Yuan. *Fuzzy sets and fuzzy logic : Theory and Applications*. Prentice Hall, 1995.

[71] Kevin Knight and Steve K. Luk. Building a large-scale knowledge base for machine translation. In *AAAI '94: Proceedings of the twelfth national conference on Artificial intelligence (vol. 1)*, pages 773–778, Menlo Park, CA, USA, 1994. American Association for Artificial Intelligence.

[72] Raymond Y. K. Lau, Yuefeng Li, and Yue Xu. Mining fuzzy domain ontology from textual databases. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 156–162, Washington, DC, USA, 2007. IEEE Computer Society.

[73] Maria Angelica Leite and Ivan L. Ricarte. A framework for information retrieval based on fuzzy relations and multiple ontologies. In *IBERAMIA '08: Proceedings of the 11th Ibero-American conference on AI*, pages 292–301, Berlin, Heidelberg, 2008. Springer-Verlag.

[74] Maria Angelica Leite and Ivan L. Ricarte. Using multiple related ontologies in a fuzzy information retrieval model. In *Third Workshop on Ontologies and Their Applications*, Bahia, Brasil, 2008. Universidade Federal da Bahia.

[75] Maria Angelica A. Leite and Ivan L. M. Ricarte. Document retrieval using fuzzy related geographic ontologies. In *GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 47–54, New York, NY, USA, 2008. ACM.

[76] Maria Angelica A. Leite and Ivan L. M. Ricarte. Fuzzy information retrieval model based on multiple related ontologies. In *20th IEEE International Conference on Tools with Artificial Intelligence*, pages 309–316, Washington, DC, USA, 2008. IEEE Computer Society.

[77] Chenxi Lin, Lei Zhang, Jian Zhou, Yin Yang, and Yong Yu. Sports: Semantic+portal+service. In *ECAI 2004: Workshop on Application of Semantic Web Technologies to Web Communities, volume 107 of CEUR-WS*, 2004.

[78] James N. K. Liu. An intelligent system integrated with fuzzy ontology for product recommendation and retrieval. In *FS'07: Proceedings of the 8th Conference on 8th WSEAS International Conference on Fuzzy Systems*, pages 180–185, Stevens Point, Wisconsin, USA, 2007. World Scientific and Engineering Academy and Society (WSEAS).

[79] S. Liu, C. A. McMahon, M. J. Darlington, S. J. Culley, and P. J. wild. An automatic mark-up approach for structured document retrieval in engineering design. *The International Journal of Advanced Manufacturing Technology*, 38(3–4):418–425, 2008.

[80] S. Liu, C.A. McMahon, M.J. Darlington, S.J. Culley, and P.J. Wild. A computational framework for retrieval of document fragments based on decomposition schemes in engineering information management. *Advanced Engineering Informatics*, 20(4):401–413, 2006.

[81] Jane Lomax and Alexa T. McCray. Mapping the gene ontology into the unified medical language system: Research papers. *Comp. Funct. Genomics*, 5(4):354–361, 2004.

[82] Natalia Loukachevitch and Boris Dobrov. Evaluation of thesaurus on socio-political life as information retrieval tool. In *(LREC2002: Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 115–121, Paris, France, 2002. Association European Language Resources.

[83] Miguel R. Luaces, Jose R. Paramá, Oscar Pedreira, and Diego Seco. An ontology-based index to retrieve documents with geographic information. In *SSDBM '08: Proceedings of the 20th International Conference on Scientific and Statistical Database Management*, pages 384–400, Berlin, Heidelberg, 2008. Springer-Verlag.

[84] Miguel R. Luaces, Jose R. Parama, Oscar Pedreira, Diego Seco, and Jose R. R. Viqueira. An index structure to retrieve documents with geographic information. In *DEXA '07: Proceedings*

of the 18th International Conference on Database and Expert Systems Applications, pages 64–68, Washington, DC, USA, 2007. IEEE Computer Society.

[85] Christopher D. Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. Cambridge MIT, 1993.

[86] James Mayfield. Ontologies and text retrieval. *The Knowledge Engineering Review*, 17(1):71–75, 2002.

[87] George A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[88] Ralf Möller and Volker Haarslev. RACER: Renamed Abox and Concept Expression Reasoner. Página na internet, Hamburg University of Technology and Concordia University, Acesso em: Maio 2006. `http://www.sts.tu-harburg.de/~r.f.moeller/racer/`.

[89] Josiane Mothe, Claude Chrisment, Bernard Dousset, and Joel Alaux. Doccube: multi-dimensional visualisation and exploration of large document sets. *Journal of the American Society for Information Science and Technology*, 54(7):650–659, 2003.

[90] Jibran Mustafa, Sharifullah Khan, and Khalid Latif. Ontology based semantic information retrieval. In *Fourth International IEEE Conference on Intelligent Systems*, pages 2214–2219, Washington, DC, USA, 2008. IEEE Computer Society.

[91] Natalya F. Noy. Semantic integration: a survey of ontology-based approaches. *SIGMOD Record*, 33(4):65–70, 2004.

[92] Natalya Fridman Noy and Mark A. Musen. Prompt: Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 450–455. AAAI Press / The MIT Press, 2000.

[93] Yasushi Ogawa, Tetsuya Morita, and Kiyohiko Kobayashi. A fuzzy document retrieval system using the keyword connection matrix and a learning method. *Fuzzy Sets and Systems*, 39(2):163–179, 1991.

[94] Ontoprise. Ontobroker. Página na internet, Ontoprise GmbH, Acesso em: Abril 2009. `http://www.ontoprise.de/en/home/products/ontobroker/`.

[95] David Parry. A fuzzy ontology for medical document retrieval. In *ACSW Frontiers '04: Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation*, pages 121–126, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

[96] Gordon W. Paynter and Ian H. Witten. A combined phrase and thesaurus browser for large document collections. In *ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 25–36, London, UK, 2001. Springer-Verlag.

[97] Gordon W. Paynter, Ian H. Witten, Sally Jo Cunningham, and George Buchanan. Scalable browsing for large collections: a case study. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 215–223, New York, NY, USA, 2000. ACM Press.

[98] Christian Paz-Trillo, Renata Wassermann, and Paula P. Braga. An information retrieval application using ontologies. *Journal of the Brazilian Computer Society*, pages 17–31, Março 2006.

[99] Witold Pedrycz and Fernando Gomide. *An introduction to fuzzy sets : Analysis and Design*. MIT Press, Cambridge, Massachusetts, 1998.

[100] Witold Pedrycz and Fernando Gomide. *Fuzzy Systems Engineering: Toward Human–Centric Computing*. John Wiley & Sons, Inc, 2007.

[101] Raquel Pereira, Ivan Ricarte, and Fernando Gomide. Fuzzy relational ontological model in information search systems. In *Elie Sanchez. (Org.). Fuzzy Logic and The Semantic Web*, pages 395–412, Amsterdan, 2006. Elsevier B. V.

[102] Raquel Carlos Pereira. Modelo Ontológico Relacional *Fuzzy* em Sistemas de Recuperação de Informação Textual. Tese de mestrado, Faculdade de Engenharia Elétrica e de Computação, UNICAMP, Novembro 2004.

[103] José R. Pérez-Agüera and Lourdes Araujo. Query expansion with an automatically generated thesaurus. *Lecture Notes in Computer Science*, 4224:771–778, 2006.

[104] Bruce W. Perry. *Java Servlet & JSP Cookbook*. O'Reilly Media Inc., 2004.

[105] H. Sofia Pinto, Asunción Gómez-Pérez, and João P. Martins. Some issues on ontology integration. In *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem Solving Methods*, 1999.

[106] Projeto SISGA. Mapa do Clima no Brasil. Página na internet, Universidade Regional de Blumenau, Acesso em: Junho 2008. `http://www2.inf.furb.br/sisga/educacao/ensino/mapaClima.php`.

[107] Suresh Rajagopalan, Ramesh Rajamani, Ramesh Krishnaswamy, and Sridhar Vijendran. *Java Servlet Programming Bible*. Wiley Publishing Inc., 2003.

[108] Ivan L. M. Ricarte and Fernando A. C. Gomide. A reference model for intelligent information search. In *Nikravesh, M.; Zadeh, L. A.; Azvine, B.; Yager, R.R. (Org.). Enhancing the Power of Internet - Studies in Fuzziness and Soft Computing*, pages 327–346. Heidelberg: Springer-Verlag, 2004.

[109] Stuart Russel and Peter Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, 2003.

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

[111] SAX Project. About SAX. Página na internet, SourceForge.Net, Acesso em: janeiro 2009. `http://www.saxproject.org/about.html`.

[112] Urvi Shah, Tim Finin, and Anupam Joshi. Information retrieval on the semantic web. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 461–468, New York, NY, USA, 2002. ACM Press.

[113] Ben Shneiderman. Discovering business intelligence using treemap visualizations. In *Beye-Network*, Boulder, CO, USA, 2006. Powell Media, LLC.

[114] Michael K. Smith, Chris Welty, and Deborah L. McGuinness. *OWL Web Ontology Language Guide*. W3C - World Wide Web Consortium, February 2008.

[115] Von-Wun Soo, Chen-Yu Lee, Chung-Cheng Li, Shu Lei Chen, and Ching chih Chen. Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 61–72, Washington, DC, USA, 2003. IEEE Computer Society.

[116] Von-Wun Soo, Chen-Yu Lee, Jaw Jium Yeh, and Ching chih Chen. Using sharable ontology to retrieve historical images. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 197–198, New York, NY, USA, 2002. ACM Press.

