UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE
ÁREA DE CIÊNCIA DA COMPUTAÇÃO

Fábio Augusto de Santana Silva

UM MODELO DE RECUPERAÇÃO DE INFORMAÇÃO PARA A WEB SEMÂNTICA

São Luís-MA
2009

FÁBIO AUGUSTO DE SANTANA SILVA

UM MODELO DE RECUPERAÇÃO DE INFORMAÇÃO PARA A WEB SEMÂNTICA.

Dissertação de Mestrado apresentada ao curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão, como parte dos requisitos para a obtenção do título de Mestre em Engenharia de Eletricidade, na área de Ciência da Computação.

Orientadora: Prof[a]. Dra. Rosário Girardi

São Luís-MA
2009

Silva, Fábio Augusto de Santana

Um modelo de recuperação de informação para a web semântica/ Fábio Augusto de Santana Silva. – São Luís, 2009.

122 f.

Orientadora: Profa. Dra. Rosário Girardi

Dissertação (Mestrado) – Curso de Pós Graduação em Engenharia da Eletricidade – Ciências da Computação, Universidade Federal do Maranhão, 2009.

1. Web semântica. 2. Recuperação na web. 3. Organização da informação na web. I. Título.

CDU 004.423.4:007

UM MODELO DE RECUPERAÇÃO DE INFORMAÇÃO PARA A WEB SEMÂNTICA.

FÁBIO AUGUSTO DE SANTANA SILVA

Dissertação aprovada em 18 de maio de 2009.

Profa. Maria del Rosário Girardi, Ph. D.
(Orientadora)

Prof. Stanley Loh, Dr.
(Membro da Banca Examinadora)

Prof. Osvaldo Ronald Saavedra Mendez, Dr.
(Membro da Banca Examinadora)

A minha esposa Renata, que me deu incentivo e apoio durante essa caminhada, mesmo abrindo mão de nossos momentos.

A minha mãe Neyde, que sacrificou muitos dos seus momentos para que eu tivesse uma formação para enfrentar os desafios da vida.

AGRADECIMENTOS

A todos que acreditaram em mim e também tornaram esse trabalho possível.

A equipe de informática do Hospital Sarah – São Luís pelo apoio e incentivo.

A todos os colegas do mestrado e do grupo de pesquisa, em especial a, Adriana Leite, Uiratan Cavalcante, Djefferson Smith Santos Maranhão e à Professora Girardi pela paciência e apoio.

Ao Lucas Drumond, que trabalhou comigo nas publicações, e cujos trabalhos me ajudaram a compor esta dissertação.

“A cada instante há que sacrificar o que somos ao que podemos vir a ser.”

Charles Bos

RESUMO

Várias técnicas para extrair significado de textos com o objetivo de construir representações internas mais precisas, tanto para itens de informação quanto para consultas em sistemas de recuperação já foram propostas. Contudo, faltam modelos de recuperação baseados em semântica que especifiquem abstrações apropriadas para essas técnicas. Este trabalho apresenta um modelo de recuperação baseado no conhecimento que explora o conteúdo semântico dos itens de informação. A representação interna dos itens de informação é baseada em grupos de interesse do usuário chamados de “casos semânticos”. O modelo também define um critério para a recuperação dos itens de informação e uma função para ordenar os resultados obtidos que utiliza medidas de similaridade baseadas na distância semântica entre os elementos das representações internas. O modelo foi instanciado em um sistema construído para o domínio jurídico tributário usando a ontologia ONTOTRIB, uma extensão da ontologia genérica ONTOJURIS, que permite a instanciação de instrumentos jurídico-tributários. Os resultados obtidos nos testes realizados neste domínio específico apontaram uma melhoria da precisão em relação a um sistema baseado em palavras-chave.

Palavras-chave: Modelo de Recuperação de Informação, Web Semântica, Filtragem de Informação, Ontologia Tributária.

ABSTRACT

Several techniques for extracting meaning from text in order to construct more accurate internal representations of both queries and information items in retrieval systems have been already proposed. However, there is a lack of semantic retrieval models to provide appropriate abstractions of these techniques. This work proposes a knowledge--based information retrieval model that explores the semantic content of information items . The internal representation of information items is based on user interest groups, called “semantic cases”. The model also defines a criteria for retrieve information items and a function for ordering the results that uses similarity measures based on semantic distance between semantic cases items. The model was instantiated by a sample system built upon the tributary legal domain using the specialization of the ONTOJURIS, a generic legal ontology, called ONTOTRIB. Legal normative instruments can be instantiated in a knowledge base by ONTOTRIB classes. The results obtained for this specific domain showed an improvement in the precision rates compared to a keyword-based system.

Keywords: Information Retrieval Model, Semantic Web, Information Filtering, Tributary Ontology.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

API	Application Program Interface
DOSE	Distributed Open Semantic Elaboration
FI	Filtragem de Informação
GESEC	Grupo de Pesquisa em Engenharia de Software e Engenharia do Conhecimento
HTML	Hypertext Markup Language
ICMS	Imposto sobre Circulação de Mercadoria e Prestação de Serviços de Comunicação e de Transporte Interestadual e Intermunicipal de Transporte
IE	Imposto de Exportação
IEEE	Institute of Electrical and Electronics Engineers, Inc.
II	Imposto de Importação
IPI	Imposto sobre Produtos Industrializados
IPVA	Imposto sobre a Propriedade de Veículo Automotor
IR	Imposto sobre a Renda e Proventos de Qualquer Natureza
ISSQN	Imposto sobre Serviços de Qualquer Natureza
ITBI	Imposto Sobre A Transmissão De Bens Imóveis E De Direitos A Eles Relativos
ITCD	Imposto Causa Mortis e Doações
ITNG	International Conference on Information Technology
ITR	Imposto sobre A Propriedade Territorial Rural
KIM	Knowledge Information and Management
MSC	Most Specific Concept
ONTOJURIS	Ontologia para o domínio jurídico
ONTOTRIB	Ontologia para o domínio jurídico tributário
OWL	Ontology Web Language
PLN	Processamento de Linguagem Natural
RDF	Resource Description Framework
RI	Recuperação de Informação
ROSA	Retrieval Of Software Artifacts
SC	Semantic Cotopy

SHOE	Simple HTML Ontology Extensions
STF	Supremo Tribunal Federal
TF x IDF	Total Frequency-Inverse Document Frequency
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	Extensible Markup Language

SUMÁRIO

1. INTRODUÇÃO

Desde o seu surgimento, a tecnologia de informação tem sido de grande valia para as atividades humanas. A representação da informação em meio digital facilita a criação, a disponibilização, o acesso, a pesquisa e a recuperação de conteúdo. Um dos grandes repositórios de informação, de uso geral, é a World Wide Web, popularizada no início da década de 90, e que guarda uma enorme quantidade de documentos sobre os mais variados assuntos.

Os usuários, por sua vez, possuem necessidades de informação específicas e que podem ser satisfeitas a partir de um conjunto restrito de documentos presentes na internet. A descoberta da informação de forma precisa é uma tarefa árdua para os usuários (BAEZA-YATES, RIBEIRO-NETO, 1998) devido ao crescimento exponencial do número de documentos nesses repositórios, gerando uma sobrecarga de informação ("*Information Overload*"). O usuário não é capaz de lidar com a grande quantidade de documentos que continuam a ser produzidos e disponibilizados ininterruptamente (DACONTA, OBRST, SMITH, 2003).

Além disso, a Web não foi projetada de modo a permitir a publicação e recuperação de informação de modo estruturado. Embora seja um enorme repositório de informação, de fácil acesso e disponibilidade, grande parte do material publicado foi criado com tecnologias para a apresentação de conteúdo ao usuário final, oferecendo facilidades de uso e navegação através das páginas dos sítios. O conteúdo dessas páginas não possui um formato adequado para o processamento por ferramentas de software.

Diversas áreas de pesquisa surgem a partir desse cenário, como a recuperação de informação, a filtragem de informação, a classificação e a categorização de documentos. Apesar de terem objetivos distintos, todas essas áreas compartilham a mesma dificuldade em descobrir o significado das informações que estão publicadas, possibilitando um tratamento mais refinado da informação e oferecendo ao usuário um acesso direto ao conteúdo desejado. A falta de estruturação presente nas grandes bibliotecas digitais diminui a efetividade da maioria das técnicas empregadas para extrair informação requerida por um usuário em qualquer dos processos suportados pelas áreas citadas, tal como a busca de informação e a entrega de recomendações.

1.1 RELEVÂNCIA E MOTIVAÇÃO

Dentro da Ciência da Computação, a área de recuperação de informação (RI) é responsável por métodos para suplantar as dificuldades no processo de busca em ambientes como a Web. Vários modelos de recuperação de informação, como o modelo booleano, o modelo vetorial e o modelo probabilístico, foram propostos para cobrir as atividades de processar uma consulta de usuário bem como armazenar e recuperar itens de informação a partir de fontes desestruturadas (BAEZA-YATES, RIBEIRO-NETO, 1998). Modelos clássicos representam os documentos com um conjunto de palavras-chave extraídas do texto e propõem diferentes abordagens para a recuperação dos itens de informação ordenados de acordo com a sua relevância.

A efetividade dos modelos de processos baseados em palavras-chave está limitada pelo fenômeno conhecido como “barreira das palavras chaves”, ou seja, a representação interna de um item de informação através de apenas um conjunto de palavras extraídas dos textos através de técnicas estatísticas e/ou sintáticas não permite uma melhora considerável da efetividade dos sistemas de recuperação de informação e, em particular, da precisão dos seus resultados (GIRARDI, IBRAHIN, 1995). Em outras palavras, mesmo que resultados relevantes sejam retornados, eles são de pouca utilidade ao serem listados dentro de um conjunto muito grande de dados irrelevantes.

Outros problemas podem ainda ser observados ao utilizar-se palavras-chave como elemento fundamental da representação interna. O uso de palavras-chave é altamente sensível ao vocabulário utilizado, sujeito a ambigüidades e às dificuldades existentes na interpretação da linguagem natural. Os resultados obtidos são documentos completos, que podem estar distribuídos em vários sub-sítios, dificultando a avaliação do resultado retornado.

Essas limitações estimularam o desenvolvimento de várias técnicas buscando a extração de significado dos textos, como a análise semântica, de forma a obter representações internas mais precisas dos itens de informação (DEERWESTER et al., 1990) (GIRARDI, 1995) (LEACOCK, CHODOROW, 1998) (LIN, 1998) (RESNICK, 1999). Contudo, existe uma falta de modelos de recuperação

semântica provendo uma abstração apropriada para representar as atividades, produtos e técnicas envolvidas no processo de recuperação.

A Web Semântica (ANTONIOU, HARMELEN, 2004) (BERNERS-LEE, HENDLER, LASSILA, 2001) (SCHADBOLT, HALL, BERNERS-LEE, 2006) é uma extensão da Web na qual os dados são estruturados de modo a serem lidos também por máquinas e exibidos de forma amigável. A Web Semântica associa aos documentos metadados capazes de representar, descrever e contextualizar a informação de modo preciso e sem ambigüidades (TANNENBAUM, 2001). Assim, a Web será transformada em uma base de conhecimento cujo conteúdo pode ser interpretado por sistemas de recuperação, filtragem e descoberta de informação. As ferramentas de busca de informação se enquadram entre as principais ferramentas utilizadas na Web e necessitam de modelos que explorem os recursos da representação semântica do conteúdo (SCHADBOLT, HALL, BERNERS-LEE, 2006).

Com a representação semanticamente estruturada dos dados na Web, os sistemas de recuperação de informação podem fazer uso de técnicas baseadas em semântica para aumentar a sua efetividade tais como: anotação de documentos, busca baseada em conhecimento, expansão da consulta e medidas de similaridade baseadas em ontologias. Contudo, tais técnicas não são suficientes para gerar sistemas de recuperação totalmente semânticos (SCHEIR, PAMMER, LINDSTAEDT, 2007). Um mecanismo de recuperação, mesmo baseado em conhecimento, deve encontrar documentos que atendam à consulta do usuário e deve fornecer uma ordem de relevância para eles ou partes dele. Por exemplo, algumas abordagens apoiadas por uma base de conhecimento, utilizam consultas formais a essa base, mas os resultados gerados carecem de mecanismos efetivos de ordenação (WEI, BARNAGHI, BARGIELA, 2008).

Muitos dos sistemas de recuperação para a Web Semântica ainda são baseados modelo vetorial, onde itens de informação ainda são representados com palavras-chave e o processo de recuperação se baseia em métodos estatísticos não capturando o significado da informação em cada documento. O uso de palavras-chave é apoiado pelo fato de que a geração de metadados para a grande quantidade de documentos é uma tarefa de alto custo, imprecisa e ainda sem uma solução definitiva (BENZ, HOTHO, 2007) (VALLET, FERNÁNDEZ, CASTELLS, 2005). Contudo, o desenvolvimento de novas técnicas para a anotação automática

(CIMIANO, HANDSCHUH, STAAB, 2004) (HANDSCHUH, STAAB, 2003) ou semi-automática (KIRYAKOV et al., 2004) tem nos dado exemplos de que este cenário passa por significativos avanços.

Este trabalho propõe um modelo de recuperação que usa estruturas baseadas em ontologias para representar os itens de informação e também uma medida de similaridade baseada em casos semânticos para a ordenação dos resultados. Sobre este modelo, novos sistemas podem ser construídos e os autores podem aplicar diferentes técnicas para encontrar a melhor configuração para as tarefas de recuperação em um domínio.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Especificação de um modelo de recuperação de informação para a Web Semântica baseado em estruturas de representação do conhecimento.

1.2.2 Objetivos Específicos

- Desenvolvimento de um modelo que possa ser aplicado à área de recuperação de informação e que aproveite as estruturas de representação do conhecimento da Web Semântica.
- Avaliar a efetividade do modelo através do desenvolvimento de um estudo de caso com a sua instanciação em um sistema de recuperação baseado no domínio tributário.
- Avaliação das estruturas de representação do conhecimento desenvolvidas pelo grupo GESEC na instanciação daquele sistema.

1.3 ESTRUTURA DA DISSERTAÇÃO

O capítulo 2 mostra o referencial teórico observado no trabalho, discutindo os modelos clássicos de recuperação de informação e seus aspectos gerais. Ainda

nesse capítulo, é feita uma breve descrição das principais tecnologias que dão suporte à Web Semântica e que permitem a representação de conhecimento fornecendo a base conceitual para a construção do modelo proposto. O capítulo 3 apresenta os principais sistemas de recuperação de informação construídos para a Web Semântica e compara as suas características. O capítulo 4 apresenta o modelo de recuperação proposto, detalhando os seus componentes. O capítulo 5 apresenta um estudo de caso a cerca do modelo e que utiliza a ONTOTRIB, ontologia do domínio tributário desenvolvida pelo grupo GESEC. O capítulo 6 expõe as conclusões sobre o trabalho realizado, enfatizando os resultados e os trabalhos futuros.

2. MODELOS DE RECUPERAÇÃO E WEB SEMÂNTICA

Esse capítulo apresenta os aspectos gerais dos trabalhos mais representativos na área de recuperação de informação. Os primeiros trabalhos na área da recuperação de informação surgiram na década de 60, a partir do trabalho de Salton (SOUZA, ALVARENGA, 2004). A pesquisa nesta área ganhou um novo impulso a partir do surgimento e popularização da Web na década de 90. O advento de uma grande biblioteca digital exigiu que fossem desenvolvidos mecanismos de recuperação de informação para grandes bases de documentos.

Muitos modelos de recuperação de informação foram propostos desde o início da pesquisa nessa área, sendo que os modelos mais representativos são: o modelo booleano, o modelo vetorial e o modelo probabilístico. Esses modelos de recuperação utilizam representações dos itens de informação baseadas em palavras-chave que apresentam conhecidas limitações para o tratamento da informação (ANTONIOU, HARMELEN, 2004). A Web Semântica é uma área de pesquisa que desenvolve tecnologias para extrair, representar e distribuir conhecimento a partir do conteúdo dos itens de informação. Tais tecnologias são apresentadas nas últimas seções deste capítulo, bem como uma breve discussão de como essas tecnologias podem contribuir para as pesquisas na área de recuperação de informação em busca de modelos mais efetivos.

2.1 RECUPERAÇÃO DE INFORMAÇÃO

A área da recuperação de informação (RI) é responsável pelo desenvolvimento de técnicas que permitam representar, armazenar e localizar itens de informação. Enquanto os sistemas de informação tradicionais realizam a recuperação de dados, os sistemas de RI recuperam informação (BAEZA-YATES, RIBEIRO-NETO, 1998). No primeiro caso, todo o conteúdo disponível possui estruturação, uma formatação determinada e as consultas são expressas em uma linguagem formal. Os resultados representam uma resposta exata à consulta efetuada pelo usuário. No segundo caso, o conteúdo não possui uma estruturação definida, a consulta expressa de forma vaga a necessidade do usuário e o resultado

obtido pelo processo de busca não é exato. Dessa forma, os resultados apresentam um grau de imprecisão. Podemos identificar três principais tarefas envolvidas no processo de RI:

1. Representação dos elementos de informação e da consulta dos usuários;
2. Análise de similaridade entre as representações envolvidas no processo;
3. Apresentação dos resultados para o usuário final.