[117] Kleber X. S. de Souza, Adriana D. dos Santos, and Silvio R. M. Evangelista. Visualization of ontologies through hypertrees. In *CLIHC '03: Proceedings of the Latin American conference on Human-computer interaction*, pages 251–255, New York, NY, USA, 2003. ACM.

[118] Kleber Xavier Sampaio de Souza, Joseph Davis, and Silvio Roberto de Medeiros Evangelista. Aligning ontologies, evaluating concept similarities and visualizing results. *Lecture Notes in Computer Science*, 3870/2006:211–236, 2006.

[119] Stanford Medical Informatics. The Protégé Ontology Editor and Knowledge Acquisition System. Página na internet, Stanford University School of Medicine, Acesso em: Outubro 2006. `http://protege.stanford.edu/`.

[120] Mike Steckel. Ranganathan for IAs. Página na internet, Boxes and Arrows, Acesso em: Julho 2002. `http://www.boxesandarrows.com/view/ranganathan_for_ias`.

[121] G. Stolios, N. Simou, G. Stamou, and S. Kollias. Uncertainty and the semantic web. *IEEE Intelligent Systems*, 21(5):84–87, 2006.

[122] Umberto Straccia. Reasoning within fuzzy description logics. *Journal of Artificial Intelligence Research*, 14:137–166, 2001.

[123] Heiner Stuckenschmidt, Frank van Harmelen, Anita de Waard, Tony Scerri, Ravinder Bhogal, Jan van Buel, Ian Crowlesmith, Christiaan Fluit, Arjohn Kampman, Jeen Broekstra, and Erik van Mulligen. Exploring large document repositories with rdf technology: The dope project. *IEEE Intelligent Systems*, 19(3):34–40, 2004.

[124] Sari Suomela and Jaana Kekäläinen. User evaluation of ontology as query construction tool. *Information Retrieval*, 9(4):455–475, 2006.

[125] TrebleCLEF Consortium. Cross-language evaluation forum. Página na internet, TrebleCLEF Coordination Action, Acesso em: Novembro 2008. `http://www.clef-campaign.org/`.

[126] Michal Tvarozek and Maria Bielikova. Personalized faceted navigation for multimedia collections. In *SMAP '07: Proceedings of the Second International Workshop on Semantic Media Adaptation and Personalization*, pages 104–109, Washington, DC, USA, 2007. IEEE Computer Society.

[127] D. H. Widyantoro and J. Yen. A fuzzy ontology-based abstract search engine and its user studies. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, pages 1291–1294, Washington, DC, USA, 2001. IEEE Computer Society.

[128] Wikipédia. Brasil. Página na internet, Wikimedia Foundation, Acesso em: Junho 2008. `http://pt.wikipedia.org/wiki/Brasil`.

[129] Wikipédia. Classificação climática de Köppen-Geiger. Página na internet, Wikimedia Foundation, Acesso em: Junho 2008. `http://pt.wikipedia.org/wiki/Classificação_do_clima_de_Köppen`.

[130] Wikipédia. Faceted classification. Página na internet, Wikimedia Foundation, Acesso em: Abril 2009. `http://en.wikipedia.org/wiki/Faceted_classification`.

[131] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. Faceted metadata for image search and browsing. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 401–408, New York, NY, USA, 2003. ACM Press.

[132] Lei Zhang, Yong Yu, Jian Zhou, ChenXi Lin, and Yin Yang. An enhanced model for searching in semantic portals. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 453–462, New York, NY, USA, 2005. ACM.

[133] Mingquan Zhou, Guohua Geng, and Shiguo Huang. Ontology development for insect morphology and taxonomy system. In *WI-IATW '06: Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology*, pages 324–330, Washington, DC, USA, 2006. IEEE Computer Society.

[134] Leyla Zhuhadar and Olfa Nasraoui. Semantic information retrieval for personalized e-learning. In *20th IEEE International Conference on Tools with Artificial Intelligence*, pages 364–368, Washington, DC, USA, 2008. IEEE Computer Society.

Apêndice A

Preparação dos Dados Experimentais

Este apêndice mostra como os dados utilizados no experimento, para validação dos modelos de recuperação de informação, foram elaborados. O apêndice descreve a forma como as ontologias foram construídas, o processo de seleção da coleção de documentos, a preparação das consultas e a atribuição dos documentos relevantes para cada consulta.

A.1 Construção das Ontologias

Duas ontologias *lightweight fuzzy* foram construídas. O mapa do território do Brasil [106], na Fig. A.1, é considerado na construção das ontologias. Ele contém a distribuição da classificação de clima Köppen no território do Brasil. A classificação climática de Köppen [129] é o sistema de classificação global dos tipos climáticos mais utilizada em geografia, climatologia e ecologia.

A primeira ontologia, representando o domínio D_1, modela o domínio referente à divisão territorial do Brasil e possui três níveis. O nó raiz representa o conceito "Brasil", os nós descendentes representam conceitos referentes às regiões brasileiras e cada conceito de região possui os conceitos referentes aos estados constituintes como descendentes. A Fig. A.2 ilustra a ontologia *lightweight* que modela a divisão territorial do Brasil.

A segunda ontologia, representando o domínio D_2, modela o domínio referente à classificação climática de Köppen que atua sobre o Brasil e possui três níveis. O nó raiz representa o conceito "Clima", os nós descendentes representam conceitos referentes aos climas zonais existentes no Brasil e cada conceito de clima zonal possui os conceitos constituintes referentes à classificação Köppen como descendentes. A definição de climas zonais foi baseada na classificação de climas dada pelo IBGE (Instituto Brasileiro de Geografia e Estatística) [60]. A Fig. A.3 ilustra a ontologia *lightweight* que modela o domínio de clima no Brasil.

Para a realização dos experimentos de recuperação de informação cada uma das ontologias foi

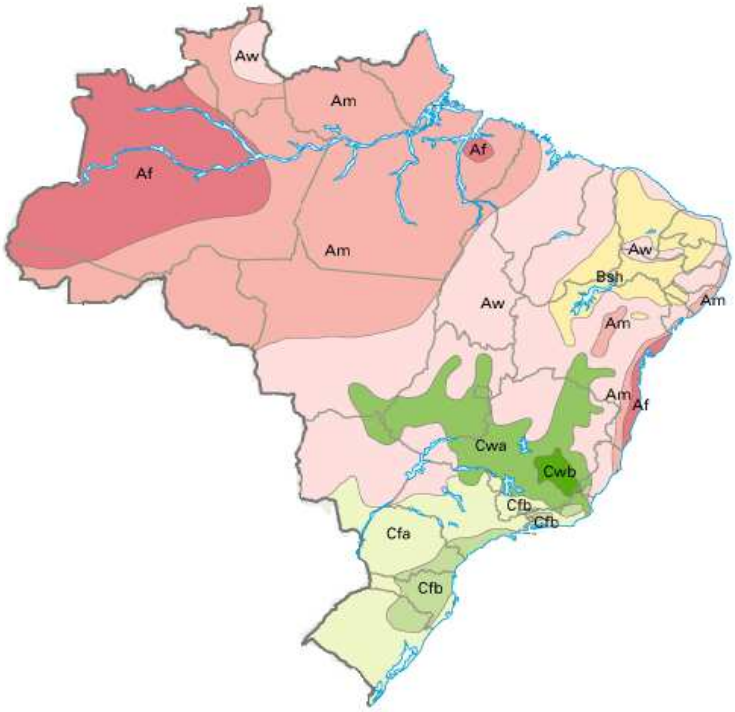

Fig. A.1: Mapa do Brasil com a distribuição climática de köppen no país. Fonte: [106].

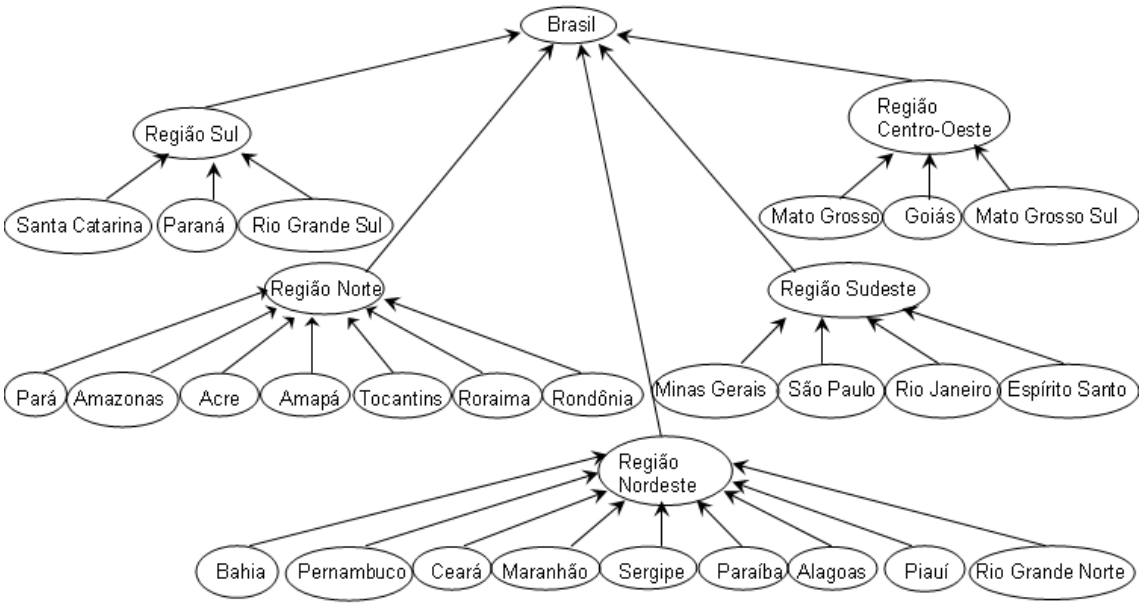

Fig. A.2: Ontologia *lightweight* com conceitos relativos à divisão territorial do Brasil.

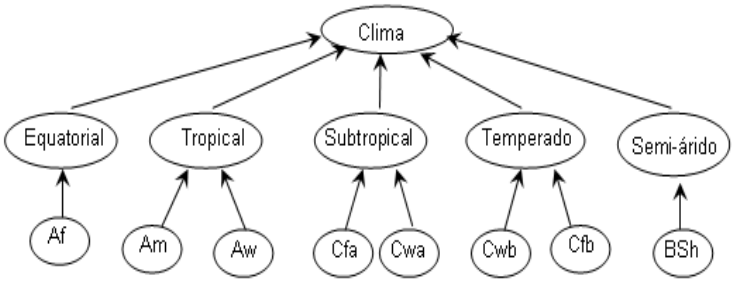

Fig. A.3: Ontologia *lightweight* com conceitos relativos à classificação climática de Köppen no Brasil.

considerada na forma *crisp* e *fuzzy*. Na forma *crisp* é considerada apenas a existência (valor 1) ou ausência (valor 0) das relações de especialização e generalização entre os conceitos das ontologias. Na forma *fuzzy* foi calculado um peso para cada relação em função da distribuição espacial das entidades representadas pelos conceitos.

Para a ontologia *lightweight* de divisão territorial do Brasil, domínio D_1, a associação de generalização *fuzzy* e a associação de especialização *fuzzy* denotam a relação espacial entre as entidades territoriais descritas pelos conceitos. Esta relação espacial é dada pela distribuição geográfica baseada na extensão territorial de cada entidade. A Tab. A.1 mostra as extensões territoriais de cada uma das entidades.

Nome da Entidade	**Área** $(\mathbf{km^2})$	**Nome da Entidade**	**Área** $(\mathbf{km^2})$
Brasil	8.516.090,74	Mato Grosso do Sul	35.816,00
Região Norte	3.854.127,03	Minas Gerais	588.528,29
Região Nordeste	1.556.141,48	Pará	1.247.689,52
Região Centro-Oeste	1.602.805,50	Paraíba	56.584,60
Região Sudeste	926.511,29	Paraná	199.314,00
Região Sul	576.505,44	Pernambuco	98.937,80
Acre	152.581,40	Piauí	252.378,00
Alagoas	27.767,00	Rio de Janeiro	43.696,05
Amapá	142.814,59	Rio Grande do Norte	52.796,79
Amazonas	1.570.745,68	Rio Grande do Sul	281.748,54
Bahia	567.295,30	Rondônia	237.576,17
Ceará	146.348,30	Roraima	224.298,98
Espírito Santo	46.077,52	Santa Catarina	95.442,90
Goiás	341.289,50	São Paulo	248.209,43
Maranhão	331.983,29	Sergipe	22.050,40
Mato Grosso	903.357,00	Tocantins	278.420,70

Tab. A.1: Extensão territorial das entidades geográficas no Brasil. Fonte: [128].

A Tab. A.2 mostra a relação de especialização *fuzzy* entre as entidades da ontologia de divisão territorial. Por exemplo, como a extensão total do território brasileiro é 8.516.090,74 km^2 e a extensão da Região Norte é de 3.854.127,03 km^2 então o valor da associação de especialização *fuzzy* $R_1^S\,(\text{Região Norte},\ \text{Brasil}) = 0,45$. A associação de especialização *fuzzy* indica que o conceito Região Norte especializa o conceito Brasil com valor igual a 0,45. Este valor indica que a "Região Norte" ocupa 45% do "Brasil". Em termos da recuperação de informação, se uma consulta consistir do conceito "Brasil" então o conceito "Região Norte" está associado ao conceito "Brasil" com um peso de $0,45$. Como a relação de generalização *fuzzy* é o inverso da relação de especialização *fuzzy* então $R_1^G\,(\text{Brasil},\ \text{Região Norte}) = 0,45$ indicando que o conceito Brasil generaliza o conceito Região Norte com valor igual a 0,45.

A Fig. A.4 ilustra as associações de especialização e generalização *fuzzy* na ontologia *lightweight* de divisão territorial do Brasil. No caso de se considerar a ontologia *lightweight* de divisão territorial do Brasil na forma *crisp* os pesos entre os conceitos assumem o valor 1.0.

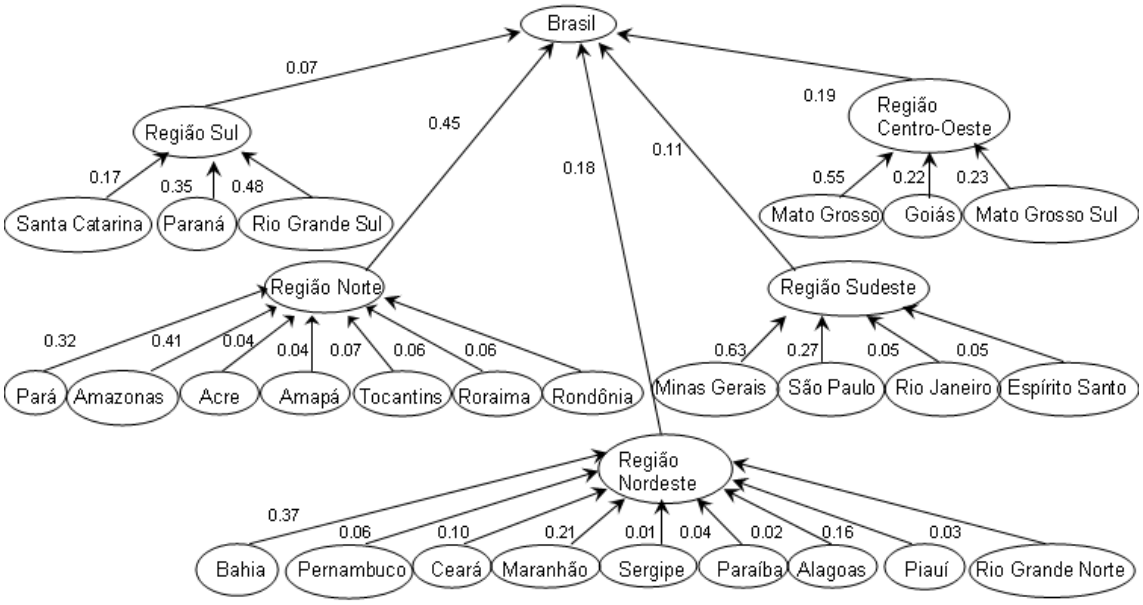

Fig. A.4: Ontologia *lightweight* com os pesos das associações de especialização e generalização *fuzzy* entre os conceitos de divisão territorial do Brasil.

Para a ontologia *lightweight* de clima do Brasil, domínio D_2, a associação de generalização *fuzzy* e a associação de especialização *fuzzy* denotam a relação espacial entre as entidades de climas descritas pelos conceitos. Esta relação espacial é dada pela distribuição climática baseada na extensão de cada entidade climática no território do Brasil. Para calcular a extensão de cada entidade o mapa da Fig. A.1 foi escaneado para obter a quantidade de pixels referente a cada uma das entidades. A Tab. A.3 mostra as extensões de cada uma das entidades de clima medidas em pixels.

Entidade Específica	Entidade Geral	Valor da relação
Região Norte	Brasil	0,45
Região Nordeste	Brasil	0,18
Região Centro-Oeste	Brasil	0,19
Região Sudeste	Brasil	0,11
Região Sul	Brasil	0,07
Acre	Região Norte	0,04
Amapá	Região Norte	0,04
Amazonas	Região Norte	0,41
Pará	Região Norte	0,32
Rondônia	Região Norte	0,06
Roraima	Região Norte	0,06
Tocantins	Região Norte	0,07
Alagoas	Região Nordeste	0,02
Bahia	Região Nordeste	0,36
Ceará	Região Nordeste	0,09
Maranhão	Região Nordeste	0,21
Pernambuco	Região Nordeste	0,06
Piauí	Região Nordeste	0,16
Paraíba	Região Nordeste	0,04
Rio Grande do Norte	Região Nordeste	0,03
Sergipe	Região Nordeste	0,01
Goiás	Região Centro-Oeste	0,22
Mato Grosso	Região Centro-Oeste	0,55
Mato Grosso do Sul	Região Centro-Oeste	0,23
Espírito Santo	Região Sudeste	0,05
Minas Gerais	Região Sudeste	0,63
Rio de Janeiro	Região Sudeste	0,05
São Paulo	Região Sudeste	0,27
Paraná	Região Sul	0,35
Rio Grande do Sul	Região Sul	0,48
Santa Catarina	Região Sul	0,17

Tab. A.2: Relação de especialização *fuzzy* para o domínio de divisão territorial.

Nome da Entidade	Área mapa (pixels)	Nome da Entidade	Área mapa (pixels)
Clima	93369	Am	33833
Equatorial	13486	Aw	25978
Tropical	59811	Cfa	7243
Subtropical	13509	Cwa	6266
Temperado	2249	Cwb	468
Semi-árido	4314	Cfb	1781
Af	13486	BSh	4314

Tab. A.3: Área das entidades de clima no território brasileiro.