A RI atende a uma necessidade pontual de informação do usuário (BECHHOFER et al., 2004) e a consulta é processada contra uma fonte de informação de conteúdo estático e desestruturado conforme mostrado na Figura 1.

Em (MANNING, RAGHAVAN, SCHÜTZE, 2008), a recuperação de informação é definida como:

"Encontrar material (usualmente documentos) de natureza desestruturada (normalmente texto) que satisfaça uma necessidade de informação a partir de grandes coleções".

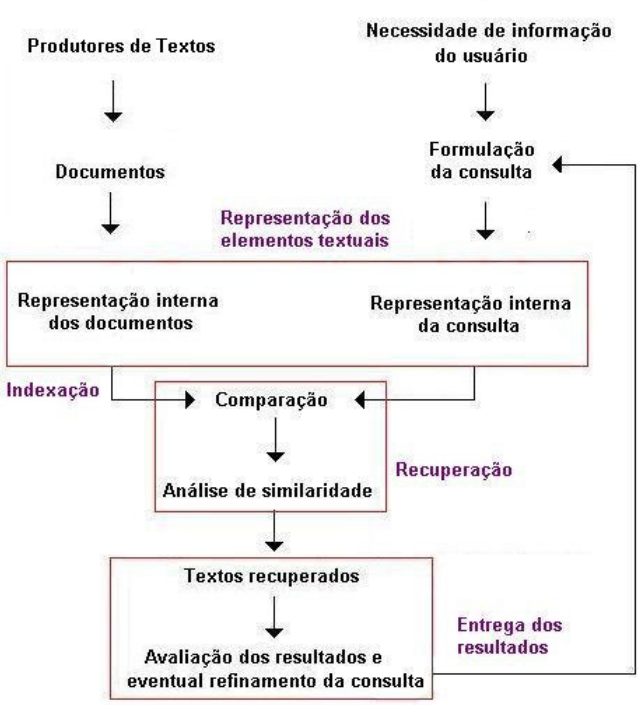

Figura 1 Processo genérico de recuperação de informação.

2.2 MODELO BOOLEANO

O modelo booleano é o mais simples dentre os modelos de recuperação de informação e se baseia na teoria dos conjuntos e álgebra booleana. Nesse modelo, os documentos são representados por termos que podem estar presentes ou ausentes em cada documento. O modelo apresenta as seguintes estruturas:

- Um documento K é representado por um conjunto de palavras-chaves tal que $K = \{k_1, k_2, \dots, k_n\}$;
- Cada termo k_i recebe um peso ω, sendo que o modelo booleano admite apenas que w pertença a {0, 1};
- Uma consulta q é composta de palavras-chave concatenadas pelos operadores booleanos and, or e not;
- O conjunto resposta R de uma consulta q será o conjunto de documentos que possuir os termos que tornem a expressão booleana da consulta verdadeira.

Por exemplo, tomemos os documentos listados nas colunas da Tabela 1 e os termos constantes nas linhas da tabela. A consulta $q = \{a\ and\ e\}$ terá como o conjunto resposta $\{D_1, D_3\}$, que são os documentos que possuem o valor 1 para os termos "a" e "e".

Tabela 1 Exemplo da representação de documentos no modelo booleano

	D_1	D_2	D_3	D_4	D_5
Termo a	1	0	1	0	0
Termo b	1	0	0	0	1
Termo c	0	0	0	1	0
Termo d	0	1	0	1	1
Termo e	1	0	1	1	0

O modelo booleano é de fácil compreensão e bastante intuitivo possibilitando que os sistemas sejam construídos de maneira simples. Porém, apresenta sérias desvantagens, como o fato de considerar os documentos apenas

como relevantes ou irrelevantes, não apresentando uma ordenação para os resultados. Por essa característica, o modelo se assemelha a um modelo de recuperação de dados e não de informação. Tentando superar esta limitação, foram propostas extensões do modelo booleano que exploram os termos da consulta para gerar uma ordenação dos resultados. Ainda assim, o modelo apresenta um desempenho inferior aos demais modelos clássicos (BAEZA-YATES, RIBEIRO-NETO, 1998).

2.3 MODELO VETORIAL

O modelo vetorial representa os documentos através de um vetor de termos em um espaço vetorial de muitas dimensões. O tamanho do espaço vetorial é dado pelo número total de termos indexados na coleção. Da mesma forma que o modelo booleano, um peso é associado a cada termo da representação, porém, o valor dado para este peso não é restrito a valores binários. Se o termo pertencer ao documento ele receberá um valor maior que zero, representando a sua importância relativa, do contrário terá o valor zero. Esses pesos são utilizados para calcular o grau de similaridade entre os documentos, ou seja, o documento possui uma relevância apenas parcial em relação à consulta feita pelo usuário. O modelo foi criado por Salton (SALTON, WONG, YANG, 1975), sendo o modelo de recuperação de informação mais utilizado devido ao seu bom desempenho no processo de recuperação e à sua simplicidade conceitual. O modelo vetorial pode ser descrito dessa forma:

- Seja K o conjunto formado pelos termos existentes em uma coleção. Temos que $K = \{k_1, k_2, \ldots, k_n\}$ definirá um espaço vetorial de n dimensões, em que cada termo representa uma dimensão ortogonal do espaço;
- Seja D um documento da coleção. Então $D = \{\omega_1, \omega_2, \ldots, \omega_n\}$ sendo ω_i o peso do termo i no documento D.
- Seja Q uma consulta submetida à coleção. Então $Q = \{\omega_1, \omega_2, \ldots, \omega_n\}$ sendo ω_i o peso do termo i na consulta Q.

A Figura 2 apresenta um exemplo de como seria a representação interna de três documentos em um espaço vetorial com os termos t_1 e t_2. A figura apresenta

os pesos relativos aos termos representados. Pode-se observar que o peso atribuído ao termo t_1 na consulta não possui um valor binário, e também a ortogonalidade entre os termos, com o termo t_1 sendo apresentado na coordenada x, enquanto que o termo t_2 é apresentado na coordenada y do gráfico.

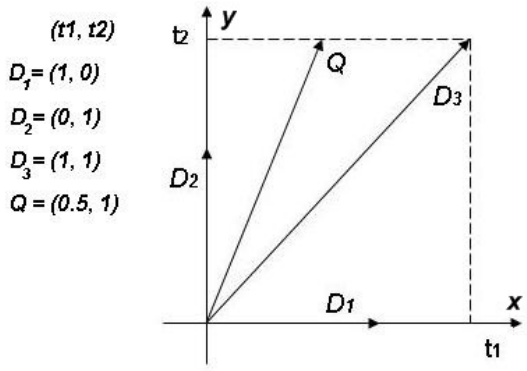

Figura 2 Exemplo da representação interna no modelo vetorial

O grau de similaridade entre uma consulta e um documento no modelo vetorial é dada pela similaridade entre os vetores que os representam. O ângulo entre os vetores nos dá a medida da divergência entre os vetores, sendo que o cosseno do ângulo nos dá um valor numérico para tal divergência. A medida do cosseno atribui o valor 0 para vetores ortogonais e o valor 1 para vetores idênticos. Sendo os vetores de tamanho idêntico, o cosseno do ângulo é dado pelo produto interno dos vetores.

Definição 1: A similaridade entre dois vetores $\overrightarrow{D}$ e $\overrightarrow{Q}$, de mesmo tamanho (SALTON, BUCKLEY, 1988), é dada por:

$$Sim\left(\overrightarrow{D}, \overrightarrow{Q}\right) = \sum_{t_i \in Q,D} \omega_{t_{iQ}} * \omega_{t_{iD}}$$

onde $\omega_{t_{iQ}}$ é o valor do peso de índice i no vetor que representa a consulta e $\omega_{t_{iD}}$ é o valor do peso de índice i no vetor que representa o documento.

A forma de cálculo dos pesos associados aos termos não faz parte da definição, mas é um fator importante para a efetividade do modelo. Apesar das diversas fórmulas já propostas e discutidas, a medida mais comumente adotada pela

maioria dos sistemas é baseada na estatística de ocorrência dos termos na coleção, chamada $tfx\ idf$ (*term frequency, inverse document frequency*). A fórmula pode ser definida como:

- Seja $tf(i,d)$ a freqüência do termo de índice i no documento d;
- Seja $idf(i) = \log(N/N_i)$ o inverso da freqüência do termo de índice i na coleção.

onde N é o total de itens da coleção e N_i o total de itens em que o termo de índice i ocorre.

Definição 2: O peso do termo i no documento d é dado por:

$$tf(i,d)\ x\ idf(i)$$

O modelo vetorial apresenta como vantagens a possibilidade de ordenação dos resultados gerando uma ordem de relevância, o valor obtido com o casamento parcial dá uma medida de quão similar o documento é em relação à consulta e a fórmula simples melhora o desempenho da recuperação nos sistemas construídos. A principal crítica ao modelo vetorial é que ele não considera a dependência entre os termos dos documentos, uma vez que os termos são independentes e ortogonais. Apesar dessa deficiência, os resultados obtidos pelo modelo são quase sempre superiores a outras abordagens ou , ao menos, se aproximam dos melhores resultados obtidos.

2.4 MODELO PROBABILÍSTICO

O modelo probabilístico não trabalha com pesos como os dois modelos apresentados anteriormente, mas com a probabilidade de que um documento *d* seja relevante para uma dada consulta *q,* de acordo com a representação interna de ambos. O modelo assume que existe na coleção um conjunto ideal que contém exatamente os documentos relevantes à consulta. O problema do modelo probabilístico é definir este conjunto ideal em termos de suas propriedades, o que seria equivalente à consulta. Ou seja, não se conhece, à priori, as características desse conjunto ideal e não há como gerar as probabilidades desejadas. Portanto, na

prática, é necessário gerar conjuntos testes para estimar as probabilidades de relevância para uma determinada coleção.

Formalmente, o modelo descreve os documentos e as consultas considerando pesos binários (0 ou 1) que representam a ausência ou a presença de termos da coleção. Então, considerando que $P(+R_Q|D)$ é a probabilidade de que o documento D seja relevante para a consulta Q e que $P(-R_Q|D)$ é a probabilidade de que o documento D não seja relevante para a consulta Q, pode-se considerar que um documento D é relevante para consulta Q se $P(+R_Q|D) > P(-R_Q|D)$.

Definição 3: A ordem de relevantes (peso) no modelo probabilístico é dada através da seguinte equação:

$$\omega_{D|Q} = \frac{P(+R_Q|D)}{P(-R_Q|D)}$$

Definição 4: A ordem de relevantes após a aplicação do teorema de Bayes na equação da Definição 3, é dada por:

$$Similaridade(Q, D) = \sum_{i=1}^{N} x_i * \omega_{qi}$$

onde

$x_i \in \{0, 1\}$ e $\omega_{qi} = \log\left(\frac{(r_{Qi*}(1-s_{Qi}))}{(s_{Qi*}(1-r_{Qi}))}\right)$,

$r_{Qi} = P(x_i = 1 \mid +R_Q)$ e

$s_{Qi} = P(x_i = 1 \mid -R_Q)$.

Os termos r_{Qi} e s_{Qi} correspondem à probabilidade de que o termo de índice i ocorra no documento, dado que o documento D seja ou não, respectivamente, relevante para a consulta Q.

O modelo probabilístico apresenta como principal vantagem permitir a ordenação probabilística das repostas, o que resulta num bom desempenho do método. Por outro lado, o comportamento dependente da precisão das

probabilidades e a ausência do cálculo da freqüência do termo nos documentos são as principais desvantagens deste modelo.

2.5 AVALIAÇÃO DA EFETIVIDADE DA RECUPERAÇÃO

A avaliação da efetividade dos modelos de recuperação tem sido um dos maiores desafios da comunidade de pesquisa nessa área (SINGHAL, 2001). As medidas tradicionalmente aceitas para realizar a avaliação são a *revocação* (*recall*) e a *precisão* (*precision*). *Revocação* é a fração de documentos relevantes recuperados, e avalia a capacidade do sistema em recuperar a informação desejada. *Precisão* é a fração dos documentos recuperados que é relevante, e avalia se o sistema é capaz de ignorar material irrelevante. Esses conceitos podem ser calculados conforme as definições (5) e (6):

Seja I uma requisição de informação a uma determinada coleção satisfeita por um conjunto de itens relevantes R, sendo $|R|$ o total de itens relevantes. Seja A o conjunto resposta dado por um determinado modelo de recuperação, sendo $|A|$ o total de itens desse conjunto. Seja $|R_a|$ o número de elementos do conjunto resultante da interseção entre R e A.

Definição 5: Revocação

$$Revocação = \frac{|R_a|}{|R|}$$

Definição 6: Precisão

$$Precisão = \frac{|R_a|}{|A|}$$

Um ponto a ser observado é que o conjunto A, que é o conjunto resposta, não é visto pelo usuário de uma só vez. Os documentos desse conjunto são avaliados de acordo com a ordem de relevância dada. Por isso, a precisão é melhor avaliada quando observada sobre diferentes taxas de revocação. Essa forma de

observação gera uma curva característica da relação (GIRARDI, 1995) (Figura 3).

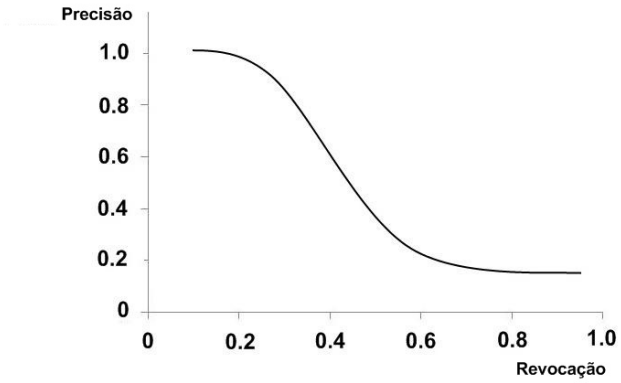

Figura 3 Curva típica de uma avaliação revocação x precisão (GIRARDI, 1995)

Um bom sistema de recuperação de informação deve ter bons índices de *revocação* e *precisão*, porém, isto não é facilmente observado. À medida que a *revocação* aumenta, quando procuramos novos itens relevantes, a *precisão* diminui. As técnicas utilizadas para aumentar a *precisão* em geral diminuem a *revocação* e vice-versa (GIRARDI, 1995), portanto, a efetividade na recuperação depende do equilíbrio entre esses fatores.

2.6 A WEB SEMÂNTICA

A definição clássica para a Web Semântica foi dada em (BERNERS-LEE, HENDLER, LASSILA, 2001), no artigo em que apresenta uma nova visão para a Web, em que "*a Web Semântica não é uma Web separada, mas uma extensão da Web atual, na qual a informação possui um significado bem definido, possibilitando que computadores e seres humanos trabalhem em cooperação*".

A Web Semântica ainda está em evolução, mas já apresenta uma série de tecnologias para a publicação de conteúdo, de modo que possam ser entendidos por máquinas, a fim de melhorar a efetividade do acesso à informação na Web (DRUMOND, GIRARDI, 2006). Essas tecnologias substituem o HTML e o uso de aplicativos proprietários na Web por linguagens que transportam tanto o conteúdo quanto a descrição deste conteúdo, em outras palavras, usam metadados associados aos documentos. Esses metadados capturam parte da semântica da informação e podem ser processados por sistemas automatizados.

As técnicas para a representação do conhecimento da Inteligência Artificial contribuíram para o surgimento das tecnologias necessárias à estruturação da Web Semântica (BERNERS-LEE, HENDLER, LASSILA, 2001). Dessa forma, mesmo que a visão descrita no artigo inicial não seja ainda possível de ser implementada, técnicas para suportar funções imaginadas estão sendo aprimoradas e criadas, dando suporte a aplicações para esse novo ambiente. O W3C (*World Wide Web Consortium*) apresentou, no ano de 2000, uma proposta para a representação do conhecimento neste novo ambiente baseada em camadas (ANTONIOU, HARMELEN, 2004) (Figura 4).