A Tab. A.4 mostra a relação de especialização *fuzzy* entre as entidades da ontologia do domínio de clima. Por exemplo, como a extensão total do clima tropical é de 59811 pixels e a extensão do clima de Köppen Am é de 33833 pixels então o valor da associação de especialização *fuzzy* $R_2^S\,(\text{Am},\ \text{Tropical}) = 0.57$. A associação de especialização *fuzzy* indica que o conceito Am especializa o conceito Tropical com valor igual a 0,57. Como a relação de generalização *fuzzy* é o inverso da relação de especialização *fuzzy* então $R_2^G\,(\text{Tropical},\ \text{Am}) = 0.57$ indicando que o conceito Tropical generaliza o conceito Am com valor igual a 0,57.

Entidade Específica	Entidade Geral	Valor da relação
Equatorial	Clima	0,14
Tropical	Clima	0,64
Subtropical	Clima	0,14
Temperado	Clima	0,02
Semi-árido	Clima	0,05
Af	Equatorial	1,0
Am	Tropical	0,57
Aw	Tropical	0,43
Cfa	Subtropical	0,54
Cwa	Subtropical	0,46
Cwb	Temperado	0,21
Cfb	Temperado	0,79
BSh	Semi-árido	1,0

Tab. A.4: Relação de especialização *fuzzy* para o domínio de clima no Brasil.

A Fig. A.5 ilustra as associações de especialização e generalização *fuzzy* na ontologia *lightweight* de clima do Brasil. No caso de se considerar a ontologia *lightweight* de clima do Brasil na forma *crisp* os pesos entre os conceitos assumem o valor 1.0.

O relacionamento entre as ontologias é dado pela distribuição do clima no território brasileiro

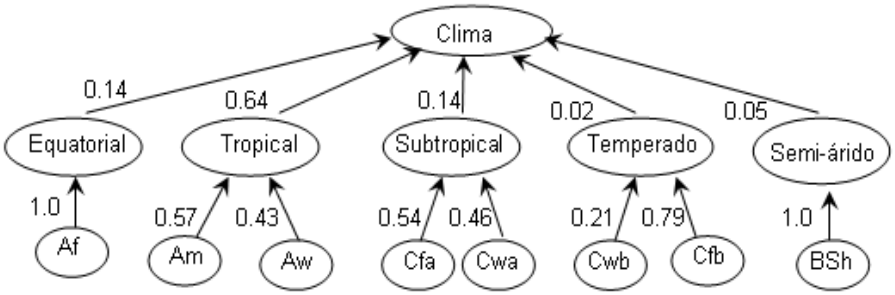

Fig. A.5: Ontologia *lightweight* com os pesos das associações de especialização e generalização *fuzzy* entre os com conceitos relativos ao domínio de clima no Brasil.

como observado no mapa da Fig. A.1. O relacionamento entre as ontologias é representado pela associação positiva *fuzzy*. O relacionamento entre os conceitos da ontologia *lightweight* de divisão territorial do Brasil, domínio D_1, e a ontologia *lightweight* de clima do Brasil, domínio D_2, é representado pela relação positiva *fuzzy* R_{12}^P. O relacionamento entre os conceitos da ontologia *lightweight* de clima do Brasil, domínio D_2, e a ontologia *lightweight* de divisão territorial do Brasil, domínio D_1, é representado pela relação positiva *fuzzy* R_{21}^P.

Em cada uma das relações positivas *fuzzy* o relacionamento é estabelecido em dois níveis. O primeiro nível ocorre entre os conceitos que representam as regiões do Brasil e os conceitos que representam os climas zonais. O segundo nível ocorre entre os conceitos que representam os estados brasileiros e os conceitos que representam os climas da classificação de Köppen. Na Fig. A.6 as linhas tracejadas ilustram os dois níveis de relacionamento entre as ontologias.

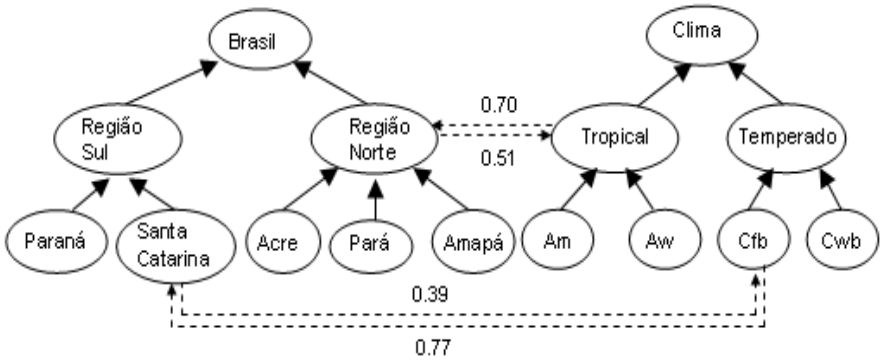

Fig. A.6: Associações positivas entre as ontologias de divisão territorial e clima.

Para calcular os dois níveis de relacionamentos o mapa foi escaneado e o valor da distribuição de cada tipo de clima Köppen nos estados do Brasil foram registrados. A Tab. A.5 mostra os valores coletados para os climas da classificação de Köppen para os estados das regiões Norte e Nordeste do Brasil e a Tab. A.6 mostra os valores coletados para os estados das regiões Centro-Oeste, Sul e

Sudeste.

O primeiro nível da associação positiva *fuzzy* ocorre entre os conceitos que representam as regiões brasileiras e os conceitos que representam os climas zonais. O valor desta relação é dada pela distribuição do clima zonal em cada região. A Tab. A.7 mostra a área total das entidades territoriais compiladas a partir das Tabs. A.5 e A.6, medidas em pixels. A Tab. A.8 mostra a área que os climas zonais ocupam em cada uma das regiões brasileiras.

Com os dados das Tabs. A.7 e A.8 é possível calcular o valor do primeiro nível da associação positiva *fuzzy*. Por exemplo, a área total do clima Tropical no Brasil é de 59.811 pixels. A área total de clima Tropical na Região Norte é 30.616 pixels. Assim o valor da associação positiva *fuzzy* entre a Região Norte e o clima tropical é $R_{12}^{P}\left(\text{Região Norte}, \text{ Tropical}\right) = 0.51$. Isto significa que o conceito Região Norte implica o conceito Tropical pela associação positiva *fuzzy* com o valor 0.51. A Tab. A.9 mostra os valores da associação positiva *fuzzy* entre os conceitos de região do domínio de divisão territorial e os conceitos de clima zonal do domínio de clima.

Os dados das Tabs. A.7 e A.8 também são utilizados para calcular o valor da associação positiva *fuzzy* entre os conceitos de clima zonal do domínio de clima e os conceitos de região do domínio de divisão territorial. Por exemplo a área total da Região Norte é de 43.795 pixels e a área de clima Tropical na Região Norte é de 30.616 pixels. Assim o valor da associação positiva *fuzzy* entre o clima Tropical e a Região Norte é $R_{21}^{P}\left(\text{Tropical}, \text{ Região Norte}\right) = 0.70$. Isto significa que o conceito Tropical implica no conceito Região Norte pela associação positiva *fuzzy* com o valor 0.70. A Tab. A.10 mostra os valores da associação positiva *fuzzy* entre os conceitos de clima zonal do domínio de clima e os conceitos de região do domínio de divisão territorial.

O segundo nível da associação positiva *fuzzy* ocorre entre os conceitos que representam os estados brasileiros e os conceitos que representam os climas da classificação de Köppen. O valor desta relação é dada pela distribuição do clima da classificação de Köppen em cada estado.

As Tabs. A.5 e A.6 mostram a área que os climas da classificação de Köppen ocupam em cada um dos estados brasileiros e Tab. A.7 mostra a área total das entidades territoriais compiladas a partir das Tabs. A.5 e A.6, medidas em pixels. Por exemplo o clima Cfb ocorre nos estados Rio Grande do Sul, Santa Catarina, Paraná, São Paulo e Espírito Santo. A extensão total do clima Köppen Cfb no Brasil é de 1.781 pixels. A extensão do clima Köppen Cfb no estado Santa Catarina é de 693 pixels. Assim a associação positiva *fuzzy* entre o estado de Santa Catarina e o clima Köppen Cfb é $R_{12}^{P}\left(\text{Santa Catarina}, \text{ Cfb}\right) = 0.39$ denotando que o estado de Santa Catarina é associado ao clima Köppen Cfb pelo valor de 0.39. Isto significa que o conceito Santa Catarina implica no conceito Cfb pela associação positiva *fuzzy* com o valor 0.39. A Tab. A.11 mostra os valores da associação positiva *fuzzy* entre os conceitos de estado, do domínio de divisão territorial, e os conceitos de clima Köppen, do domínio de clima.

Região	Estado	Clima Zonal	Clima Köppen	Área (pixels)
Norte	Acre	Equatorial	Af	509
Norte	Acre	Tropical	Am	1095
Norte	Amapá	Tropical	Am	1508
Norte	Amazonas	Tropical	Am	5867
Norte	Amazonas	Equatorial	Af	12447
Norte	Pará	Tropical	Am	14093
Norte	Pará	Equatorial	Af	197
Norte	Rondônia	Tropical	Am	2745
Norte	Roraima	Tropical	Am	1515
Norte	Roraima	Equatorial	Af	26
Norte	Roraima	Tropical	Aw	817
Norte	Tocantins	Tropical	Am	132
Norte	Tocantins	Tropical	Aw	2844
Nordeste	Alagoas	Tropical	Aw	58
Nordeste	Alagoas	Semi-árido	BSh	68
Nordeste	Alagoas	Tropical	Am	94
Nordeste	Bahia	Equatorial	Af	307
Nordeste	Bahia	Tropical	Am	558
Nordeste	Bahia	Tropical	Aw	3483
Nordeste	Bahia	Semi-árido	BSh	1143
Nordeste	Bahia	Subtropical	Cwa	304
Nordeste	Ceará	Tropical	Aw	397
Nordeste	Ceará	Semi-árido	BSh	1146
Nordeste	Maranhão	Tropical	Am	799
Nordeste	Maranhão	Tropical	Aw	2787
Nordeste	Paraíba	Tropical	Am	14
Nordeste	Paraíba	Tropical	Aw	173
Nordeste	Paraíba	Semi-árido	BSh	284
Nordeste	Pernambuco	Tropical	Am	34
Nordeste	Pernambuco	Tropical	Aw	380
Nordeste	Pernambuco	Semi-árido	BSh	505
Nordeste	Piauí	Tropical	Aw	2051
Nordeste	Piauí	Semi-árido	BSh	659
Nordeste	Rio Grande Norte	Tropical	Aw	97
Nordeste	Rio Grande Norte	Semi-árido	BSh	441
Nordeste	Sergipe	Tropical	Aw	115
Nordeste	Sergipe	Tropical	Am	11
Nordeste	Sergipe	Semi-árido	BSh	68

Tab. A.5: Área de clima Köppen nos estados das Regiões Norte e Nordeste do Brasil.

Região	Estado	Clima Zonal	Clima Köppen	Área (pixels)
Centro-oeste	Goiás	Tropical	Aw	1775
Centro-oeste	Goiás	Subtropical	Cwa	1985
Centro-oeste	Mato Grosso	Tropical	Am	5306
Centro-oeste	Mato Grosso	Tropical	Aw	4805
Centro-oeste	Mato Grosso	Subtropical	Cwa	348
Centro-oeste	Mato Grosso do Sul	Tropical	Aw	3228
Centro-oeste	Mato Grosso do Sul	Subtropical	Cwa	211
Centro-oeste	Mato Grosso do Sul	Subtropical	Cfa	550
Sudeste	Espírito Santo	Temperado	Cfb	12
Sudeste	Espírito Santo	Subtropical	Cwa	30
Sudeste	Espírito Santo	Tropical	Aw	311
Sudeste	Espírito Santo	Tropical	Am	62
Sudeste	Minas Gerais	Tropical	Aw	2118
Sudeste	Minas Gerais	Subtropical	Cfa	359
Sudeste	Minas Gerais	Subtropical	Cwa	3186
Sudeste	Minas Gerais	Temperado	Cwb	468
Sudeste	Minas Gerais	Temperado	Cfb	11
Sudeste	Rio de Janeiro	Temperado	Cfb	113
Sudeste	Rio de Janeiro	Subtropical	Cfa	175
Sudeste	Rio de Janeiro	Subtropical	Cwa	85
Sudeste	São Paulo	Tropical	Aw	539
Sudeste	São Paulo	Subtropical	Cfa	1638
Sudeste	São Paulo	Subtropical	Cwa	117
Sudeste	São Paulo	Temperado	Cfb	257
Sul	Paraná	Subtropical	Cfa	1868
Sul	Paraná	Temperado	Cfb	295
Sul	Rio Grande do Sul	Subtropical	Cfa	2444
Sul	Rio Grande do Sul	Temperado	Cfb	400
Sul	Santa Catarina	Subtropical	Cfa	209
Sul	Santa Catarina	Temperado	Cfb	693

Tab. A.6: Área de clima Köppen nos estados das Regiões Centro-oeste, Sudeste e Sul do Brasil.

Nome da Entidade	Área (pixels)	Nome da Entidade	Área (pixels)
Brasil	93369	Mato Grosso do Sul	3989
Região Norte	43795	Minas Gerais	6142
Região Nordeste	15976	Pará	14290
Região Centro-Oeste	18208	Paraíba	471
Região Sudeste	9481	Paraná	2163
Região Sul	5909	Pernambuco	919
Acre	1604	Piauí	2710
Alagoas	220	Rio de Janeiro	373
Amapá	1508	Rio Grande do Norte	538
Amazonas	18314	Rio Grande do Sul	2844
Bahia	5795	Rondônia	2745
Ceará	1543	Roraima	2358
Espírito Santo	415	Santa Catarina	902
Goiás	3760	São Paulo	2551
Maranhão	3586	Sergipe	194
Mato Grosso	10459	Tocantins	2976

Tab. A.7: Área das entidades geográficas no domínio de divisão territorial do Brasil.

Região	Clima zonal	Área (pixels)
Centro-oeste	Tropical	15114
Centro-oeste	Subtropical	3094
Nordeste	Tropical	11051
Nordeste	Semi-árido	4314
Nordeste	Equatorial	307
Nordeste	Subtropical	304
Norte	Equatorial	13179
Norte	Tropical	30616
Sudeste	Tropical	3030
Sudeste	Temperado	861
Sudeste	Subtropical	5590
Sul	Temperado	1388
Sul	Subtropical	4521

Tab. A.8: Área de clima zonal nas regiões do Brasil.

Região	Clima zonal	Valor da relação
Região Centro-Oeste	Subtropical	0.23
Região Centro-Oeste	Tropical	0.25
Região Nordeste	Equatorial	0.02
Região Nordeste	Semi-árido	1.00
Região Nordeste	Subtropical	0.02
Região Nordeste	Tropical	0.18
Região Norte	Equatorial	0.98
Região Norte	Tropical	0.51
Região Sudeste	Subtropical	0.41
Região Sudeste	Temperado	0.38
Região Sudeste	Tropical	0.05
Região Sul	Subtropical	0.33
Região Sul	Temperado	0.62

Tab. A.9: Valor da associação positiva *fuzzy* entre os conceitos de região, no domínio de divisão territorial, e os conceitos de clima zonal no domínio de clima.

Os dados das Tabs. A.5, A.6 e A.7 também são utilizados para calcular o valor da associação positiva *fuzzy* entre os conceitos de clima da classificação de Köppen do domínio de clima e os conceitos de estado do domínio de divisão territorial. Por exemplo a área total do estado de Santa Catarina é de 900 pixels e a área de clima Köppen Cfb em Santa Catarina é de 693 pixels. Assim o valor da associação positiva *fuzzy* entre o clima Köppen Cfb e Santa Catarina é $R_{21}^{P}\,(\text{Cfb},\ \text{Santa Catarina}) = 0.77$. Isto significa que o conceito Cfb implica no conceito Santa Catarina pela associação positiva *fuzzy* com o valor 0.77.

A Tab. A.12 mostra os valores da associação positiva *fuzzy* entre os conceitos de clima Köppen do domínio de clima e os conceitos de estado do domínio de divisão territorial.

A.2 Coleção de Documentos

A coleção de documentos é composta de uma amostra de metadados de 129 documentos selecionados da base de documentos do domínio de agrometeorologia. Esta base é mantida pela Embrapa [33] e contém aproximadamente 17.000 documentos. Esta amostra possui os metadados dos documentos quem contêm apenas um ou uma combinação de conceitos das ontologias. O arquivo contendo a amostra com os metadados dos documentos, no formato XML, pode ser encontrado na página da tese[1] .

[1] `http://www.dca.fee.unicamp.br/~ricarte/MORFuzz/`

Clima Zonal	Região	Valor da relação
Equatorial	Região Nordeste	0.02
Equatorial	Região Norte	0.30
Semi-árido	Região Nordeste	0.27
Subtropical	Região Centro-Oeste	0.17
Subtropical	Região Nordeste	0.02
Subtropical	Região Sudeste	0.59
Subtropical	Região Sul	0.77
Temperado	Região Sudeste	0.09
Temperado	Região Sul	0.23
Tropical	Região Centro-Oeste	0.83
Tropical	Região Nordeste	0.69
Tropical	Região Norte	0.70
Tropical	Região Sudeste	0.32

Tab. A.10: Valor da associação positiva *fuzzy* entre os conceitos de clima zonal, no domínio de clima, e os conceitos de região no domínio de divisão territorial.

O arquivo XML, com os metadados dos documentos, é percorrido pela ferramenta SAX (Simple API for XML) [111] para construir as relações que associam os documentos aos conceitos de cada domínio. Os metadados referentes à identificação dos documentos (campo <docid>) e os seus resumos (campo <conteudo>) são considerados para coletar as informações para indexação. A máquina de busca Lucene do projeto Apache foi utilizada para criar um índice para armazenar as informações sobre os termos e os documentos. Para cada conceito das ontologias o índice criado pelo Lucene é pesquisado para adquirir as informações necessárias para o cálculo da medida de *tf-idf*, descrita na seção 4.2.2, para todos os documentos em relação a este conceito.

A amostra constituída por 129 documentos inclui todos os conceitos das ontologias construídas e acredita-se que seja o mínimo necessário para provar o potencial da forma de organização de conhecimento e do método de expansão de consulta propostos nesta tese. Além disto este tamanho de amostra viabiliza a associação de documentos relevantes à cada consulta por um especialista de domínio.

Para gerar a amostra da coleção de documentos uma pesquisa por cada conceito c_{ky}, do domínio D_k, é realizada na coleção de agrometeorologia. O conjunto de documentos retornados DOC_{ky}, $1 \leq k \leq K$, K é o número de domínios e $1 \leq y \leq |D_k|$ é armazenado. Para o experimento $K = 2$, $|D_1| = 32$ e $|D_2| = 14$. A amostra de documentos foi montada refletindo a proporção do número de documentos existentes para cada conceito. A contribuição $Con_{ky} \in [0, 1]$ que conceito c_{ky} representa sobre o número total de documentos retornados, considerando todos os conceitos, é dada pela Eq. A.1.