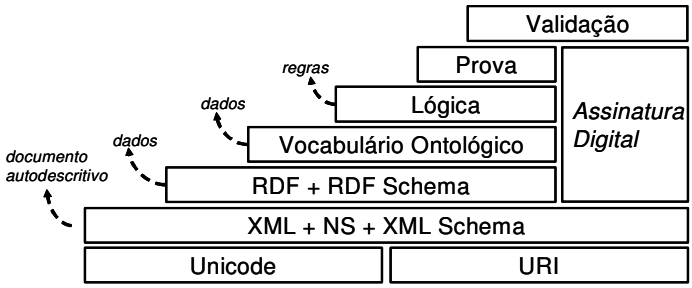

Figura 4 Camadas da Web Semântica (ANTONIOU, HARMELEN, 2004)

As camadas da estrutura da representação do conhecimento são sobrepostas de modo que os níveis superiores sejam compatíveis com os níveis inferiores e que um agente que trabalhe em um determinado nível seja capaz de trabalhar com os níveis mais baixos. A compatibilidade reversa é parcial, pois os agentes que trabalham em um nível são capazes de compreender parte da estrutura do nível superior.

As principais tecnologias envolvidas serão brevemente descritas nas próximas seções, evidenciando-se o papel que cada uma exerce e como se relaciona com as camadas próximas.

2.6.1 URI (*Uniform Resource Identifier*)

URI (*Uniform Resource Identifier*) (BERNERS-LEE, FIELDING, MASINTER, 1998) é uma seqüência de caracteres que identifica um recurso físico ou abstrato. O URI busca identificar os itens através da representação dos seus mecanismos de acesso ao invés do nome ou outro atributo. A uniformidade sugerida

advém do fato de podermos representar qualquer tipo de recurso através destes mecanismos. O URI possui uma sintaxe organizada hierarquicamente a partir do elemento mais importante para o menos importante. Podemos assim definir diversos recursos, de diferentes tipos, dentro de um mesmo contexto.

```
ftp://ftp.is.co.za/rfc/rfc1808.txt
http://www.ietf.org/rfc/rfc2396.txt
ldap://[2001:db8::7]/c=GB?objectClass?one
```

Figura 5 Exemplos de URIs.

2.6.2 XML

XML (ANTONIOU, HARMELEN, 2004) (BRAY et al., 2004) é uma linguagem de marcação que utiliza marcações (*tags*), similar à linguagem HTML, mas com a função de descrever a estrutura do conteúdo transportado. As marcações podem ser aninhadas para compor partes de um documento. Enquanto o HTML possui marcações fixas, que focam na apresentação do conteúdo, o XML permite que o criador defina suas próprias marcações e o formato do documento de acordo com a sua necessidade. A Figura 6 mostra um exemplo de como a sentença "$I\ Just\ got\ a\ new\ pet\ dog$" poderia ser representada em XML.

```
<sentence>
<person webid="http://example.com/#johnsmith">I</person>
just got a new pet
<animal>dog</animal>.
</sentence>
```

Figura 6 Exemplo de representação de uma sentença em XML

Um documento em XML se inicia com um cabeçalho informando que o arquivo se trata de um XML, a versão e o padrão de codificação de caracteres utilizado. É possível também definir referências a recursos externos que serão utilizados ao longo do documento. A principal parte de um arquivo XML são os elementos, que são marcações definidas pelo usuário informando o que está sendo transportado. O formato de um elemento é uma marcação que abre o elemento, o conteúdo do elemento e uma marcação fechando o elemento. Na Figura 6, *<animal> dog </animal>* é um elemento do arquivo XML.

Além do elemento em si, diversas outras estruturas podem ser utilizadas para transportar conteúdo em XML. Um elemento pode ser descrito em termos de atributos e não apenas de um valor simples, comentários, instruções de processamento e formatação também fazem parte da sintaxe XML.

Um documento XML é considerado bem formado caso obedeça às regras sintáticas definidas pela linguagem. A linguagem não impõe restrições semânticas ao conteúdo expresso, mas sua estrutura pode ser restringida por meio de um esquema (XML Schema) ou de um dicionário de tipos (XML DTD). Dessa forma, um documento XML pode ser bem formado e, mesmo assim, não ser um documento válido ao não obedecer à estrutura especificada para ele. Tendo uma estrutura bem definida e de conteúdo flexível, o XML é um meio para transporte de dados: simples, independente da plataforma e da aplicação.

2.6.3 RDF

RDF (Resource Description Framework) (ANTONIOU, HARMELEN, 2004) (BECKETT, 2004) é um modelo para a descrição de recursos da Web ou referenciados na Web. Um recurso pode ser tanto um documento da Web, como um conceito descrito dentro de um contexto. O RDF se propõe a ser uma estrutura que permita eliminar a ambigüidade na troca de informações, ou seja, garantir a interoperabilidade, tornando o conteúdo dos documentos expressos em RDF interpretável independentemente do auxílio de seu criador. Para isso, o RDF possibilita a descrição de um contexto para os recursos.

O RDF se baseia na premissa de que os recursos podem ser descritos através de declarações (MANOLA, MILLER, 2004). Cada declaração é expressa em termos das seguintes partes: sujeito (o recurso), predicado (a propriedade) e o objeto (valor da propriedade). Essa forma de expressão pode ser vista como uma tripla RDF ou mesmo como um predicado binário numa formulação em lógica. O RDF pode também ser expresso através de um grafo no qual o arco é o predicado e os nodos são o sujeito e o objeto. O recurso é identificado através de um URI. Em particular, são utilizados URIrefs que são referências para partes específicas de uma URI. Uma URIref é uma URI seguida de um “#” e um identificador da parte referenciada. O objeto pode ser um valor literal ou mesmo um URI. A propriedade

identifica a relação existente entre ambos. Para expressar e transportar o conjunto de relações, o RDF define uma sintaxe XML particular chamada RDF/XML. Na Figura 7, temos um exemplo em que o recurso é a página HTML *http://www.example.org/index.html*, o predicado é a sua data de criação e o objeto é o valor da propriedade *August 16, 1999*.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#"
xmlns:exterms="http://www.example.org/terms/">
<rdf:Description
rdf:about="http://www.example.org/index.html">
<exterms:creation-date>August 16, 1999</exterms:creation-date>
</rdf:Description>
</rdf:RDF>
```

Figura 7 Exemplo de RDF/XML que mostra a data de criação da página HTML

O RDF/XML permite a formulação de declarações acerca de outras declarações através do mecanismo de reificação. Esse mecanismo associa identificadores únicos a uma declaração que pode, assim, ser referenciada em outras declarações.

O RDF provê um modelo para expressar declarações a cerca de recursos de modo genérico. Para expressar recursos a cerca de um domínio específico é importante que seja definido um vocabulário que o descreva. O RDF Schema (BRICKLEY, GUHA, 2004) é um conjunto de recursos com significado especial que descreve outros recursos em termos de classes e propriedades. O RDF Schema aumenta a capacidade de expressão do RDF possibilitando a representação de estruturas hierárquicas.

```
ex:MotorVehicle    rdf:type    rdfs:Class
```

Figura 8 Definição de uma classe através do RDF Schema

O RDF Schema não disponibiliza o vocabulário em si, mas sim, os meios para descrever um vocabulário qualquer. Uma classe é descrita como um recurso cuja propriedade *rdf:type* tem o valor *rdfs:Class*, conforme exemplo na Figura 8. Uma instância, por sua vez, possui a propriedade *rdf:type* com um valor referente a um recurso definido como uma classe. Existem recursos para definir sub-classes,

propriedades, valores que uma propriedade pode assumir, sub-propriedades e o vínculo entre uma propriedade a uma classe. Uma propriedade é independente da classe, que pode ter zero ou mais propriedades associadas. A mesma propriedade pode ser compartilhada por mais de uma classe, porém, isto determina que os recursos que possuam esta propriedade são instâncias de ambas as classes.Na prática isto limita o uso de propriedades genéricas como, por exemplo, “Pai” que não pode ser usada ao descrever seres humanos e animais mesmo que ambas as classes possuam exatamente a mesma propriedade.

Enquanto o XML se concentra na estrutura de transporte da informação, o RDF modela aspectos referentes à descrição da informação transportada. Porém, as capacidades do RDF, mesmo ampliadas com a adoção do RDF Schema, não são suficientes para expressar muitas das características dos domínios reais.

2.6.4 Ontologias

O termo ontologia é utilizado em diversas áreas do conhecimento. Tem sua origem na Grécia Antiga, quando foi introduzido por Aristóteles. O termo, inicialmente usado na filosofia, foi incorporado por cientistas da computação e da ciência da informação para expressar a especificação de uma conceitualização. “*Uma ontologia é uma especificação formal explicita de uma conceitualização compartilhada”* (GRUBER, 1995). Essa definição é considerada como um senso comum para a área (GUARINO, 1998). Segundo a definição, os conceitos são abstrações simplificadas extraídas de um domínio específico. Uma ontologia, normalmente, consiste de um conjunto de classes organizadas hierarquicamente descrevendo um domínio (GÓMEZ-PÉREZ, BENJAMINS, 1999). Essa descrição é explícita, pois não está embutida em uma ferramenta, tendo suas propriedades e relações definidas e acessíveis. A ontologia é compartilhada ao representar um consenso sobre o conhecimento obtido sobre esse domínio. A ontologia fornece um esqueleto sobre o qual se estrutura uma base de conhecimento. Os conceitos, instâncias, relações e restrições devem ser representadas de modo explícito, utilizando-se de uma estrutura formal, isto é, que possa ser reconhecida e interpretada por entidades de software.

Para se trabalhar com ontologias é necessário conhecer a sua estrutura formal. A representação do conhecimento na Web Semântica herda o seu formalismo das pesquisas em Inteligência Artificial, e utiliza, por exemplo, elementos presentes nas redes semânticas, sistemas de frames e lógica de descrição (DING et al., 2007). Este trabalho utiliza a mesma notação adotada por (MAEDCHE, 2002).

Definição 7: Componentes formais de uma ontologia

$$\mathcal{O} \coloneqq (C, R, H^C, A^{\mathcal{O}})$$

onde:

- C é o conjunto de conceitos da ontologia. Os conceitos representam as entidades do domínio sendo modelado. Eles são designados por um ou mais termos em linguagem natural e normalmente são referenciados dentro da ontologia por um identificador único.
- $\mathrm{H^C} \subseteq \mathrm{C} \times \mathrm{C}$ é um conjunto das relações taxonômicas entre os conceitos que definem uma hierarquia ou taxonomia dos mesmos. Mais formalmente $\mathrm{H^C}$ é definido como $\mathrm{H^C} \coloneqq \{(\mathrm{c_i}, \mathrm{c_j}) | \mathrm{c_i}, \mathrm{c_j} \in \mathrm{C} \wedge \mathrm{c_i} \sqsubseteq \mathrm{c_j}\}$.
- R é o conjunto das relações não taxonômicas entre os conceitos.
- $A^{\mathcal{O}}$ é um conjunto de axiomas, normalmente formalizados em alguma linguagem lógica. Tais axiomas são regras que permitem checar a consistência da ontologia e deduzir novos conhecimentos a partir da ontologia através de algum mecanismo de inferência.

Uma ontologia é responsável por especificar a conceitualização de um domínio. As entidades e relações de um domínio, bem como os fatos a respeito dessas entidades são representados em termos dessa conceitualização. Uma base de conhecimento reúne a conceitualização e as entidades com suas diversas relações em um único repositório. Mais uma vez utilizando a notação dada por (MAEDCHE, STAAB, 2002).

Definição 8: Componentes de uma base de conhecimento

$$KB \coloneqq (\mathcal{O}, \mathcal{I}, inst, inst^R)$$

onde:

- $\mathcal{O}$ é uma ontologia;
- $\mathcal{I}$ é um conjunto de instâncias;
- $inst$ é uma função $inst: I \rightarrow C$ que mapeia instâncias de $\mathcal{I}$ para conceitos;
- $inst^R$ é uma função $inst^R: I \rightarrow R$ que mapeia instâncias para relações da ontologia $\mathcal{O}$.

2.6.5 OWL

As linguagens para a descrição de ontologias devem ter características especiais (ANTONIOU, HARMELEN, 2004):

- Sintaxe bem definida: processável por máquinas;
- Semântica formal: expressar significado de modo preciso;
- Suporte eficiente a raciocínio: validação do conhecimento;
- Expressivo poder de expressão: relações complexas.

OWL (*Ontology Web Language*) (ANTONIOU, HARMELEN, 2004) (BECHHOFER et al., 2004) é uma linguagem para a descrição de ontologias recomendada pelo W3C para aplicações para a Web Semântica. É uma forma de representação mais rica do que RDF permitindo descrever relações entre as classes e características das propriedades, disjunção entre classes, cardinalidade, igualdade e enumerações. A expressividade do OWL permite representar conhecimento ontológico e possibilita a aplicação de raciocínio lógico por entidades computacionais. A Figura 9 mostra um exemplo com informações sobre a classe "*aluno"*, sendo esta uma sub classe de "*pessoa*" e equivalente à classe "*estudante"*.