Estado	Köppen	Valor	Estado	Köppen	Valor
Acre	Af	0.04	Minas Gerais	Cwb	1.00
Acre	Am	0.03	Pará	Af	0.01
Alagoas	Bsh	0.02	Pará	Am	0.42
Amapá	Am	0.04	Paraíba	Aw	0.01
Amazonas	Af	0.92	Paraíba	Bsh	0.07
Amazonas	Am	0.17	Paraná	Cfa	0.26
Bahia	Af	0.02	Paraná	Cfb	0.17
Bahia	Am	0.02	Pernambuco	Aw	0.01
Bahia	Aw	0.13	Pernambuco	Bsh	0.12
Bahia	Bsh	0.26	Piauí	Aw	0.08
Bahia	Cwa	0.05	Piauí	Bsh	0.15
Ceará	Aw	0.02	Rio de Janeiro	Cfa	0.02
Ceará	Bsh	0.27	Rio de Janeiro	Cfb	0.06
Espírito Santo	Aw	0.01	Rio de Janeiro	Cwa	0.01
Espírito Santo	Cfb	0.01	Rio Grande do Norte	Bsh	0.10
Goiás	Aw	0.07	Rio Grande do Sul	Cfa	0.34
Goiás	Cwa	0.32	Rio Grande do Sul	Cfb	0.22
Maranhão	Am	0.02	Rondônia	Am	0.08
Maranhão	Aw	0.11	Roraima	Am	0.04
Mato Grosso	Am	0.16	Roraima	Aw	0.03
Mato Grosso	Aw	0.18	Santa Catarina	Cfa	0.03
Mato Grosso	Cwa	0.06	Santa Catarina	Cfb	0.39
Mato Grosso do Sul	Aw	0.12	São Paulo	Aw	0.02
Mato Grosso do Sul	Cfa	0.08	São Paulo	Cfa	0.23
Mato Grosso do Sul	Cwa	0.03	São Paulo	Cfb	0.14
Minas Gerais	Aw	0.08	São Paulo	Cwa	0.02
Minas Gerais	Cfa	0.05	Sergipe	Bsh	0.02
Minas Gerais	Cfb	0.01	Tocantins	Aw	0.11
Minas Gerais	Cwa	0.51			

Tab. A.11: Valor da associação positiva *fuzzy* entre os conceitos de estado, no domínio de divisão territorial, e os conceitos de clima Köppen no domínio de clima.

Köppen	Estado	Valor	Köppen	Estado	Valor
Af	Acre	0.32	Aw	São Paulo	0.21
Af	Amazonas	0.68	Aw	Sergipe	0.59
Af	Bahia	0.05	Aw	Tocantins	0.96
Af	Pará	0.01	Bsh	Alagoas	0.31
Af	Roraima	0.01	Bsh	Bahia	0.21
Am	Acre	0.68	Bsh	Ceará	0.74
Am	Alagoas	0.43	Bsh	Paraíba	0.60
Am	Amapá	1.00	Bsh	Pernambuco	0.55
Am	Amazonas	0.32	Bsh	Piauí	0.24
Am	Bahia	0.10	Bsh	Rio Grande do Norte	0.82
Am	Espírito Santo	0.15	Bsh	Sergipe	0.35
Am	Maranhão	0.22	Cfa	Mato Grosso do Sul	0.14
Am	Mato Grosso	0.51	Cfa	Minas Gerais	0.06
Am	Pará	0.99	Cfa	Paraná	0.86
Am	Paraíba	0.03	Cfa	Rio de Janeiro	0.47
Am	Pernambuco	0.04	Cfa	Rio Grande do Sul	0.86
Am	Rondônia	1.00	Cfa	Santa Catarina	0.23
Am	Roraima	0.64	Cfa	São Paulo	0.64
Am	Sergipe	0.06	Cfb	Espírito Santo	0.03
Am	Tocantins	0.04	Cfb	Paraná	0.14
Aw	Alagoas	0.26	Cfb	Rio de Janeiro	0.30
Aw	Bahia	0.63	Cfb	Rio Grande do Sul	0.14
Aw	Ceará	0.26	Cfb	Santa Catarina	0.77
Aw	Espírito Santo	0.75	Cfb	São Paulo	0.10
Aw	Goiás	0.47	Cwa	Bahia	0.06
Aw	Maranhão	0.78	Cwa	Espírito Santo	0.07
Aw	Mato Grosso	0.46	Cwa	Goiás	0.53
Aw	Mato Grosso do Sul	0.81	Cwa	Mato Grosso	0.03
Aw	Minas Gerais	0.35	Cwa	Mato Grosso do Sul	0.05
Aw	Paraíba	0.37	Cwa	Minas Gerais	0.52
Aw	Pernambuco	0.41	Cwa	Rio de Janeiro	0.23
Aw	Piauí	0.76	Cwa	São Paulo	0.05
Aw	Rio Grande do Norte	0.18	Cwb	Minas Gerais	0.08
Aw	Roraima	0.35			

Tab. A.12: Valor da associação positiva *fuzzy* entre os conceitos de clima Köppen, no domínio de clima, e os conceitos de estado no domínio de divisão territorial.

$$Con_{ky} = \frac{|DOC_{ky}|}{\sum_{k=1}^{\mathrm{K}} \sum_{y=1}^{|\mathrm{D_k}|} |DOC_{ky}|} \tag{A.1}$$

O número Num_{ky} de documentos associados ao conceito c_{ky} necessários para compor a amostra de 129 documentos é dado pelo produto aritmético da Eq. A.2. O número Num_{ky} de documentos associados ao conceito c_{ky} é retirado do topo da lista de documentos no conjunto DOC_{ky}.

$$Num_{ky} = 129 \cdot Con_{ky} \tag{A.2}$$

A.3 Preparação das Consultas

O conjunto de consultas contém 83 consultas e é composto por consultas contendo apenas um conceito de cada ontologia assim como consultas contendo combinações de dois conceitos de ambas as ontologias nos diferentes níveis. Os conceitos são conectados com operadores AND ou OR. Para cada uma das consultas a amostra de 129 documentos é analisada e os documentos relevantes são selecionados por um especialista do domínio. Para associar os documentos relevantes para cada consulta cada documento na amostra é examinado, pelo especialista, e é atribuído aos conceitos das ontologias criando um novo conjunto de documentos $DOCN_{ky}$ a ser associado ao conceito c_{ky}. As atribuições consideraram as seguintes diretivas:

- Considerando ambas as ontologias um documento relacionado a um conceito mais específico é atribuído ao seu conceito mais geral. Neste caso documentos relacionados aos conceitos de estado são associados aos conceitos de regiões correspondentes e os documentos relacionados a conceitos de clima köppen são associados aos conceitos de clima zonal correspondentes. Por exemplo: um documento contendo o conceito Santa Catarina é associado ao conceito Região Sul e um documento contendo o conceito Cfb e associado ao conceito Temperado.

- Considerando a ontologia de clima, um documento relacionado apenas ao conceito zonal é associado aos seus conceitos köppen correspondentes pois o conceito zonal representa as características mais gerais dos conceitos Köppen mais específicos. Por exemplo: documentos contendo apenas o conceito Tropical são associados aos conceitos Aw e Am.

- Um documento relacionado a um clima zonal e a um conceito territorial também é associado ao conceito Köppen específico relacionado ao conceito territorial (este conhecimento é obtido do mapa da Fig. A.1).Por exemplo: um documento contendo o conceito Temperado e o conceito Santa Catarina é associado ao conceito Cfb; um documento contendo o conceito Tropical e o

conceito Região Norte também é associado ao conceito Am uma vez que o conceito Am é o conceito Tropical específico que ocorre no território dado pelo conceito Região Norte.

- De acordo com o conhecimento do especialista do domínio e pela observação do mapa da Fig. A.1 outras associações são consideradas e refletidas na associação de documentos aos conceitos como por exemplo: o clima Köppen Bsh e o seu clima zonal correspondente Semi-árido ocorrem apenas em alguns estados da Região Nordeste. Então todos os documentos contendo apenas os conceitos Semi-árido ou Bsh são atribuídos ao conceito Região Nordeste e aos estados onde este tipo de clima ocorre. Por outro lado documentos que tratem de aspectos relacionados à seca em estados onde ocorre o clima Semi-árido são associados ao conceito Semi-árido e Bsh. O clima Köppen Cwb ocorre apenas no estado de Minas Gerais, da Região Sudeste, então os documentos contendo apenas o conceito Cwb são atribuídos aos conceitos Minas Gerais e Região Sudeste. Documentos contendo os conceitos Tropical ou Equatorial e relacionados à Região Amazônica são atribuídos aos conceitos dos estados incluídos nesta região e que possuem estes tipos de climas.

Uma vez que a nova atribuição de documentos aos conceitos é realizada os documentos relevantes para cada consulta são estabelecidos baseado em operações no conjunto de documentos $DOCN_{ky}$ em função da especificação da consulta. Os documentos relevantes para uma consulta contendo apenas um conceito c_{ky} é o próprio conjunto $DOCN_{ky}$. Se a consulta é do tipo $(c_{ki} \text{ and } c_{kj})$ onde $1 \leq k \leq 2$ e $1 \leq i, j \leq |D_k|$ então o conjunto de documentos relevantes para a consulta é dado por $\{DOCN_{ki} \cap DOCN_{kj}\}$. Se a consulta é do tipo $(c_{ki} \text{ or } c_{kj})$ então o conjunto de documentos relevantes é dado por $\{DOCN_{ki} \cup DOCN_{kj}\}$.

A atribuição de documentos relevantes para os conceitos da ontologia *lightweigh* referente ao domínio de clima é mostrada na Tab. A.13. Os números que representam os documentos são dados pelo metadado <docid>, no arquivo XML que contém os metadados dos documentos da amostra.

A atribuição de documentos relevantes para os conceitos da ontologia *lightweigh* referente ao domínio de divisão territorial do Brasil é mostrada na Tab. A.14.

Tab. A.14: Documentos relevantes para os conceitos da ontologia de divisão territorial do Brasil.

Divisão territorial	**Identificador dos documentos relevantes**
Amazonas	8, 9, 10, 11, 18, 22, 28, 36, 37
Acre	9, 10, 11, 18, 22, 28, 38
Amapá	9, 11, 18, 22, 28, 39
Roraima	9, 11, 18, 22, 28, 40

Continua na próxima página

Tab. A.14 –continuação da tabela anterior

Divisão territorial	Identificador dos documentos relevantes
Rondônia	11, 18, 22, 28, 41, 42
Tocantins	43
Pará	8, 9, 11, 18, 20, 21, 22, 24, 25, 28,44, 105, 123
Rio de Janeiro	45, 46, 123
Espírito Santo	25, 47, 48, 49, 123
Minas Gerais	2, 3, 5, 7, 19, 25, 50, 51, 52, 53, 111
São Paulo	1, 16, 25, 54, 55, 56, 57, 58, 59, 102, 123
Bahia	7, 25, 29, 30, 31, 32, 33, 34, 35, 60, 61, 62, 63, 64, 65, 106, 108, 110, 111, 112, 113, 114, 115, 116, 117,121
Pernambuco	0, 25, 29, 30 , 31, 32, 34, 35, 66, 67, 68, 106, 108, 111, 115, 121
Maranhão	7, 27, 69, 123
Ceará	25, 27, 29, 30, 31, 32, 34, 35, 70, 71, 72,73, 75, 106, 107, 108, 109, 111, 115
Rio Grande do Norte	25, 29, 30, 31, 32, 34, 35, 74, 75, 106, 107, 108, 111
Paraíba	25, 29, 30, 31, 32, 34, 35, 75, 76, 106, 108, 111, 113, 115
Sergipe	25, 29, 30, 31, 32, 34, 35, 65, 77, 78, 106, 111
Alagoas	29, 30, 31, 32, 34, 35, 79, 106,115, 111
Piauí	7, 27, 29, 30, 31, 32, 33, 34, 35, 80, 81, 82, 106, 107, 111, 115
Mato Grosso	6, 7, 11, 19, 28, 42, 83
Mato Grosso do Sul	6, 7, 19, 28 , 84
Goiás	7, 19, 28, 85, 86
Paraná	13, 87, 88, 89, 90, 123,124,125,127
Rio Grande do Sul	4, 14, 15, 91, 92, 93, 94, 118,119,120, 121, 122, 123, 125,126,127
Santa Catarina	95, 96, 121, 125, 127, 128
Região norte	8, 9, 10, 11, 12, 17, 18, 20, 21, 22, 24, 25, 28 ,36, 37, 38, 39, 40, 41, 42, 43, 44, 97, 105, 123
Região Nordeste	0, 7, 12, 17, 26, 27, 29, 30, 31, 32, 33, 34, 35, 60,61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76,77,78, 79, 80, 81, 82, 98, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
Região Sul	4, 12, 13, 14, 15, 17, 58, 87, 88, 89, 90, 91,92, 93, 94, 95, 96, 100, 101, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128

Continua na próxima página

Tab. A.14 –continuação da tabela anterior

Divisão territorial	Identificador dos documentos relevantes
Região Sudeste	1, 2, 3, 5, 7, 12, 16, 17, 19, 25, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 102, 111, 123
Região Centro-Oeste	6, 7, 11, 12, 19, 28, 42, 83, 84, 85, 86, 103, 104
Brasil	0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41,42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60,61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76,77,78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 100, 101, 105, 98, 99, 102, 103, 104, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 121, 123, 124, 125, 126, 127, 128

O conjunto de consultas utilizadas no experimento bem como a atribuição de documentos relevantes para cada uma delas é mostrada na Tab. A.15.

Tab. A.15: Tabela de documentos relevantes para cada uma das consultas.

No.	Consulta	Documentos relevantes
1	Pará and Amapá	9, 11, 18, 22, 28
2	Bahia and Pernambuco	25, 29, 30, 31, 32, 34, 35, 106, 108, 111, 115, 121
3	Amazonas and Acre	9, 10, 11, 18, 22, 28
4	Santa Satarina and Paraná	125, 127
5	Cwa and Cwb	16, 17, 50, 51, 52
6	Am and Aw	9, 17, 18, 19, 23, 25, 26, 28, 60, 62, 79, 83, 99
7	Bsh and Cfa	121, 123
8	Região Centro-Oeste and Região Norte	11, 12, 28, 42
9	Região Sudeste and Região Sul	12, 17, 58, 123
10	Região Norte and Região Nordeste	12, 17, 123
11	Equatorial and Tropical	8, 9, 11, 60, 61, 62, 64, 99
12	Temperado and Subtropical	1, 16, 17, 50, 51, 54, 55, 56, 58, 87, 121
13	Semi-árido and Tropical	27, 60, 62, 65, 66, 67, 68, 69, 70, 71, 79, 99, 108, 116, 121

Continua na próxima página

Tab. A.15 – continuação da tabela anterior

No.	Consulta	Documentos relevantes
14	Região Sul and Rio Grande do Sul	4, 14, 15, 91, 92, 93, 94, 118, 119, 120, 121, 122, 123, 125, 126, 127
15	Região Norte and Pará	8, 9, 11, 18, 20, 21, 22, 24, 25, 28, 44, 105, 123
16	Região Sul and São Paulo	58, 123
17	Região Nordeste and Minas Gerais	7, 111
18	Semi-árido and Bsh	0, 12, 27, 29, 30, 31, 32, 33, 34, 35, 60, 62, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
19	Subtropical and Cfa	1, 4, 16, 17, 50, 51, 54, 55, 56, 58, 87, 118, 119, 120, 121, 122, 123, 124
20	Temperado and Cwa	1, 16, 17, 50, 51, 52, 121
21	Tropical and Af	9, 11, 60, 61, 62, 64, 99
22	Cwb and Minas Gerais	2, 3, 25, 50, 51, 52
23	Pará and am	8, 9, 11, 18, 20, 21, 22, 24, 25, 28
24	Pernambuco and Bsh	0, 29, 30, 31, 32, 34, 35, 66, 67, 68, 108, 111, 115, 121
25	Mato Grosso and Tropical	6, 7, 11, 19, 28, 42, 83
26	Ceará and Semi-árido	27, 29, 30, 31, 32, 34, 35, 70, 71, 72, 73, 75, 106, 107, 108, 109, 111, 115
27	São Paulo and Subtropical	1, 16, 54, 55, 56, 58, 123
28	Amazonas and Equatorial	8, 9, 10, 11
29	Região Norte and Tropical	8, 9, 11, 17, 18, 20, 21, 22, 24, 25, 28, 36, 39, 41, 42, 43
30	Região Sudeste and Temperado	1, 2, 3, 7, 16, 17, 25, 50, 51, 52, 54, 55, 56, 58
31	Região Nordeste and Semi-árido	0, 12, 27, 29, 30, 31, 32, 33, 34, 35, 60, 62, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123

Continua na próxima página

Tab. A.15 – continuação da tabela anterior

No.	Consulta	Documentos relevantes
32	Região Norte and Am	8, 9, 11, 17, 18, 20, 21, 22, 24, 25, 28, 36, 39, 41, 42, 43
33	Região Sudeste and Cwa	1, 5, 16, 17, 50, 51, 52, 123
34	Região Sul and Cfa	4, 17, 58, 87, 118, 119, 120, 121, 122, 123, 124
35	Região Centro-Oeste and Aw	6, 7, 19, 28, 83
36	Região Sudeste and Cwb	2, 3, 16, 17, 25, 50, 51, 52
37	Pará or Amapá	8, 9, 11, 18, 20, 21, 22, 24, 25, 28, 39, 44, 105, 123
38	Bahia or Pernambuco	0, 7, 25, 29, 30, 31, 32, 33, 34, 35, 60, 61, 62, 63, 64, 65, 66, 67, 68, 106, 108, 110, 111, 112, 113, 114, 115, 116, 117, 121
39	Amazonas or Acre	8, 9, 10, 11, 18, 22, 28, 36, 37, 38
40	Santa Catarina or Paraná	13, 87, 88, 89, 90, 95, 96, 121, 123, 124, 125, 127, 128
41	Cwa or Cwb	1, 2, 3, 5, 13, 14, 15, 16, 17, 25, 50, 51, 52, 60, 62, 86, 99, 121, 123
42	Am or Aw	6, 7, 8, 9, 11, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 36, 39, 41, 42, 43, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 79, 80, 81, 82, 83, 99, 116
43	Bsh or Cfa	0, 1, 4, 12, 16, 17, 27, 29, 30, 31, 32, 33, 34, 35, 50, 51, 54, 55, 56, 58, 60, 62, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 87, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124
44	Região Centro-Oeste or Região Sudeste	1, 2, 3, 5, 6, 7, 11, 12, 16, 17, 19, 25, 26, 28, 42, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 83, 84, 85, 86, 102, 103, 104, 111, 123