```
<owl:Class rdf:ID="aluno">
<rdfs:subClassOf rdf:resource="#pessoa"/>
<owl:disjointWith rdf:resource="#professor"/>
<owl:equivalentClass rdf:resource="#estudante"/>
</owl:Class>
```

Figura 9 Exemplo de sintaxe OWL

A OWL se divide em três sub-linguagens (ANTONIOU, HARMELEN, 2004) (DING et al., 2007) (BECHHOFER et al., 2004) que variam de acordo com o seu poder expressivo e necessidade de suporte a raciocínio:

- OWL Lite: É a versão mais simples de OWL. Permite classificação hierárquica e a definição de restrições simples.
- OWL DL: É uma versão da OWL que possui uma correspondência com a lógica de descrições. Limitam-se algumas construções como representar uma classe como uma instância ou propriedade e vice-versa. Preserva-se, no entanto, o suporte eficiente ao raciocínio.
- OWL Full: Versão completa do OWL que garante toda a expressividade e total compatibilidade com o RDF. Como desvantagem, o suporte ao raciocínio pode se tornar um problema intratável.

2.7 RECUPERAÇÃO BASEADA EM CONHECIMENTO

A recuperação baseada em conhecimento é uma subárea da recuperação de informação que trabalha com uma coleção de documentos representados de modo estruturado, que diferem dos documentos tradicionais da Web que não possuem estruturação ou são semi-estruturados, e de dados relacionais que trabalham em um domínio fechado com uma conceitualização implícita.

Em (MANNING, RAGHAVAN, SCHÜTZE, 2008), a recuperação de informação em documentos estruturados é chamada "*Recuperação em XML"*, pois o XML é a unidade fundamental de representação dos dados, situando-se na base da arquitetura em camadas da Web Semântica. O termo, ao considerar apenas o aspecto tecnológico, não engloba os potencialidades das demais camadas que oferecem conceitos como classes, relacionamentos, inferência e regras que podem

auxiliar nos processos de recuperação de informação. No entanto, novos desafios são colocados quando se trabalha com documentos estruturados:

- Recuperar todo o documento ou partes de um documento?
- Uma vez que é possível segmentar a informação, deve-se procurar a parte do documento que atenda exatamente o que foi solicitado na consulta do usuário?
- Como indexar as diversas partes de um documento com a finalidade de responder ao usuário de modo preciso?
- Consultas baseadas em XML não geram um conjunto de respostas ordenado;
- Como o usuário pode tomar proveito da estruturação sem a necessidade de conhecer como as informações são estruturadas?
- Como lidar com o conhecimento utilizado na criação de conteúdo que não é padronizado? É possível encontrar ontologias concorrentes e distintas sobre um mesmo domínio e criar conteúdo a partir dela.

Tais questionamentos não possuem uma única resposta e muitos trabalhos têm sido apresentados, sendo que alguns destes serão mostrados no capítulo sobre a recuperação de informação na Web Semântica.

2.8 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentados conceitos gerais sobre um modelo de recuperação de informação, mostrando os seus componentes e como são instanciados em modelos clássicos de recuperação de informação, apontando vantagens e desvantagens de cada modelo. Os modelos tradicionais, mesmo com as diversas extensões propostas (BAEZA-YATES, RIBEIRO-NETO, 1998), possuem limitações em sua efetividade em ambientes como a Web atual devido à estrutura de representação da informação utilizada. Limitações essas que são compartilhadas por outras áreas do tratamento da informação que não somente a recuperação. Por isso, novas estruturas de representação de informação são propostas para criação da Web Semântica.

Conforme apresentado neste capítulo, diversas tecnologias, tais como o XML, RDF e OWL, já estão disponíveis para a utilização de desenvolvedores, para o aumento da efetividade das aplicações atuais e a criação de novas aplicações baseadas em conhecimento. Em particular, sistemas de recuperação de informação devem trabalhar sobre um novo paradigma, tratado nesse trabalho, que é a recuperação baseada no conhecimento. Os tópicos apresentados não foram exaustivos, apresentaram apenas as características principais de cada tópico a fim de ressaltar como essas tecnologias interagem e contribuem para a representação da informação. No capítulo seguinte, são mostrados diversos sistemas de recuperação criados para a Web Semântica enfatizando as técnicas aplicadas.

3. RECUPERAÇÃO DE INFORMAÇÃO NA WEB SEMÂNTICA

Desde o surgimento da Web Semântica, os novos padrões propostos para a representação de conhecimento na Web foram aproveitados pela área de recuperação de informação como um meio de transpor as limitações encontradas com o uso de fontes desestruturadas. A ambigüidade de alguns termos pesquisados e a dificuldade para encontrar a informação desejada no conjunto de resultados retornados são problemas difíceis de resolver apenas com o uso de operações sobre o conteúdo textual (ANTONIOU, HARMELEN, 2004).

Neste capítulo são apresentados diversos sistemas construídos a partir dos padrões da Web Semântica ou endereçados à busca de conhecimento representado segundo esses padrões. Alguns destes sistemas já aparecem como representativos em outros levantamentos (SCHEIR, PAMMER, LINDSTAEDT, 2007) (WEI, BARNAGHI, BARGIELA, 2008), enquanto outros foram incluídos devido ao uso de novas abordagens ou freqüentemente citados como relevantes nos trabalhos pesquisados. Após a apresentação dos sistemas e da definição do modelo é mostrada uma discussão sobre estas características e um estudo comparativo sobre os componentes de cada sistema segundo a arquitetura genérica de um processo de recuperação de informação (BAEZA-YATES, RIBEIRO-NETO, 1998).

Dentre os sistemas de recuperação de informação disponíveis segundo os critérios citados destacam-se os seguintes:

3.1 SIMPLE HTML ONTOLOGY EXTENSIONS (SHOE)

SHOE (HEFLIN, HENDLER, 2000) é uma linguagem baseada nos padrões SGML e XML, proposta em 1998, com o intuito de embutir marcações semânticas em páginas HTML. As marcações permitem associar às páginas da Web: conceitos, instâncias de ontologias, valores literais e alegações sobre o conteúdo anotado. As marcações são, necessariamente, instâncias de ontologias disponibilizadas pelo sistema. Uma ferramenta para anotação de páginas é distribuída aos usuários para garantir a correção das marcações. O sistema possui uma implementação para a Web que pesquisa páginas anotadas em linguagem SHOE. As marcações pesquisadas são armazenadas em uma base de

conhecimento que utiliza a estrutura oferecida pela ferramenta Parka KB (HENDLER et al., 1996), capaz de realizar inferências sobre o conteúdo adquirido.

O usuário cria a consulta a partir de uma interface proprietária conforme mostrado na Figura 10. SHOE permite a utilização de múltiplas ontologias, mas cada consulta utiliza apenas uma delas como contexto para a busca submetida.

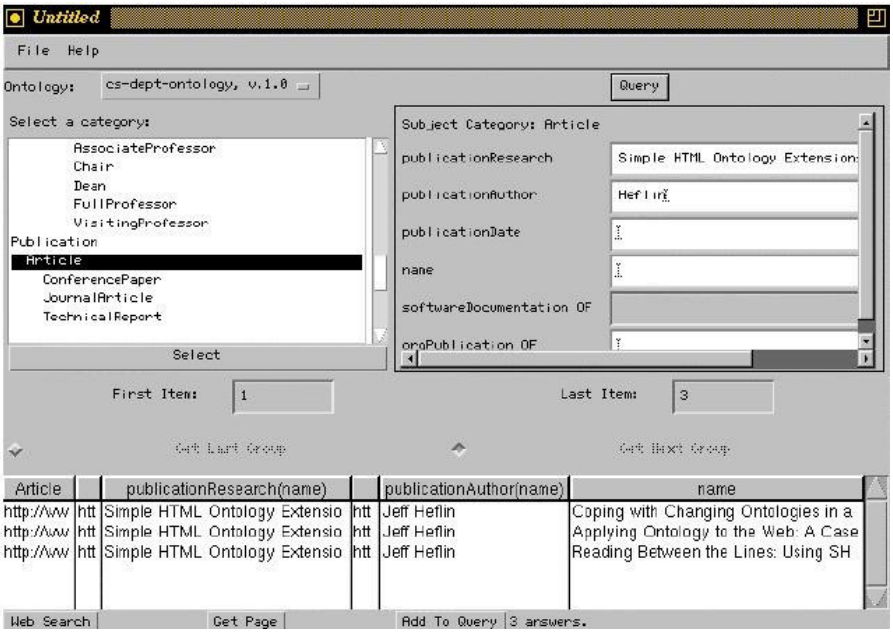

Figura 10 Interface de consulta do SHOE (HEFLIN, HENDLER, 2000)

Nesta interface, o usuário seleciona uma das ontologias da lista colocada na parte superior da interface, uma ou mais categorias dessa ontologia e valora as propriedades dessas categorias no quadro à direita. A consulta é transformada em uma expressão formal de pesquisa, neste caso a sintaxe é a Parka KB, e submetida à base de conhecimento, retornando os documentos encontrados na parte inferior da interface. Ou seja, o formulário esconde a complexidade da linguagem de consulta do usuário final. Por outro lado, o conhecimento prévio das ontologias disponíveis é necessário para a correta especificação do contexto.

Sendo um dos primeiros métodos propostos para a Web Semântica, o SHOE utiliza um modelo estruturado em cima dos padrões da base Parka KB. O desenvolvimento de padrões recomendados pelo W3C para a Web Semântica desencorajou o uso da linguagem proposta.

3.2 TAP

TAP (GUHA, MCCOOL, MILLER, 2003) é um framework para o desenvolvimento de aplicações de busca semântica. As aplicações de busca semântica, segundo a proposta, devem melhorar os resultados obtidos dos mecanismos de busca tradicionais das seguintes formas:

- Adicionando aos resultados informações extraídas das marcações semânticas;
- Contextualizando e denotando os recursos pesquisados.

O framework TAP representa conceitos e instâncias de domínio geral utilizando os padrões de representação de conhecimento sugeridos pelo W3C. Os conceitos utilizados pelo framework TAP são descritos por uma ontologia própria, básica e ampla sobre pessoas, lugares, organizações, eventos entre outros. A fonte de informação é proveniente de arquivos em formato RDF disponibilizados pelo W3C, somado a elementos retirados de páginas HTML através do uso de um tradutor que gera recursos e propriedades a partir do texto das páginas. A representação interna destas marcações é feita através de arquivos contendo triplas RDF associadas à URL de origem. Estas triplas formam um grafo (Figura 11), representando a base de conhecimento disponível para a consulta.

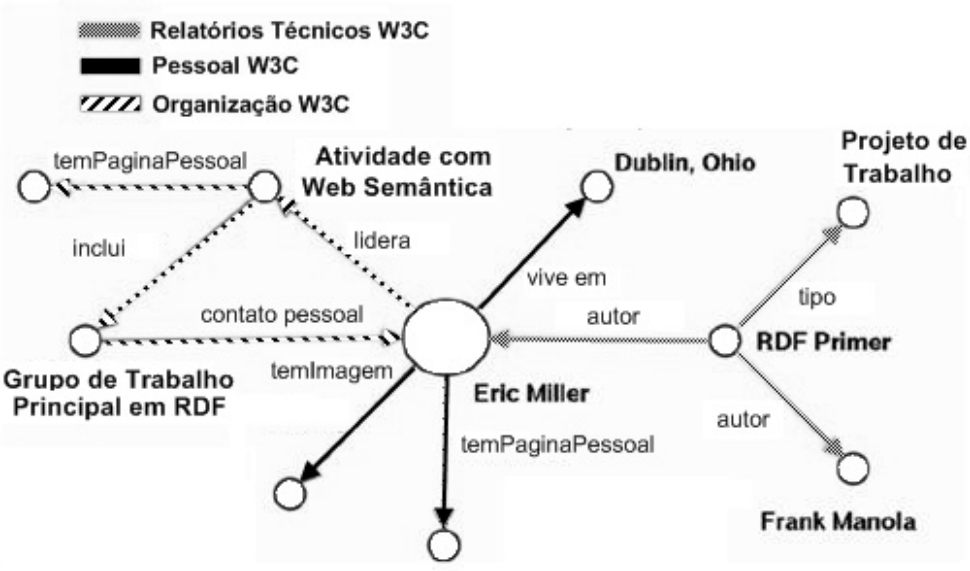

Figura 11 Grafo RDF extraído da base de dados do TAP (GUHA, MCCOOL, MILLER, 2003)

O usuário pode consultar esta base através de uma interface programável disponível no servidor TAP, chamada *GetData* que, a partir da especificação de um recurso e uma propriedade, retorna um valor para o usuário.

A consulta trabalha sempre a partir de um recurso (chamado também de classe) e uma propriedade a ser pesquisada. Para auxiliar o usuário ou aplicação que utilize o sistema, TAP disponibiliza uma ferramenta auxiliar para a identificação de itens na ontologia dado um termo textual. A consulta é processada em dois passos, sendo o primeiro a descoberta de um ou dois nodos âncora no grafo que melhor representam o recurso procurado. O segundo passo é percorrer o grafo segundo uma busca em largura, (*breadth-first order*) recuperando os recursos mais próximos ao nodo âncora. A classificação dos resultados se dá em termos da distância dos recursos recuperados para o nodo âncora. Por exemplo, uma consulta referente ao nodo $Eric\ Miller$, no grafo da Figura 11, teria como resultados, além de páginas Web recuperadas por sistemas de recuperação tradicionais, informações complementares como a atividade relacionada à pessoa pesquisada $Semantic\ Web\ Activity$, sua cidade de residência $Dublin, Ohio$ e uma obra de sua co-autoria $RDF\ Primer$.

TAP oferece uma interface simples e programável. A sua estrutura de resposta às consultas é otimizada através de métodos de indexação dos recursos e a manutenção de um registro dos recursos coletados nas páginas. O processamento da consulta utiliza uma técnica conhecida para percorrer os grafos e procura incrementar o processo com heurísticas próprias aproveitando os recursos relacionados a nodos mais populares.

3.3 KNOWLEDGE AND INFORMATION MANAGEMENT (KIM)

KIM (KIRYAKOV et al., 2004) é um sistema de recuperação de informação que realiza as funções de anotação de documentos, manutenção da base de conhecimento, indexação e recuperação de documentos.

KIM utiliza uma ontologia própria e de alto nível que define as entidades do domínio. Sendo um sistema aberto, a base de conhecimento é mais superficial, porém abrangente. Uma base de conhecimento é construída a partir desta ontologia, entidades pré-definidas obtidas a partir de fontes confiáveis e entidades obtidas

através do processo de extração de anotações. O processo de anotação percorre os documentos identificando e extraindo as “entidades nomeadas” (*named entities*), exemplificado na Figura 12. As entidades nomeadas são identificadores textuais para pessoas, lugares e organizações, bem como alguns conjuntos de valores literais.

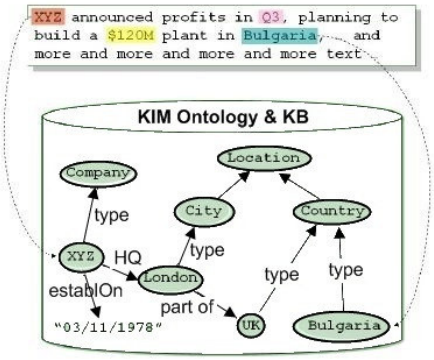

Figura 12 Anotação semântica em KIM (KIRYAKOV et al., 2004)

A representação dos documentos é feita através dos metadados, no caso as entidades nomeadas encontradas, obtidos pelo processo de anotação. Cada metadado é indexado e vinculado ao documento anotado, a uma instância da base de conhecimento e ao seu conceito mais específico na hierarquia dessa base. O formato para a base de conhecimento e anotações é o RDF e a ferramenta utilizada é o Sesame. Algumas regras são aplicadas para estender as relações da base.

As consultas são inseridas através de uma interface, do tipo formulário, que identifica relações e entidades nomeadas existentes na base de conhecimento. O processo de recuperação utiliza um framework genérico, Lucene, que realiza a busca e recuperação dos documentos com uma abordagem híbrida entre os modelos booleano e vetorial (HATCHER, GOSPODNETIC, 2005).

O principal foco do sistema KIM é o processo de anotação semântica que utiliza técnicas de processamento de linguagem natural, interfaces de apoio para anotações do usuário e módulos de atualização da base de conhecimento. O processo de recuperação adiciona termos da ontologia à representação dos documentos, o que leva a resultados mais precisos. Ao mesmo tempo, o método de recuperação é baseado na freqüência e o uso de formulário fornece uma interface para consultas simples baseadas nas entidades descritas na base de conhecimento.

3.4 ROCHA

O modelo proposto em (ROCHA, SCHWABE, ARAGAO, 2004) se fundamenta na propagação de ativação (*spread activation*) em uma rede semântica. Para isto, a base de conhecimento é vista como uma rede semântica que é descrita em termos de relações nomeadas (propriedades) que são avaliadas em termos sub-simbólicos através de pesos (Figura 13), que traduzem a intensidade do relacionamento entre os conceitos e as instâncias. Na parte superior da Figura 13 são mostradas as relações conceituais do domínio (*schema*), na parte intermediária são mostradas as instâncias dos conceitos e as suas relações na base de conhecimento (*schema instance*) e, por fim, o grafo final da base de conhecimento é criado (*instances graph*) atribuindo-se pesos às relações representadas por arcos orientados.

Os pesos associados às relações entre os conceitos são calculados a partir das seguintes fórmulas:

- Similaridade: A similaridade entre as instâncias dos conceitos é calculada por uma fórmula que compara os vínculos existentes nas instâncias, isto é, se existirem muitos vínculos comuns então as instâncias são próximas.
- Especificidade: Apresenta uma medida que indica se a ligação entre os conceitos é mais ou menos comum no grafo baseado na quantidade de instâncias vinculadas entre os conceitos.
- Combinada: Produto das medidas anteriores e que apresentou melhor resultado nos testes segundo os autores.

A rede semântica forma uma fonte de informação na qual os nodos possuem um conjunto de expressões para o conteúdo representado. O processo de busca se inicia a partir da consulta do usuário feita através de palavras-chaves. Estas palavras-chave são submetidas à base formada por nodos da rede semântica da mesma forma que um sistema de busca clássico. O resultado é um conjunto de nodos ordenados conforme a similaridade textual entre as representações. Um determinado número de nodos mais bem classificados é usado para a ativação do algoritmo e para cada um desses nodos é informado também o seu respectivo valor

de ativação. Um algoritmo de propagação de ativação é utilizado para percorrer o grafo e encontrar as instâncias mais próximas do conceito da consulta.

O método proposto gera uma representação quantitativa das relações existentes na base de conhecimento e explora os caminhos entre os conceitos para satisfazer ao critério de busca. Os aspectos explorados são a similaridade entre os conceitos e a especificidade das instâncias relacionadas.