Continua na próxima página

Tab. A.15 – continuação da tabela anterior

No.	Consulta	Documentos relevantes
45	Região Sudeste or Região Sul	1, 2, 3, 4, 5, 7, 12, 13, 14, 15, 16, 17, 19, 25, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 100, 101, 102, 111, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128
46	Região Norte or Região Nordeste	0, 7, 8, 9, 10, 11, 12, 17, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 97, 98, 99, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
47	Região Sul or Rio Grande do Sul	4, 12, 13, 14, 15, 17, 58, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 100, 101, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128
48	Região Norte or Pará	8, 9, 10, 11, 12, 17, 18, 20, 21, 22, 24, 25, 28, 36, 37, 38, 39, 40, 41, 42, 43, 44, 97, 105, 123
49	Região Sul or São Paulo	1, 4, 12, 13, 14, 15, 16, 17, 25, 54, 55, 56, 57, 58, 59, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 100, 101, 102, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128
50	Região Centro-Oeste or Minas Gerais	2, 3, 5, 6, 7, 11, 12, 19, 25, 28, 42, 50, 51, 52, 53, 83, 84, 85, 86, 103, 104, 111
51	Equatorial or Tropical	3, 6, 7, 8, 9, 10, 11, 12, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 36, 39, 41, 42, 43, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 79, 80, 81, 82, 83, 99, 108, 116, 121

Continua na próxima página

Tab. A.15 – continuação da tabela anterior

No.	Consulta	Documentos relevantes
52	Temperado or Subtropical	1, 2, 3, 4, 5, 7, 13, 14, 15, 16, 17, 25, 50, 51, 52, 54, 55, 56, 58, 60, 62, 86, 87, 95, 99, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128
53	Semi-árido or Bsh	0, 12, 27, 29, 30, 31, 32, 33, 34, 35, 60, 62, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
54	Subtropical or Cfa	1, 4, 5, 16, 17, 50, 51, 54, 55, 56, 58, 60, 62, 86, 87, 99, 118, 119, 120, 121, 122, 123, 124
55	Temperado or Cwa	1, 2, 3, 5, 7, 13, 14, 15, 16, 17, 25, 50, 51, 52, 54, 55, 56, 58, 60, 62, 86, 87, 95, 99, 121, 123, 125, 126, 127, 128
56	Tropical or Af	3, 6, 7, 8, 9, 10, 11, 12, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 36, 39, 41, 42, 43, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 79, 80, 81, 82, 83, 99, 108, 116, 121
57	Minas Gerais or Cwb	2, 3, 5, 7, 13, 14, 15, 16, 17, 19, 25, 50, 51, 52, 53, 111
58	Pará or Am	8, 9, 11, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 36, 39, 41, 42, 44, 60, 61, 62, 64, 79, 83, 99, 105, 123
59	Pernambuco or Bsh	0, 12, 25, 27, 29, 30, 31, 32, 33, 34, 35, 60, 62, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
60	Mato Grosso or Tropical	3, 6, 7, 8, 9, 11, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 36, 39, 41, 42, 43, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 79, 80, 81, 82, 83, 99, 108, 116, 121

Continua na próxima página

Tab. A.15 – continuação da tabela anterior

No.	Consulta	Documentos relevantes
61	Ceará or Semi-árido	0, 12, 25, 27, 29, 30, 31, 32, 33, 34, 35, 60, 62, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
62	São Paulo or Temperado	1, 2, 3, 7, 13, 14, 15, 16, 17, 25, 50, 51, 52, 54, 55, 56, 57, 58, 59, 87, 95, 102, 121, 123, 125, 126, 127, 128
63	Bahia or Equatorial	7, 8, 9, 10, 11, 12, 25, 29, 30, 31, 32, 33, 34, 35, 60, 61, 62, 63, 64, 65, 99, 106, 108, 110, 111, 112, 113, 114, 115, 116, 117, 121
64	Região Norte or Tropical	3, 6, 7, 8, 9, 10, 11, 12, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 36, 37, 38, 39, 40, 41, 42, 43, 44, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 79, 80, 81, 82, 83, 97, 99, 105, 108, 116, 121, 123
65	Região Sudeste or Temperado	1, 2, 3, 5, 7, 12, 13, 14, 15, 16, 17, 19, 25, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 87, 95, 102, 111, 121, 123, 125, 126, 127, 128
66	Região Nordeste or Semi-árido	0, 7, 12, 17, 26, 27, 29, 30, 31, 32, 33, 34, 35, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 98, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
67	Região Norte or Am	8, 9, 10, 11, 12, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 36, 37, 38, 39, 40, 41, 42, 43, 44, 60, 61, 62, 64, 79, 83, 97, 99, 105, 123
68	Região Sudeste or Cwa	1, 2, 3, 5, 7, 12, 16, 17, 19, 25, 26, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 62, 86, 99, 102, 111, 121, 123

Continua na próxima página

Tab. A.15 – continuação da tabela anterior

No.	Consulta	Documentos relevantes
69	Região Sul or Cfa	1, 4, 12, 13, 14, 15, 16, 17, 50, 51, 54, 55, 56, 58, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 100, 101, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128
70	Região Centro-Oeste or Cfb	1, 6, 7, 11, 12, 13, 14, 15, 16, 17, 19, 25, 28, 42, 54, 55, 56, 58, 83, 84, 85, 86, 87, 95, 103, 104, 121, 125, 126, 127, 128
71	Amazonas	8, 9, 10, 11, 18, 22, 28, 36, 37
72	Rio Grande do Sul	4, 14, 15, 91, 92, 93, 94, 118, 119, 120, 121, 122, 123, 125, 126, 127
73	Bahia	7, 25, 29, 30, 31, 32, 33, 34, 35, 60, 61, 62, 63, 64, 65, 106, 108, 110, 111, 112, 113, 114, 115, 116, 117, 121
74	Cfa	1, 4, 16, 17, 50, 51, 54, 55, 56, 58, 87, 118, 119, 120, 121, 122, 123, 124
75	Bsh	0, 12, 27, 29, 30, 31, 32, 33, 34, 35, 60, 62, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
76	Af	9, 10, 11, 12, 60, 61, 62, 64, 99
77	Aw	6, 7, 9, 17, 18, 19, 23, 25, 26, 27, 28, 43, 50, 51, 53, 54, 55, 56, 57, 58, 60, 62, 63, 65, 66, 67, 68, 69, 70, 71, 79, 80, 81, 82, 83, 99, 116
78	Região Centro-Oeste	6, 7, 11, 12, 19, 28, 42, 83, 84, 85, 86, 103, 104
79	Região Nordeste	0, 7, 12, 17, 26, 27, 29, 30, 31, 32, 33, 34, 35, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 98, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123

Continua na próxima página

Tab. A.15 – continuação da tabela anterior

No.	Consulta	Documentos relevantes
80	Região Sudeste	1, 2, 3, 5, 7, 12, 16, 17, 19, 25, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 102, 111, 123
81	Semi-árido	0, 12, 27, 29, 30, 31, 32, 33, 34, 35, 60, 62, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
82	Tropical	3, 6, 7, 8, 9, 11, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 36, 39, 41, 42, 43, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 79, 80, 81, 82, 83, 99, 108, 116, 121
83	Subtropical	1, 4, 5, 16, 17, 50, 51, 54, 55, 56, 58, 60, 62, 86, 87, 99, 118, 119, 120, 121, 122, 123, 124

Conceito de Clima	Identificador dos documentos relevantes
Bsh	0, 12, 27, 29, 30, 31, 32, 33, 34, 35, 60, 62, 65, 66, 67, 68, 69,70, 71, 72, 73, 74, 75, 79, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
Cfb	1, 13, 14, 15, 16, 17, 25, 54, 55, 56, 58, 87, 95, 121, 125, 126, 127, 128
Cwb	2,3, 13, 14, 15, 16, 17, 25, 50, 51, 52
Cfa	1, 4, 16, 17, 50, 51, 54, 55, 56, 58, 87, 118, 119, 120, 121, 122, 123, 124
Cwa	1,5, 16, 17, 50, 51, 52, 60, 62, 86, 99, 121, 123
Aw	6, 7, 9, 17, 18, 19, 23, 25, 26, 27, 28, 43, 50, 51, 53, 54, 55, 56, 57, 58, 60, 62, 63, 65, 66, 67, 68, 69, 70, 71, 79, 80, 81, 82, 83, 99, 116
Am Af	8, 9,11, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 36, 39, 41, 42, 60, 61, 62, 64, 79, 83, 99 9, 10, 11, 12, 60, 61, 62, 64, 99
Equatorial	8, 9, 10, 11, 12, 60, 61, 62, 64, 99
Temperado	1,2, 3, 7, 13, 14, 15, 16, 17, 25, 50, 51, 52, 54, 55, 56, 58, 87, 95, 121, 125,126,127, 128
Subtropical	1, 4, 5, 16, 17, 50, 51, 54, 55, 56, 58, 60, 62, 86, 87, 99, 118,119, 120, 121, 122, 123, 124
Tropical	3, 6, 7, 8, 9, 11, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 36, 39, 41, 42, 43, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 79, 80, 81, 82, 83, 99, 108,116, 121
Semi-árido	0, 12, 27, 29, 30, 31, 32, 33, 34, 35, 60, 62, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 99, 106,107,108, 109. 110, 111, 112, 113, 114, 115, 116, 117, 121, 123
Clima	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 39, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 86, 87, 95, 99, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128

Tab. A.13: Documentos relevantes para os conceitos da ontologia de classificação climática no Brasil.