O grafo de conceitos embute parte da semântica, mas a busca em si, não utiliza diretamente as estruturas semânticas. São gerados valores numéricos para as arestas do grafo e é difícil explicar porque um determinado documento foi recuperado, ou seja, quais os critérios que levaram o documento a ser considerado relevante. Uma vantagem do método é a capacidade de lidar com consultas mais complexas, envolvendo diversos conceitos.

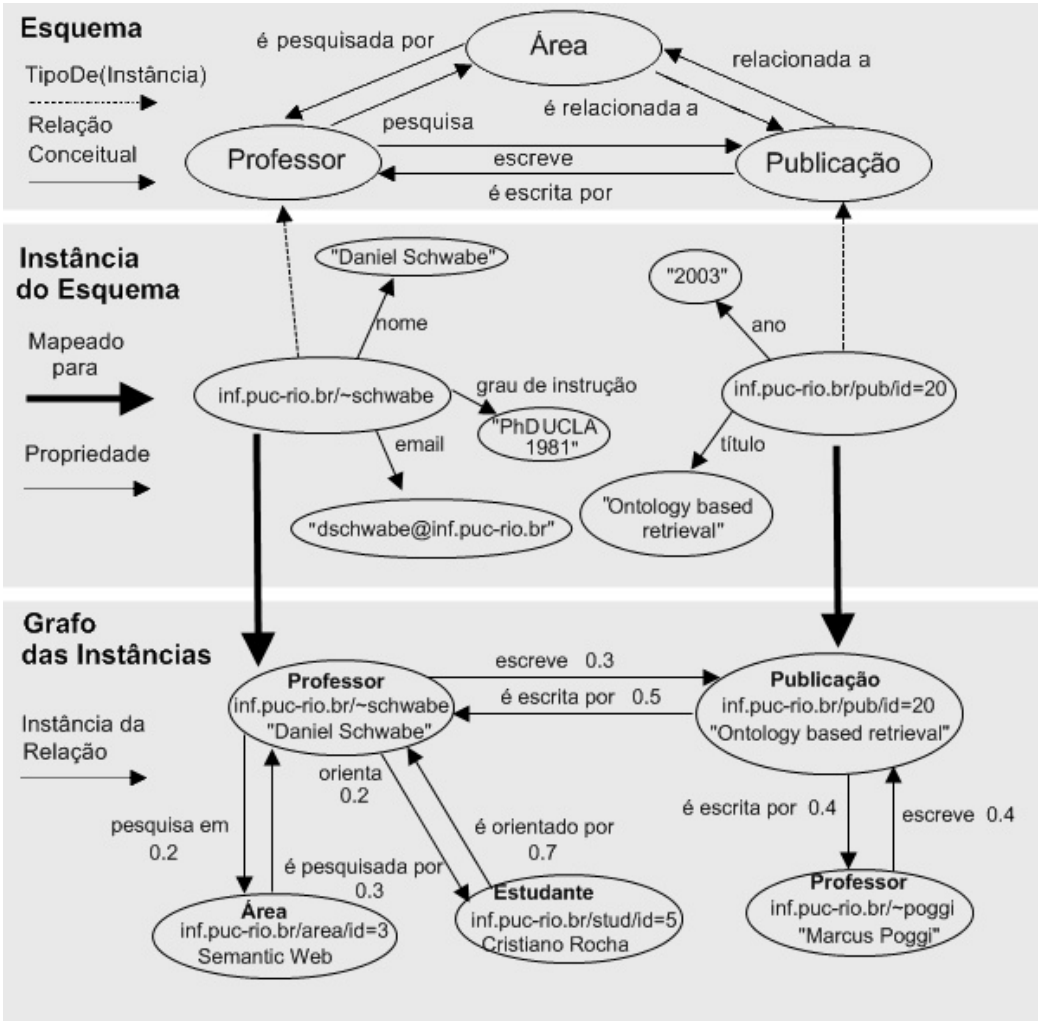

Figura 13 Exemplo de um grafo de instâncias (ROCHA, SCHWABE, ARAGAO, 2004)

3.5 DISTRIBUTED OPEN SEMANTIC ELABORATION (DOSE)

O modelo especificado para o sistema DOSE (BONINO, CORNO, FARINETTI, 2004) utiliza dois operadores que trabalham sobre o relacionamento hierárquico entre os conceitos: generalização e especialização (*focalization*). O método é baseado no modelo vetorial e permite a expansão da consulta pelo acréscimo de novos termos relevantes. As heurísticas desenvolvidas até então operam somente sobre os relacionamentos hierárquicos.

A proposta é um método de expansão da consulta baseado na navegação ontológica. A consulta é formulada com termos, pesos e um número de documentos relevantes desejados. O objetivo é recuperar, ao menos, o número de documentos desejados pela aplicação que submete a consulta. Caso a consulta não atinja o número de documentos desejados, os termos da consulta são avaliados para determinar a necessidade de generalização ou especialização. Neste último caso, para cada conceito é verificado se o número N_{rc} de documentos recuperados para o conceito excede um parâmetro de navegação O pré-estabelecido. Se sim, então o termo não é expandido. Se não, ou seja, $N_{rc} < O$, temos um segundo parâmetro F tal que:

- Se $N >= F$ então o conceito é expandido aplicando-se a propriedade “especialização”.
- Se $N < F$ então o conceito é expandido aplicando-se a propriedade “generalização”.

Se o total é maior que o parâmetro então se necessita complementar os resultados com valores mais específicos. Se o total recuperado é um número menor que o parâmetro F então se necessita de uma busca mais ampla.

O sistema DOSE pode ser encarado como um método de expansão da consulta, uma vez que o modelo de busca principal continua sendo o vetorial. O método estabelece ações para o sistema baseado em parâmetros de satisfação do usuário pré-definidos. Estas ações realizam inferência, até o momento baseadas na taxonomia, na base de dados para melhorar os resultados.

3.6 AQUALOG

AquaLog (LOPEZ, PASIN, MOTTA, 2005) é um modelo que combina técnicas de processamento de linguagem natural (PLN) e ontologias para processar consultas expressas em linguagem natural sobre uma base de documentos provendo respostas precisas às questões ao invés de apresentar documentos relevantes.

As marcações semânticas dos documentos são realizadas automaticamente pelo AquaLog através de técnicas de mineração de texto.

A arquitetura do sistema representa a consulta do usuário em termos de triplas RDF de acordo com um formato padrão, similar a asserções em lógica, para a representação de consultas: <sujeito, predicado, objeto>. Adicionalmente, AquaLog classifica as triplas segundo categorias pré-definidas. Essas categorias irão determinar a forma de processamento e resposta da consulta. As triplas RDF são então refeitas em termos dos conceitos presentes na ontologia de domínio. Por último, as triplas são processadas junto à base de conhecimento para obtenção da resposta.

O AquaLog possui um método complexo, pois envolve o processamento de linguagem natural, reconhecimento de entidades e processos de inferência para a tradução da consulta em termos da representação ontológica (Figura 14). O foco não é a recuperação de informação e sim a montagem de um sistema de perguntas e respostas sobre tópicos de um domínio. Em avaliações preliminares o sistema se mostrou capaz de resolver questões mais simples. Entretanto, mostrou uma dependência das técnicas de PLN para a obtenção de melhores resultados.

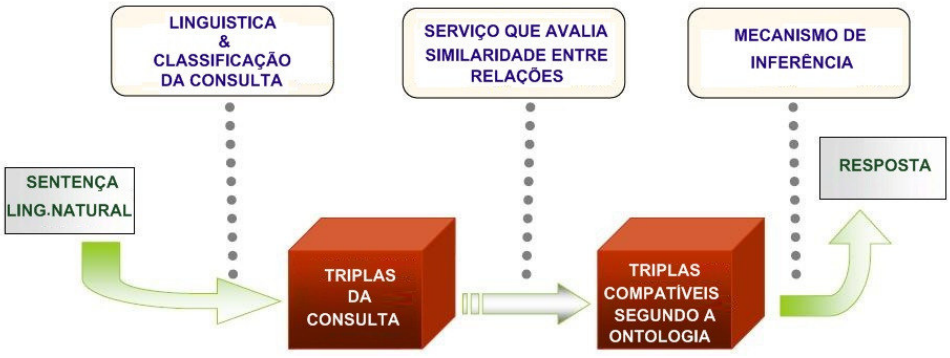

Figura 14 Modelo de processamento do AquaLog (LOPEZ, PASIN, MOTTA, 2005)

3.7 SEMANTIC SEARCH (SEMSEARCH)

A idéia central por trás do SemSearch (LEI, UREN, MOTTA, 2006) é facilitar a formulação da consulta por parte do usuário. Para isto, o sistema se baseia no padrão de consulta do Google e, em seguida, realiza um mapeamento dos termos da consulta para os elementos do domínio. A consulta é expressa a partir de uma palavra chave especial que denota o assunto (conceito) consultado e uma combinação das demais palavras chaves. O conceito é identificável através do operador ":" e a combinação de termos pode conter os operadores lógicos "and" e "or". O padrão é dado da seguinte forma: "subject:keyword1 and/or keyword2 and/or keyword3 ...". Por exemplo, para pesquisar notícias sobre a eleição americana podemos formular a seguinte consulta: "notícias:eleições Obama". Tanto o assunto quanto as palavras chave são mapeados para elementos do domínio comparando-os aos rótulos identificadores de cada elemento. Valores literais também são utilizados nessa comparação de padrões. As entidades semânticas podem ser conceitos, instâncias e propriedades. A comparação citada é feita de modo simplificado e não utiliza estruturas taxonômicas como o *WordNet* (MILLER et al., 2006) aproveitando elementos da sintaxe para obter uma representação mais exata dos termos.

A partir das entidades semânticas extraídas da busca, são formuladas consultas à base de conhecimento de modo a obter uma resposta precisa e auto-explicativa para a consulta do usuário. As formas de consulta são classificadas como simples e complexa.

- **Consulta Simples**: São consultas formuladas com no máximo duas palavras chave. Uma vez encontradas as entidades na base, são gerados pares: conceito x palavra chave. De acordo com a combinação do tipo das entidades semânticas dos elementos do par, ou seja, a classe, instância ou relacionamento, é escolhido o gabarito (*template*) apropriado para que a consulta possa recuperar as respostas com os elementos esperados. Por exemplo: se o assunto for uma classe e a palavra chave uma classe devem ser recuperadas todas as relações diretas entre instâncias destas duas classes. As consultas são geradas na linguagem Sesame SeRQL (ou outra linguagem formal para bases de

conhecimento baseadas em RDF) e o operador de união utilizado para gerar o resultado final agrupando os resultados de cada combinação das palavras-chave.

- **Consulta Complexa**: É realizada da mesma forma que a consulta simples, mas respeitando-se algumas limitações para atender a requisitos de desempenho. O número de consultas combinando-se todos os termos é o produto entre o número de casamentos obtidos para cada termo. Para reduzir o total de consultas geradas adotam-se algumas heurísticas, em especial: limitar a quantidade de mapeamentos entre o termo informado e a sua respectiva classe na base, e preencher os gabaritos apenas com as classes mais específicas (*msc*) de cada instância.

A classificação dos resultados é dada em função da distância entre o termo procurado e a classe encontrada para representá-lo e em função do número de termos da consulta que o resultado satisfaz.

Uma grande vantagem da abordagem é a facilidade de expressão da consulta. Porém, a sua tradução, apesar de simples, gera um grande número de combinações para consultas complexas.

3.8 OWL INFORMATION RETRIEVAL (OWLIR)

OWLIR (DING et al., 2005) (SHAH, FININ, JOSHI, 2002) tenta integrar diferentes visões da Web: documentos e anotações. Para isto, os documentos são representados de modo híbrido incluindo texto e marcações semânticas.

O sistema foi desenvolvido e testado em um domínio específico: eventos. Uma ontologia, em DAML+OIL, foi construída a partir de uma definição do domínio conhecida para suportar os processos de anotação e busca da ferramenta. Tanto a consulta do usuário quanto as descrições do evento são anotadas segundo esta ontologia.

A extração das anotações é feita a partir da identificação de dados relevantes do texto do documento pela ferramenta AeroText. Através da API Java da ferramenta é possível criar as triplas RDF com as marcações extraídas. Os itens presentes nas triplas possuem um formato proprietário sob o qual termos e marcações são representados de modo único e indistinto. O mecanismo de inferência na indexação é criado a partir de regras em JESS para raciocinar sobre as

instâncias e conceitos da ontologia. Por exemplo, um filme do homem-aranha pode gerar uma tripla informando se tratar de um filme de ação. O fluxo geral do processo é mostrado na Figura 15.

O sistema de recuperação é criado sobre a API do HAIRCUT que permite ao OWLIR personalizar o esquema de indexação de documentos para o seu formato de representação. A API possui ainda interfaces para busca por vários métodos entre eles o booleano e o vetorial. As consultas podem ser feitas informando itens requeridos, permitidos e não permitidos nos documentos procurados. Podem ser pesquisados tanto marcações específicas quanto texto livre, ou uma combinação de ambos. OWLIR utiliza o mecanismo de recuperação tradicional para encontrar os documentos relevantes e, em seguida, aplica a consulta formal para obter os resultados finais.

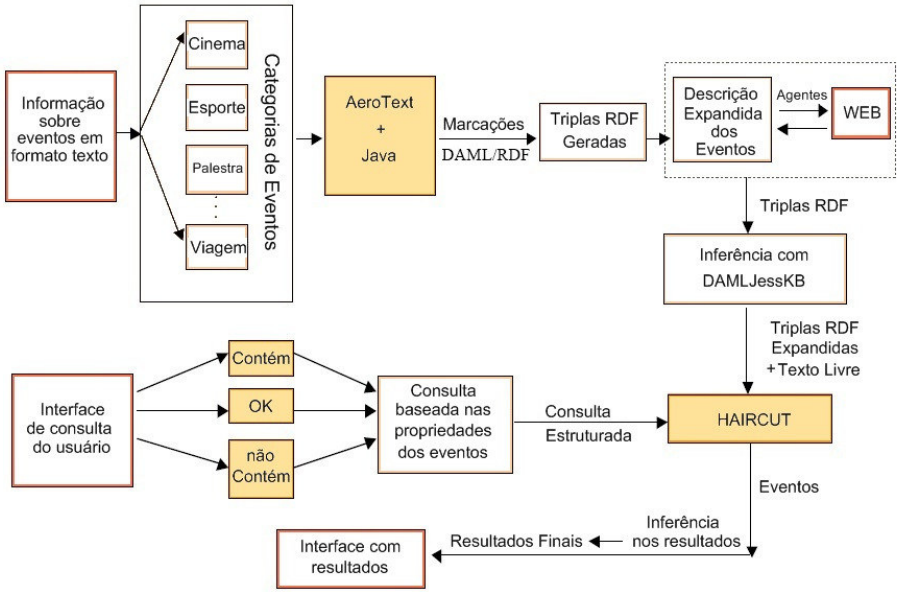

Figura 15 Fluxo de processo em OWLIR (SHAH, FININ, JOSHI, 2002)

3.9 BEAGLE++

Beagle++ (CHIRITA et al., 2006) é um sistema de busca que utiliza a representação semântica para incrementar a precisão em relação aos resultados obtidos pelo sistema de recuperação tradicional Beagle. O domínio de aplicação de ambos os sistemas é a busca em discos rígidos de computadores pessoais, e para isso, o Beagle++ utiliza uma ontologia contextual que descreve o conhecimento

sobre a estrutura de mensagens eletrônicas, arquivos do sistema operacional e publicações científicas.

A fonte de informação utiliza pelo sistema Beagle++ é composta por documentos instanciados a partir da ontologia contextual e inclui: mensagens eletrônicas, arquivos e páginas Web armazenadas localmente no computador. Cada documento é representado no formato RDF e são mantidos índices para os metadados e para os termos em texto puro. Os dados sobre os arquivos são primeiramente capturados através do conteúdo das mensagens e de campos semi-estruturados. Posteriormente, são adicionadas à fonte de informação dados, dispostos em triplas RDF, sobre a atividade do usuário capturados através dos metadados associados a cada documento. Essa captura é possível com o monitoramento da atualização de documentos a partir de eventos notificados pelo kernel do Linux.

A análise de similaridade do Beagle++ utiliza um sistema de ordenação (*ranking*) que combina o modelo tradicional, que utiliza o método de cálculo de freqüência dos termos nos textos dos documentos, e o esquema de classificação dos metadados chamado "*Object Rank*". Para calcular o "*Object Rank*", é necessário pontuar como cada elemento do domínio influencia os demais. Uma ontologia é transformada em um grafo com setas e pesos para os relacionamentos entre os recursos de modo a refletir a transferência de responsabilidade sobre a informação. A pontuação para cada recurso é dada por uma fórmula similar à fórmula do *Page Rank* (BRIN, PAGE, 1998). O valor final de ordenação é dado pelo produto do resultado de ambos os esquemas de classificação para um único recurso.

3.10 QUIZRDF

O QuizRDF (DAVIES, KROHN, WEEKS, 2002) é proposto como um sistema misto, que combina a pesquisa por metadados com a pesquisa textual. Seus autores consideram que a Web possui um número reduzido de páginas com anotações. Portanto, o sistema deve ser capaz de trabalhar com e sem anotações. Consideram também que a representação RDF não substitui o documento original. O uso de triplas RDF como fonte de recuperação produz um índice de recuperação baixo e indesejável no início da busca.

O sistema mescla termos e metadados em um único índice com elementos representados através de triplas RDF para obter resultados mais precisos. As triplas guardam a classe do documento, sua identificação na base de conhecimento e o valor de suas propriedades. Páginas não anotadas também são representadas utilizando-se a classe e uma propriedade padrão para representar os valores literais encontrados.

O esquema para a submissão da consulta considera que o usuário utiliza uma mistura de busca e exploração para obter a informação desejada. O usuário, neste caso, seria capaz de explorar apenas parte do conhecimento descrito em uma ontologia.

A consulta pode ser feita através da seleção de classes da ontologia ou a partir de consulta textual. A consulta através de uma classe permite recuperar os documentos cuja representação possua a classificação escolhida. A consulta textual, por outro lado, realiza uma pesquisa pelo método de freqüência de termos no conjunto de anotações trazendo uma lista ordenada com os documentos encontrados. O sistema examina a representação dos documentos recuperados para identificar as classes por eles referenciadas. Esta lista de classes é apresentada ao usuário para que este possa refinar a consulta por um critério mais específico filtrando classes, valorando propriedades e explorando os documentos relativos a um determinado conceito.

3.11 VALLET

Vallet et al (VALLET, FERNÁNDEZ, CASTELLS, 2005) apresentaram um modelo de recuperação baseado em ontologias. Neste modelo são definidas três classes principais em uma ontologia: conceitos do domínio em si, classificação taxonômica que define uma categoria (ou área de conhecimento) para o assunto e a classificação do documento que se refere ao formato de distribuição do texto tal como mensagem eletrônica, notícia, relatório e outros. Cada documento é representado por instâncias da classe *Annotation* que possui propriedades relativas ao processo de anotação e instâncias das três classes principais.

As anotações são utilizadas nos processos de recuperação e ordenação dos resultados. Desta forma, são associados pesos para cada instância da

representação do documento baseado na freqüência de ocorrência dos mesmos no documento.

A consulta é inserida no sistema como uma expressão RDQL que pode referenciar qualquer combinação de conceitos presentes na hierarquia das três classes principais. Tal consulta é submetida à base de conhecimento do sistema e um conjunto de instâncias é recuperado. Este conjunto serve como termos em uma consulta processada na base de documentos que recupera todos os documentos anotados com tais instâncias. Um processo de inferência na base baseado em regras e na hierarquia dos conceitos permite expandir a consulta. A classificação é feita por uma fórmula que combina a freqüência de ocorrência das instâncias com a freqüência de ocorrência dos termos em texto livre extraídos da consulta original em RDQL. Uma visão geral do processo pode ser vista na Figura 16.

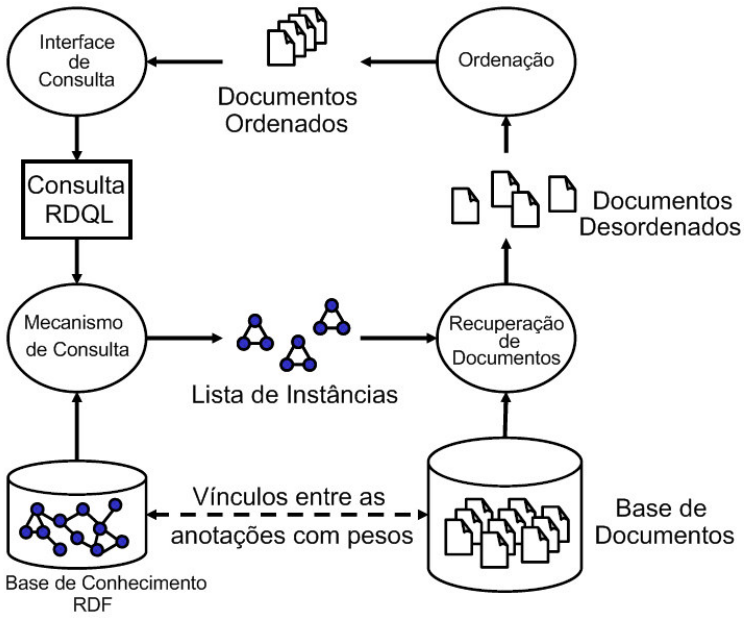

Figura 16 Visão geral do processo de recuperação (VALLET, FERNÁNDEZ, CASTELLS, 2005)

O uso de texto livre no modelo é justificado por três argumentos: o problema da conversão dos textos atuais em metadados, o grau de acuidade da representação em metadados em relação ao texto original, ou seja, haveria uma perda de informação após o processo de representação e, por último, a própria ocorrência de texto livre em elementos da base de conhecimento.

3.12 SIM-DL

SIM-DL (JANOWICZ, 2006) é uma medida de similaridade para a lógica de descrição que pode ser usada por sistemas de recuperação de informação geográfica. Apesar de não descrever um sistema de recuperação por completo, este trabalho é significativo por duas razões. Em primeiro lugar por utilizar uma definição de contexto, para os conceitos a serem comparados, similar à do modelo aqui proposto. Em segundo lugar, pelo fato da medida de similaridade comparar conceitos a partir de sua definição em lógica de descrição e não apenas baseado na distância semântica.

Os autores definem um conjunto de passos para a aplicação da medida de similaridade (JANOWICZ et al., 2007). Dentro deste conjunto, o primeiro passo é selecionar quais os conceitos estarão aptos a serem comparados com um conceito procurado. Para isto, é utilizado um contexto no qual os conceitos alvo devem ser necessariamente subconceitos de um conceito geral definido manualmente pelo usuário ou extraído por uma análise de outros elementos contextuais da consulta.

A medida descrita em SIM-DL calcula a similaridade entre dois conceitos C e D a partir das disjunções $C_1 \sqcup \ldots \sqcup C_n$ e $D_1 \sqcup \ldots \sqcup D_n$. A fórmula utilizada, mostrada na Definição 9, compara as definições de ambos os conceitos somando o valor da similaridade calculada para os pares com a mesma estrutura semântica. Um peso ω é utilizado para indicar a importância relativa do par comparado. Este peso pode ser obtido, por exemplo, com o uso de alguma função de probabilidade.

Definição 9: Similaridade entre dois conceitos proposta em (JANOWICZ, 2006).

$$sim_U(C, D) = \sum_{(C_i, D_j) \in SI} \omega_{ij} * sim_i(C_i, D_j)$$

A similaridade entre um par destes conceitos é dada em função da similaridade entre os componentes de um conceito em lógica de descrição conforme a equação da Definição 10:

Definição 10: Similaridade entre os componentes de um conceito proposta em (JANOWICZ, 2006).

$$sim_i(C,D) = \frac{1}{\sigma}\Bigg(\sum_{(A,B)\in SP} sim_p(A,B) + \sum_{(R,S)\in SE} sim_e\big(exists_R(C), exists_S(D)\big) + \sum_{(R,S)\in SF} sim_f\big(forall_R(C), forall_S(D)\big) + \sum_{(R,S)\in SMIN} sim_m\big(min_R(C), min_S(D)\big) + \sum_{(R,S)\in SMAX} sim_m\big(max_R(C), max_S(D)\big)\Bigg)$$

Desta forma, um conceito formado a partir de conceitos atômicos é comparado a um conceito similar formando um par válido para a aplicação da medida de similaridade. O mesmo critério é aplicado a papéis, operadores existenciais, restrições por valor e restrições de cardinalidade.

3.13 ESTUDO COMPARATIVO DOS SISTEMAS

As características de alguns dos sistemas mostrados neste capítulo já foram comparadas (DING et al., 2007) (HEFLIN, HENDLER, 2000) segundo as principais técnicas e os resultados esperados. Encontramos nestes sistemas uma diversidade de técnicas e de abordagens empregadas para resolver problemas tais como incrementar as taxas de revocação, a capacidade de responder questões objetivas, a obtenção de resultados mais precisos e a entrega de informação estruturada além daquela presente no texto dos documentos em si. No entanto, não existe uma definição clara sobre como deve ser um sistema de recuperação de informação para a Web Semântica (DING et al., 2007).

Esse estudo comparativo procura identificar como cada técnica utilizada nos sistemas apresentados contribui para melhorar cada componente de um sistema de recuperação de informação. Os principais componentes são a representação interna das consultas e documentos, as operações para a recuperação e a função para a ordenação dos resultados (BAEZA-YATES, RIBEIRO-NETO, 1998). As

seções 3.13.1, 3.13.2 e 3.13.3 detalham as características principais que contribuem para a melhoria de cada componente do processo.

3.13.1 Representação Interna

O uso de palavras-chaves na representação interna dos itens de informação é uma característica ainda presente nos sistemas avaliados. Por um lado, existe o fato de que grande parte do conteúdo disponível, seja na Web ou em outros repositórios, não possui marcações semânticas conforme desejado. Mesmo o conteúdo já anotado com marcações enfrenta desafios para o seu uso efetivo. Por exemplo, como garantir a fidedignidade dos metadados dada à atualização do conteúdo dos documentos? Além disso, a geração automática destes metadados pode não ser suficientemente precisa, ou ao menos tão precisa quanto os termos dos documentos. Por fim, dentro da própria representação semântica, temos a presença de valores literais, expressos em texto livre, que podem servir como fonte de informação relevante (VALLET, FERNÁNDEZ, CASTELLS, 2005). Por outro lado, o uso de conceitos permite a criação de interfaces de consulta programáveis, isto é, interfaces utilizáveis por agentes de software, realizar inferências sobre o conteúdo representado gerando novo conhecimento e oferece contextos que auxiliam na resolução de ambigüidades na interpretação do conteúdo.

Ao comparar a forma de representação dos documentos nos sistemas pesquisados (Tabela 2), notamos que a grande maioria utiliza processos de inferência na criação da representação interna. Esses processos partem de um conjunto de termos inicialmente extraídos dos documentos e utilizam uma ontologia ou uma base de conhecimento para a obtenção de novos termos que complementem e expandam a representação.

Em diversos sistemas ainda são utilizadas palavras-chave na representação interna. Os sistemas QuizRDF (DAVIES, KROHN, WEEKS, 2002) e OWLIR (DING et al., 2005) (SHAH, FININ, JOSHI, 2002) são abordagens híbridas que utilizam as técnicas da Web Semântica na etapa de representação dos elementos de informação mesclando palavras-chave e conceitos nos itens de informação da coleção. Nestes sistemas, os documentos são conjuntos de triplas no padrão RDF que guardam tanto o texto puro quanto os conceitos. Estes conjuntos

são expandidos através do uso de inferência permitindo a ampliação da base de documentos disponível para a recuperação. Rocha et al (ROCHA, SCHWABE, ARAGAO, 2004) apresentaram um sistema em que os atributos de conceitos e instâncias são utilizados como texto, da mesma forma que os modelos clássicos de recuperação, para representar nodos em uma rede semântica.

Tabela 2 Comparação das características da representação interna nos sistemas de recuperação analisados

Sistema	Palavras-Chave	Conceitos	Instâncias	Representação	Inferência
SHOE	N	S	S	KB	N
TAP	N	S	S	Grafo	S
KIM	N	S	S	Vetor	S
ROCHA	N	S	S	Grafo	S
DOSE	N	S	N	Vetor + KB	S
AQUALOG	N	S	S	RDF	S
SEMSEARCH	N	S	S	Vetor + KB	S
OWLIR	S	S	S	RDF	S
BEAGLE++	S	S	S	Vetor + Grafo	S
QUIZRDF	S	S	S	RDF	N
VALLET	S	S	S	Vetor	S
SIM-DL	N	S	S	KB	N

BEAGLE++ (CHIRITA et al., 2006) e VALLET (VALLET, FERNÁNDEZ, CASTELLS, 2005) também são abordagens híbridas, porém trabalham os conceitos e instâncias separadamente das palavras-chave na representação interna. Eles mantêm, conceitualmente, a representação textual em paralelo à representação semântica. Dados estatísticos sobre ambas as formas de representação são combinadas na análise de similaridade desses sistemas. Todos os demais sistemas apresentados representam os documentos em função dos conceitos e instâncias presentes na base de conhecimento.

As opções de representação que mantêm de alguma forma a representação textual se apóiam no fato de que a geração e manutenção automática de metadados confiáveis ainda é uma realidade distante (MAEDCHE, STAAB, 2002). Porém, o desenvolvimento de novas técnicas para anotação automática (CIMIANO, 2006) (CIMIANO, HANDSCHUH, STAAB, 2004) ou semi-automática (KIRYAKOV et al., 2004) tem nos dado exemplos de que este cenário passa por significativos avanços. A geração automática ou semi-automática dos metadados, o aumento da qualidade dos metadados disponíveis, a definição critérios para a sua avaliação, o uso de metadados controlados e que seguem uma conceitualização

específica, o casamento de ontologias são tópicos de pesquisa ainda recente. Pode-se questionar se o uso de palavras-chave se deve mais às limitações das técnicas atuais de representação semântica. Neste sentido, o desenvolvimento de sistemas de recuperação que aproveitem as maiores possibilidades encontradas com o uso de metadados contribuirá para responder a esta questão.

3.13.2 Formulação da Consulta

Um dos grandes desafios para os sistemas de recuperação endereçados à Web Semântica é oferecer uma interface de consulta que combine facilidade e poder de expressão. Em um sistema de recuperação tradicional a formulação da consulta é feita através de palavras-chave, o que aproxima o usuário do texto livre que é sua forma natural de expressão. Entretanto, em um sistema apoiado por metadados, podemos fazer pesquisas diretas por seus elementos, por exemplo, por conceitos específicos da base de dados, ou endereçar consultas formais à base de conhecimento para recuperar instâncias de interesse. No caso em que conceitos são utilizados, temos a vantagem de que os conceitos são descritos através de ontologias e, portanto, estão dentro de um contexto, o que garante uma interpretação correta dos termos da consulta. No caso em que consultas formais são utilizadas, ainda temos o contexto disponível na base, sendo que, a consulta formal nos retorna resultados precisos sobre o tópico procurado. No entanto, tanto a busca por conceitos quanto a busca através de consultas formais são próprias do universo do sistema de recuperação e não do domínio do usuário. Conhecer a estrutura de representação do conhecimento usada pelo sistema e, assim, formular corretamente a consulta é um pré-requisito difícil de ser atendido para a maioria dos usuários. Os sistemas pesquisados procuram superar a limitação na expressão da consulta vista acima de diversos modos mostrados sinteticamente na Tabela 3.

Consultas formais, expressas em linguagens como o RDQL, são de difícil montagem para um usuário leigo. Para esconder tal complexidade, alguns sistemas, como SHOE (HEFLIN, HENDLER, 2000) e KIM (KIRYAKOV et al., 2004), disponibilizam formulários próprios em que expõem de modo gradual os elementos existentes na base de conhecimento. Uma desvantagem no uso dos formulários é a limitação do poder de expressão do usuário (LEI, UREN, MOTTA, 2006).

Tabela 3 Comparação do esquema de recuperação

Sistema	Consulta	Método	Inferência
SHOE	Formulário	RDQL	N
TAP	Texto	Navegação no Grafo	S
KIM	Formulário	Booleano	N
ROCHA	Texto	Navegação no Grafo	S
DOSE	Texto	Booleano	S
AQUALOG	Texto	Booleano	S
SEMSEARCH	Formato Google	RDQL	S
OWLIR	Texto ou Metadados	Booleano	S
BEAGLE++	Texto	Booleano	S
QUIZRDF	Texto ou Metadados	Booleano	S
VALLET	Formal	RDQL	S
SIM-DL	Metadados	Booleano	S

O uso de palavras-chave ainda é o modo de formulação da consulta mais comum. São empregadas as mesmas técnicas de análise morfológica dos sistemas tradicionais, porém os termos são mapeados para conceitos e instâncias da base de conhecimento em uma etapa extra do processo de recuperação. Nessa etapa, é possível o uso de inferência para aumentar o número de elementos a serem procurados na base de dados. Uma desvantagem neste mapeamento é a necessidade de retirada da ambigüidade no caso do mapeamento encontrar mais de um conceito para o mesmo termo informado (GUHA, MCCOOL, MILLER, 2003).

Propostas intermediárias entre as abordagens por formulário e por palavras-chaves são encontradas em AquaLog (LOPEZ, PASIN, MOTTA, 2005) e SemSearch (LEI, UREN, MOTTA, 2006). Esses sistemas definem arcabouços ou estruturas nas quais os conceitos encontrados são encaixados gerando, respectivamente, questões e consultas formais a serem submetidas aos seus sistemas de recuperação.

Em Vallet et al (VALLET, FERNÁNDEZ, CASTELLS, 2005) são utilizadas consultas formais mas a questão da montagem da consulta não é abordada pelo sistema em si. Essa abordagem é similar a dos modelos clássicos de recuperação que não tratam da extração das palavras-chave em si, mas apenas definem como estas são organizadas e utilizadas pelos demais componentes dos modelos.

3.13.3 Análise de Similaridade

Recuperar documentos, partes estruturadas dos documentos, atributos estruturados ou uma combinação destes elementos incorpora uma visão particular do que é um sistema de recuperação e determina a aplicabilidade de técnicas de ordenação ou não. A Tabela 4 apresenta o tipo de elemento retornado por cada um dos sistemas discutidos.

Tabela 4 Elementos recuperados pelos sistemas

Sistema	Recuperação
SHOE	Documentos
TAP	Dados
KIM	Documentos
ROCHA	Documentos
DOSE	Documentos
AQUALOG	Dados
SEMSEARCH	Documentos
OWLIR	Documentos
BEAGLE++	Documentos
QUIZRDF	Documentos
VALLET	Documentos
SIM-DL	Dados

A busca por documentos é a abordagem mais adotada pelos sistemas discutidos, sendo que, a recuperação dos documentos que referenciam as instâncias pesquisadas é o método mais utilizado em um processo similar ao apresentado nos modelos tradicionais de recuperação. Em poucos casos, a pesquisa em cima de dados da base de conhecimento gera dados complementares à pesquisa textual e não adotam critérios de relevância. Nesse último caso, a recuperação de informação na Web Semântica seria uma atividade complementar da recuperação de informação em si, não substituindo os modelos tradicionais de recuperação.

Dada a dificuldade do usuário em expressar corretamente a sua necessidade de informação, os resultados obtidos pelos sistemas de recuperação são freqüentemente imprecisos (BAEZA-YATES, RIBEIRO-NETO, 1998). A precisão do sistema de recuperação apurada a partir dos documentos recuperados depende da análise de similaridade. Esta imprecisão nem sempre é considerada nos sistemas de recuperação para a Web Semântica. O uso de uma estrutura formal de consulta gera um conjunto resposta não ordenado, excluindo-se, assim, a etapa da

análise de similaridade. A Tabela 5 apresenta um resumo das estratégias aplicadas na análise de similaridade dos sistemas mostrados neste capítulo.

Tabela 5 Comparação dos métodos da análise de similaridade

Sistema	$tf\ x\ idf$	Abordagem Combinada
SHOE	N	-
TAP	N	-
KIM	S	-
ROCHA	S	*Spread Activation*
DOSE	S	-
AQUALOG	N	-
SEMSEARCH	S	-
OWLIR	S	-
BEAGLE++	S	*Object Rank*
QUIZRDF	S	-
VALLET	S	-
SIM-DL	N	*Similaridade entre conceitos*

Nos sistemas tradicionais, o principal critério de relevância é a freqüência de ocorrência dos termos em cada documento, combinado com o inverso da freqüência de ocorrência dos termos na coleção, conhecido como $tf\ x\ idf$ (BAEZA-YATES, RIBEIRO-NETO, 1998). Ou seja, a fórmula é o produto entre uma medida que observa as características de interesse e uma medida que observa se o documento possui uma característica que o destaca dos demais itens da coleção. Tal critério é mantido em vários dos sistemas avaliados de duas formas: na primeira temos o uso da mesma fórmula estatística, porém em lugar de termos verifica-se a ocorrência de instâncias em cada documento. Na segunda forma, o critério estatístico baseado em texto é combinado com outras estratégias de ordenação.

Outra abordagem para a ordenação dos resultados em ambiente hipertexto é o *Page Rank* (BRIN, PAGE, 1998). Por ele, os resultados são ordenados segundo um critério de popularidade da página Web dentro do grafo criado a partir dos vínculos entre as páginas. Sistemas, como BEAGLE++ (CHIRITA et al., 2006) e Rocha et al (ROCHA, SCHWABE, ARAGAO, 2004), se inspiraram nesta idéia para criar uma forma de ordenação que aproveitasse as relações semânticas da base de conhecimento. Eles criam redes semânticas a partir da base de conhecimento, atribuem pesos para as relações e usam algoritmos de propagação de ativação para computar a importância de cada nodo para uma determinada consulta.

Vimos assim que muitos dos principais sistemas avaliados usam métodos de ordenação com abordagens estatísticas. Em apenas três dos sistemas pesquisados temos a distância semântica entre os conceitos sendo considerada como critério para determinar a relevância dentro da análise de similaridade.

3.14 CONSIDERAÇÕES FINAIS

Os sistemas apresentados neste capítulo empregam diversas técnicas que suportam plataformas híbridas, com texto e metadados, ou puramente baseadas em metadados. O uso de tais técnicas permite, de acordo com os resultados demonstrados nos textos pesquisados, a melhoria da efetividade desses sistemas em relação a sistemas meramente baseados em palavras-chave. O conhecimento contido tanto nos documentos, quanto na consulta, são explorados por processos de inferência para expansão dos elementos da representação. Isto faz com que os sistemas sejam capazes de responder com maior precisão a consultas em que algum termo procurado não aparece explicitamente.

O estudo comparativo mostra que o formato empregado na representação dos documentos é o principal meio de aprimorar o desempenho dos novos sistemas em relação aos métodos clássicos de recuperação e que os processos de recuperação e análise de similaridade contribuem de modo menos efetivo para uma melhoria do desempenho dos sistemas de recuperação. Em particular, o processo de recuperação de documentos ainda se baseia nos modelos clássicos de recuperação apoiado em dados estatísticos. A recuperação e a análise de similaridade nos sistemas avaliados, com algumas exceções, utilizam adaptações das técnicas usadas nos modelos tradicionais. Novas técnicas, como por exemplo, o uso da distância semântica entre os elementos da base de conhecimento, ainda que combinado com valores estatísticos, pode refletir relações entre os elementos que antes eram desconsideradas. De fato, outras técnicas podem ser aproveitadas para incrementar os resultados obtidos por esta etapa aproveitando as novas tecnologias baseadas no conhecimento.

Entretanto, a simples aplicação dessas técnicas não é suficiente para gerar sistemas de recuperação para uma plataforma semântica (SCHEIR, PAMMER, LINDSTAEDT, 2007). Os sistemas apresentados utilizam processos próprios e

totalmente definidos para realizar a tarefa de recuperação. Enquanto, alguns deles (KIRYAKOV et al., 2004) (LOPEZ, PASIN, MOTTA, 2005) colocam maior ênfase na instanciação dos conceitos e instâncias do que no próprio processo de recuperação. O próximo capítulo apresenta um modelo para a recuperação de informação que utiliza processos de inferência na criação da representação interna e define componentes em alto nível para realizar a análise de similaridade.

4. UM MODELO DE RECUPERAÇÃO DE INFORMAÇÃO PARA A WEB SEMÂNTICA

A Web Semântica disponibiliza um conjunto de técnicas para a manipulação da informação que podem beneficiar a área de recuperação da informação. Dentre as principais técnicas encontram-se o uso de metadados, linguagens de representação para a construção de ontologias, linguagens de consulta, e mecanismos de raciocínio, entre outras. Apesar das diversas tecnologias disponíveis, existe uma falta de abstrações apropriadas para representar produtos e atividades envolvidas no processo de recuperação de informação de modo a extrair maiores vantagens dessas tecnologias.

Este capítulo apresenta um modelo para a recuperação de informação baseado em conhecimento que explora o conteúdo semântico dos itens de informação. O conhecimento utilizado pelo modelo é descrito em termos de ontologias, que provêem um vocabulário único e uma estruturação para os conceitos e relacionamentos existentes em um domínio de aplicação. Além disso, a construção do modelo também é baseada em ontologias que descrevem os formatos de representação e a interação do modelo proposto com o domínio. Conseqüentemente, o mecanismo de busca e ordenação previsto no modelo deve ser capaz de realizar operações e cálculos em função do conteúdo semântico que compõe a representação dos itens de informação.

A forma de organização do domínio proposta, define como os elementos do modelo são processados na busca e no cálculo de similaridade. Contudo, a criação destes elementos e as técnicas para comparar os componentes de uma base de conhecimento são pontos previstos pelo modelo, mas não especificados e detalhados. Cada instanciação do modelo irá gerar um novo sistema a partir de diferentes implementações para estes pontos de variações disponíveis. Algumas das possíveis abordagens para a implementação destes pontos são citadas na seção sobre a análise de similaridade.

A primeira parte deste capítulo apresenta uma visão geral de um modelo de recuperação e define, formalmente, os componentes do modelo proposto. A segunda parte trata da organização do domínio em função de casos de interesse do usuário, chamados de “casos semânticos”. Os principais componentes do modelo

são também detalhados: a representação dos elementos suportada por ontologias, a recuperação dos itens de informação através de suas estruturas semânticas e a análise de similaridade. Na última parte do capítulo, o processo de recuperação de informação é descrito através de etapas apoiadas nos elementos do modelo aqui apresentado.

4.1 VISÃO GERAL DO MODELO PROPOSTO

Neste trabalho, é adotada a definição de um modelo de recuperação de informação dado por (BAEZA-YATES, RIBEIRO-NETO, 1998).

Definição 11: Um modelo de recuperação de informação genérico pode ser descrito pela quádrupla:

$$(D, Q, \mathcal{F}, R(d, q))$$

onde:

- D é o conjunto das representações internas dos documentos;
- Q é o conjunto das representações internas das consultas;
- $\mathcal{F}$ é um *framework* para modelar as representações dos documentos, consultas e seus relacionamentos;
- $R(d, q)$ é uma função para medir a similaridade entre os itens de informação e a consulta.

O modelo apresentado acima apresenta componentes genéricos para modelos de sistemas de recuperação. Diversos modelos criados, como o modelo do espaço vetorial e o modelo booleano (BAEZA-YATES, RIBEIRO-NETO, 1998), por exemplo, redefinem esses componentes para obtenção de um modelo que possa ser diretamente instanciado. Conforme mostrado na Definição 11, D e Q são conjuntos genéricos e terão a mesma função independentemente do modelo a ser representado. O *framework*, por sua vez, é um componente do modelo genérico de recuperação de informação que assume o papel central ao se criar um novo modelo. É através do detalhamento deste *framework* que novos modelos de recuperação de

informação são especificados. Ele é responsável por definir como serão representados os elementos que irão compor os conjuntos D e Q. A partir dessa definição são escolhidas as estruturas necessárias para guardar as representações internas das consultas e dos documentos, operadores e regras para gerar essas representações e a forma de relacioná-las entre si. Ao lado do *framework*, a função de similaridade $R(d,q)$ é o outro componente a ser especificado por novos modelos. Em geral, a função caracteriza o modelo criado (BAEZA-YATES, RIBEIRO-NETO, 1998), definindo como as representações são comparadas e a sua formulação é influenciada pelos demais componentes do *framework*.

O modelo proposto neste trabalho utiliza as abstrações definidas para os componentes D, Q e $R(d,q)$, conforme o modelo genérico (Definição 11), porém, especifica o framework a partir de componentes responsáveis por armazenar o conhecimento acerca do domínio e um conjunto de operações a serem utilizadas para criar as estruturas previstas pelo modelo.

Definição 12: Modelo de recuperação de informação proposto nesse trabalho.

$$(D, Q, SS, KB, OP, RF(d,q), R(d,q))$$

Onde:

- D e Q são os conjuntos com as representações internas dos documentos e das consultas, similar ao modelo genérico;
- SS é o conjunto de todos os casos semânticos identificados no domínio;
- KB é uma base de conhecimento utilizada para guardar as ontologias e as instâncias do domínio e da aplicação;
- OP é o conjunto de operações utilizadas na criação do modelo;
- $RF(d,q)$ é uma função para a recuperação dos documentos;
- $R(d,q)$ é uma função para medir a similaridade entre os itens de informação e a consulta.

A redefinição do *framework* inclui uma estratégia de representação baseada em casos semânticos, uma base de conhecimento e um conjunto de

operações utilizadas pelo processo de recuperação. A estratégia de representação baseada em casos semânticos define como os metadados serão organizados na representação interna dos documentos. A base de conhecimento, conforme a Definição 8, é dada como se segue:

$$KB \coloneqq (\mathcal{O}, \mathcal{I}, inst, inst^R)$$

possui uma ontologia $\mathcal{O}$, um conjunto de instâncias $\mathcal{I}$, e funções $inst$ e $inst^R$ que relacionam os elementos do domínio $\mathcal{I}$ ao conjunto-imagem $\mathcal{O}$. A ontologia foi especificada através da equação dada na Definição 7, conforme se segue:

$$\mathcal{O} \coloneqq (C, R, H^C, A^{\mathcal{O}})$$

sendo o conjunto de conceitos do domínio C, organizados em uma hierarquia H^C, com o conjunto de relações R entre os conceitos e um conjunto de axiomas $A^{\mathcal{O}}$. Essa base de conhecimento armazena os itens de informação e possibilita uma uniformidade de representação para os documentos e consultas, pois compartilham uma conceitualização comum oferecida pela ontologia. As operações incluídas no conjunto OP são aplicadas na criação dos conjuntos D e Q, na recuperação e ordenação dos itens.

O modelo aqui descrito é baseado em conhecimento, desta forma, consultas e documentos são representados através de conceitos e instâncias, sendo que esta representação pode também ser armazenada como instâncias de ontologias na base de conhecimento. As ontologias irão descrever o conhecimento do domínio da aplicação para que tanto os itens de informação, quanto as consultas possam ser corretamente interpretados. A primeira das operações a ser aplicada é uma estratégia baseada nos casos semânticos para estruturar as representações internas. Esta estrutura também está descrita em termos de uma ontologia. Os conceitos utilizados para caracterizar as instâncias encontradas nos documentos são pesquisados na base de conhecimento conforme indicado no modelo. Uma segunda operação é utilizada para montar a representação interna a partir das instâncias: é o conceito mais específico de uma instância (*most specific concept*) (BAADER et al., 2003) que é utilizado para inferir quais conceitos estão presentes em cada documento. As instâncias resultantes da aplicação da estratégia de representação baseada em casos semânticos são também armazenadas na base de

conhecimento. A terceira operação prevista pelo modelo é a especificação do conteúdo semântico de um conceito a fim de especificar um critério para a recuperação dos itens de informação. Por fim, uma medida de similaridade semântica entre conceitos completa o conjunto de operações previstas pelo modelo. Esta medida será utilizada pela função $R(d,q)$ para obter um valor numérico que determine uma ordenação para os documentos recuperados. Segundo o modelo apresentado, a função de ordenação também utiliza uma estratégia baseada em casos semânticos para gerar a ordenação dos resultados. Os componentes do modelo, a definição das operações e a aplicação das mesmas na criação dos componentes e no processo de recuperação são descritos nas seções seguintes.

4.2 CASOS SEMÂNTICOS

Um caso semântico representa uma característica de um item de informação, através da qual, os interesses de um usuário podem ser especificados (DRUMOND, GIRARDI, SILVA, 2008). Um primeiro exemplo de utilização de casos semânticos é a representação dos itens de informação criados como parte do modelo de recuperação do sistema ROSA (*Retrieval Of Software Artifacts*) (GIRARDI, 1995). O sistema recupera componentes de software utilizando técnicas de processamento de linguagem natural para analisar as suas descrições textuais. O objetivo da análise das descrições textuais é descobrir as características dos componentes de software relevantes para a recuperação dos mesmos. Tais características são representadas como conceitos do domínio tais como: ação, finalidade, plataforma e linguagem de implementação do software. Cada característica é identificada no texto em função dos elementos sintáticos e dos seus relacionamentos semânticos.

Cada item de informação pode apresentar um valor diferente para cada característica representada através de um caso semântico. Por exemplo, uma característica de um componente de software é a sua linguagem de programação, representada pelo caso semântico *linguagem*. Este caso semântico pode assumir diferentes valores em diferentes elementos de informação como, por exemplo, *Java*, *C++*, *Python*, *Prolog*, etc. Esses valores são chamados de termos do caso semântico. Os casos semânticos possíveis são descritos, no sistema ROSA, através

de frames, conforme mostrado na Figura 17. A análise de cada sentença da descrição do componente de software gera uma ou mais instâncias destes frames. As consultas, também em linguagem natural, são representadas pelo mesmo sistema de frames.

O mecanismo de análise de similaridade entre as sentenças calcula, para cada caso semântico, a distância conceitual entre os termos do caso semântico, presentes na representação da consulta, e os termos do caso semântico, presentes na representação da descrição do componente de software.

FRAME sentence IS_A_KIND_OF root_frame

CASES

Action	CATEGORY	verb
Agent	DOMAIN	component
Destination	DOMAIN	noun_phrase
Duration	DOMAIN	noun_phrase
Instrument	DOMAIN	noun_phrase
Language	DOMAIN	noun_phrase
Location	DOMAIN	noun_phrase
Manner	DOMAIN	sentence
Object	DOMAIN	noun_phrase
Platform	DOMAIN	noun_phrase
Provider	DOMAIN	noun_phrase
Purpose	DOMAIN	sentence
Source	DOMAIN	noun_phrase
Time	DOMAIN	noun_phrase.

Figura 17 Modelo de representação baseado em frames dos documentos do ROSA (GIRARDI, 1995)

Os casos semânticos representam a informação contida nos documentos em unidades mais precisas do que simples palavras-chave. As técnicas de representação de conhecimento da Web Semântica, herdadas da área de Inteligência Artificial, se prestam a um objetivo semelhante. As características representadas pelo esquema de classificação do modelo ROSA podem ser vistas como conceitos do domínio. Ou seja, os conceitos relacionados a um item de informação representam o conhecimento obtido a cerca de tal item. Adicionalmente, a definição dos conceitos apresenta um formalismo maior do que o apresentado na definição dos casos semânticos do modelo ROSA, possuindo regras para a sua formação, atributos, relações e sistemas de raciocínio embutidos. Este formalismo, proposto neste trabalho, é utilizado na definição de casos semânticos específicos para o contexto da Web Semântica.

No modelo proposto, os interesses do usuário são conceitos, sendo que, cada conceito descreve uma parte da ontologia do domínio. Assim, um caso

semântico é composto de um conceito mais geral, raiz da sua respectiva sub-hierarquia, e os seus subconceitos, uma vez que preservam as propriedades encontradas no conceito mais geral. Os conceitos gerais, também chamados de conceitos raízes dos casos semânticos, possuem um conjunto de atributos que representam diferentes e independentes aspectos do domínio. Tais atributos são preservados pelas especializações dos conceitos gerais. Esses conceitos formam os grupos de interesse do usuário. A seleção dos conceitos gerais pode ser feita com o auxílio de um especialista no domínio e deve ser mais abrangente quanto possível, de modo que todos os conceitos de interesse sejam especializações de algum dos conceitos selecionados.

Formalmente, um caso semântico é definido como um par:

Definição 13: Caso semântico

$$S = (C_r, T)$$

onde C_r é o conceito raiz do caso semântico e T é o conjunto de subconceitos de C_r, tal que:

$$T = \{C_i | C_i \sqsubseteq C_r\}$$

Um exemplo do que serão os conceitos raízes e como se dá a formação dos casos semânticos pode ser extraído da ontologia mostrada na Figura 18, na qual os termos $Person$, $Content$ e $Location$ são conceitos gerais do domínio.

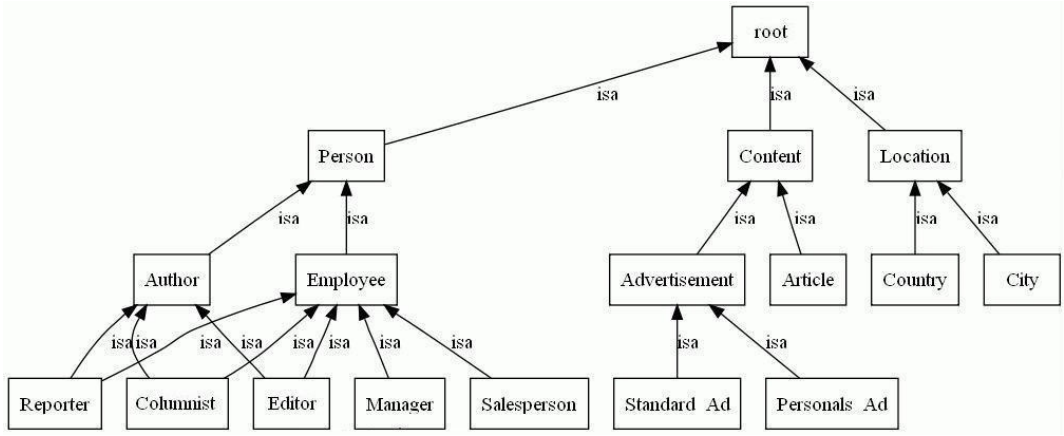

Figura 18 Trecho de uma ontologia com o domínio de um jornal

No modelo de recuperação proposto abaixo, um item de informação irá referenciar um caso semântico apenas se algum dos termos que compõem a sua representação interna for subconceito de algum conceito raiz de caso semântico. O caso semântico obtido a partir do conceito $Location$, mostrado na Tabela 6, é dado por $S_{Location= (Location,\{Location,City,Country\})}$.

Tabela 6 Casos semânticos no domínio de um jornal

Índice	Conceito Raiz	Termos
Caso 1	$Person$	$Person, Author, Employee, Reporter, Columnist, Editor, Manager, Salesperson, Director$
Caso 2	$Content$	$Content, Advertisement, Article, Standard\ Ad, Personals\ Ad$
Caso 3	$Location$	$Location, Country, City$

Diferentes documentos podem conter diferentes valores para cada um dos conceitos discriminados pela ontologia. Por exemplo, os valores *New York*, *Washington* e *Los Angeles* são valores possíveis para o conceito *City*. Em uma base de conhecimento, esses valores são armazenados como instâncias da ontologia.

4.3 REPRESENTAÇÃO INTERNA DOS ITENS DE INFORMAÇÃO E DA CONSULTA

Segundo o modelo apresentado, os seus componentes *D* e *Q* guardam representações internas dos itens de informação e da consulta. Esta seção descreve como esses elementos são representados a partir dos demais componentes do modelo baseado em conhecimento.

Conforme mostrado no capítulo sobre os modelos clássicos de recuperação, os documentos, nesses modelos, são descritos através de palavras-chave que retratam os elementos mais significativos do vocabulário extraído dos textos. Essas palavras-chave, chamadas de termos de indexação, possuem um valor relativo de importância em cada documento da coleção. Assim sendo, os documentos são representados em função do conjunto destes valores, também

chamados de pesos. A fórmula de cálculo para a obtenção destes valores irá variar de acordo com cada modelo específico.

Diferente dos modelos clássicos, a função de similaridade do modelo proposto não executará um cálculo sobre os valores que representam a importância de cada termo, mas sim, avaliará o conteúdo semântico de cada item de informação em relação ao conteúdo da consulta apresentada. Por isso, a representação desses documentos não é dada em função de pesos, mas sim, através da especificação dos conceitos e instâncias pertencentes à base de conhecimento e que também fazem parte de cada documento em particular.

O uso de conceitos e instâncias de uma base de conhecimento permite diminuir a ambigüidade na representação dos documentos. Cada elemento instanciado a partir de um documento possui uma interpretação comum e compartilhada, o que não ocorre com os termos simples. Por exemplo, em um documento textual representado através de um modelo clássico, a palavra-chave $molho$ pode se referir tanto ao ato de molhar alguém, ou seja, ao significado dado pelo verbo molhar, quanto a um item culinário, ou seja, o substantivo molho. Isto não ocorre ao trabalharmos com uma base de conhecimento, pois, temos o conceito referenciado associado à instância em questão.

A conceitualização de um domínio pode conter diferentes aspectos, com características distintas para descrevê-los. A identificação destes aspectos pode ser feita através de casos semânticos associados ao domínio conforme apresentado acima. A representação dos elementos de informação do modelo proposto é dada em função dos casos semânticos existentes no domínio que são referenciados por cada documento. É preciso então definir a forma de organização de cada documento e como alocar os conceitos e instâncias presentes em um documento criando assim a sua representação interna. Essa alocação é realizada de acordo com os casos semânticos e será detalhada através de algumas definições para maior clareza e compreensão.

Definição 14: Seja c um conceito pertencente ao conjunto de conceitos da ontologia da base de conhecimento, $c \in C$, e j o índice para um caso semântico do domínio. É dito que c está relacionado com o caso semântico S_j se o conceito c estiver relacionado ao conjunto de termos do caso semântico T_j, isto é, $c \in T_j$.

A definição dada expressa uma relação entre um conceito e o caso semântico. Por exemplo, o conceito $Country$ pertence ao caso semântico de índice 3 no domínio de um jornal, conforme descrito na Tabela 6, pois pertence ao conjunto dos termos desse caso semântico.

Um documento contém, na maioria dos casos, além de conceitos, instâncias da base de conhecimento. Uma instância representa um indivíduo e não pode ser mapeada diretamente para um caso semântico uma vez que a definição desse último é feita através de conceitos da ontologia do domínio. O mapeamento então é feito através do conceito mais próximo da instância e que contenha o maior número de propriedades que o descreva, ou seja, o seu conceito mais específico.

O conceito mais específico é uma definição extraída da teoria da lógica de descrição, cujo formalismo influenciou o desenvolvimento das principais linguagens para a Web Semântica (HORROCKS et al., 2007). As linguagens ontológicas comumente adotadas possuem uma representação que encontra equivalência nas representações definidas pela lógica de descrição. Uma base de conhecimento, conforme definida em capítulo anterior, é um componente de um sistema em lógica de descrição também composto por conceitos e indivíduos. O mapeamento entre as representações distintas para uma base de conhecimento é possível para versões mais simples de OWL, tais como OWL Lite e OWL DL. Uma ontologia é comumente vista como uma descrição da estrutura do domínio sendo assim equivalente ao conjunto de conceitos em lógica de descrição, chamado de *T-Box*. Situações particulares, que retratam indivíduos (ou instâncias conforme apresentado na definição da base de conhecimento), são mapeadas por outro conjunto chamado de *A-Box*. Na nomenclatura usada pela lógica de descrição, o conceito mais específico de uma instância é dado conforme definição seguinte (BAADER et al., 2003):

Definição 15: Um conceito *c* é o conceito mais específico de um indivíduo *a*, denotado por $msc(a)$, pertencente a *A-Box* se a for uma instância de c, isto é, $c(a)$, e para qualquer outro conceito d, $\forall d$, tal que a seja instância de d, $d(a)$, temos que $c \sqsubseteq d$.

Exemplificando a Definição 15, suponhamos que exista uma instância de um colunista de nome $Thomas\ Friedman$ que, no domínio de um jornal exemplificado na Figura 18, é uma instância dos conceitos $Author$, $Person$ e $Columnist$. O conceito mais específico dessa instância é $Columnist$, pois os demais conceitos $Author$ e $Person$ são seus super conceitos, isto é, $Columnist \sqsubseteq Author$ e $Columnist \sqsubseteq Person$.

Para relacionar uma instância a um caso semântico devemos relacionar a instância aos conceitos que fazem parte da hierarquia do caso semântico e que sejam capazes de descrever as propriedades da instância. Além disso, para reunir a maior quantidade de instâncias para caracterizar o documento, devemos evitar utilizar o uso de um conceito para descrever a instância que seja uma generalização do conceito geral do caso semântico. Ou seja, o conceito relacionado estaria fora da hierarquia do caso semântico definida a partir de um conceito geral. Para atender a este requisito e simplificar o critério de seleção do conceito utilizamos o conceito mais específico da instância msc como forma de relacionar a instância ao caso semântico.

Definição 16: Seja i uma instância pertencente ao conjunto de instâncias I da base de conhecimento, $i \in I$, e j o índice para um caso semântico do domínio. A instância i está relacionada ao caso semântico S_j se o seu conceito mais específico estiver também relacionado ao caso semântico S_j, $msc(i) \in T_j$.

Voltando ao exemplo anterior, com a instância chamada $Thomas\ Friedman$, temos que: $msc(Thomas\ Friedman) = Columnist$. Com isto, a instância $Thomas\ Friedman$ pertencerá ao caso semântico de índice 1 na Tabela 6, pois o seu conceito mais específico está presente no conjunto de termos do caso 1.

A partir dessa última definição, podemos também concluir que os conceitos que caracterizam o documento podem estar explicitamente citados no texto ou podem estar descrevendo alguma instância. O conjunto dos conceitos representativos de um documento é definido conforme a seguir.

Definição 17: Seja C_d o conjunto de conceitos, n o número total de instâncias e I_d o conjunto de instâncias tal que $I_d = \{i_1, \dots, i_n\}$, todos presentes em

um documento d. O conjunto de conceitos que descrevem as instâncias do documento é dado por $C_i = \{msc(i_1), \dots, msc(i_n)\}$. O conjunto CR_d dos conceitos representativos do documento será então a união entre os conceitos presentes no documento e os conceitos descritivos das instâncias, $CR_d = C_d \cup C_i$.

Partindo-se também do exemplo dado na Figura 18, vamos supor que um documento apresente os conceitos $Employee$ e $Article$ em seu conteúdo, além da instância $Thomas\ Friedman$. Unindo-se o conjunto $C_d = \{Employee, Article\}$ ao conjunto C_i formado pelo $msc(Thomas\ Friedman)$, tal que $C_i = \{Columnist\}$, teremos que $CR_d = \{Employee, Article, Columnist\}$.

A representação interna proposta irá agrupar os conceitos e instâncias do documento de acordo com os casos semânticos identificados no domínio. Para isso, organiza tais elementos em pares de conjuntos de acordo com cada caso semântico. Cada par contém:

- Um conjunto com os conceitos representativos presentes no documento, relativos a um determinado caso semântico;
- Um conjunto com as instâncias presentes no documento, relativas a um determinado caso semântico.

Definição 18: Seja m o número de casos semânticos do domínio, e SS o conjunto de todos os casos semânticos obtidos tal que $SS = \{S_1, \dots, S_m\}$. O documento d_k é representado por um conjunto de pares P, sendo cada par associado a um caso semântico, $d_k = \{P_{k1}, \dots, P_{km}\}$. Sendo j o índice para um dos casos semânticos, o par P_{kj} contém: os conceitos representativos encontrados em d_k e as instâncias encontradas em d_k pertencentes ao caso semântico S_j, tal que $P_{kj} = (CR_{kj}, I_{kj})$.

Expandindo-se a Definição 18, a representação de um documento d_k, segundo o modelo proposto, é dada em função da seguinte equação cujos componentes foram definidos acima:

$$d_k = \{(CR_{k1}, \mathrm{I}_{k1}), \dots, (CR_{km}, I_{km})\}$$

Para exemplificar o formato de representação definida acima, é utilizada uma sentença que possa ser instanciada a partir dos conceitos presentes na ontologia que descreve o domínio de um jornal mostrado na Figura 18. Dada a sentença *"Thomas Friedman wrote an article about New York..."* são identificados os elementos que farão parte da representação interna da sentença conforme o modelo. A Tabela 7 nos mostra os componentes e valores obtidos a partir da sentença

Tabela 7 Valores instanciados a partir de uma sentença exemplo

Componentes	**Valores Instanciados**
Instâncias	$Thomas\ Friedman, New\ York$
Conceitos	$Article$
Conceitos Adicionais	$msc(Thomas\ Friedman) = Columnist,$ $msc(New\ York) = City$
Conceitos Representativos	$Article, Columnist, City$

Uma vez que os componentes e os valores estão instanciados, o modelo os agrupa em função dos casos semânticos do domínio. A Tabela 6 apresenta os três casos semânticos escolhidos para o domínio da ontologia de um jornal. Portanto, a sentença será representada com um total de três pares contendo os valores instanciados para a mesma. A Tabela 8 mostra um exemplo de como seria a representação interna final da sentença em questão:

Tabela 8 Exemplo da representação interna do modelo proposto

SENTENÇA: "*Thomas Friedman wrote an article about New York...*"			
Caso Semântico	1	2	3
Conceito Raiz	$Person$	$Content$	$Location$
REPRESENTAÇÃO INTERNA			
Conceitos	$Columnist$	$Article$	$City$
Instâncias	$Thomas\ Friedman$	-	$New\ York$

A Tabela 8 enfatiza a relação entre os conceitos e instâncias identificados no documento e os casos semânticos do domínio colocados numa mesma coluna. Considerando o formato especificado na Definição 18, a representação é vista como um conjunto de pares:

$$d = \{(\{Columnist\}, \{Thomas\ Friedman\}), (\{Article\}, \{\}), (\{City\}, \{New\ York\})\}$$

A definição da representação interna da consulta é similar à apresentada e exemplificada para os itens de informação.

Definição 19: Seja m o número de casos semânticos do domínio, e *SS* o conjunto de todos os casos semânticos obtidos tal que $SS = \{S_1, \dots, S_m\}$. A consulta q_k é representada por um conjunto de pares P, sendo cada par associado a um caso semântico, $q_k = \{P_{k1}, \dots, P_{km}\}$. Sendo j o índice para um dos casos semânticos, o par P_{kj} contém: os conceitos representativos encontrados em q_k e as instâncias encontradas em q_k pertencentes ao caso semântico S_j, tal que $P_{kj} = (CR_{kj}, I_{kj})$.

A definição da representação interna dos itens de informação e da consulta serve como ponto de partida para a definição dos demais componentes do modelo de recuperação. Uma das principais vantagens dos modelos de recuperação clássicos é a simplicidade na representação dos documentos com o surgimento de diversos modelos a partir do uso de operações advindas de áreas como a álgebra e a estatística. O formato de representação do modelo proposto é dado em função de elementos da base de conhecimento, sendo possível tanto a manipulação dos seus termos individualmente quanto armazená-los na própria base de conhecimento e submetê-los a operações de inferência. Ou seja, o formato é baseado no conhecimento, o que possibilita a sua manipulação através dos recursos desenvolvidos para esse tipo de representação.

4.4 RECUPERAÇÃO DOS ITENS DE INFORMAÇÃO

A recuperação dos itens de informação corresponde à definição das funções $RF(d, q)$ e $R(d, q)$ do modelo proposto neste trabalho. Essas funções são responsáveis por, nessa ordem, recuperar e ordenar os itens de informação relevantes à consulta. Para definir essas funções é necessário estabelecer critérios que possam atender à necessidade de informação do usuário ao utilizar o sistema. Tal necessidade difere da consulta formulada que são os termos que o usuário informa ao sistema tentando expressar a sua necessidade de informação. Um item

de informação é relevante apenas se o usuário observar nesse item elementos que correspondam à sua necessidade de informação (MANNING, RAGHAVAN, SCHÜTZE, 2008).

A efetividade de um sistema de recuperação depende, em parte, da capacidade do usuário em expressar corretamente a sua necessidade de informação. Se a consulta submetida for mais genérica do que o necessário, haverá um alto índice de recuperação (*recall*) e uma baixa precisão (*precision*) nos resultados. Sendo a consulta mais específica, observamos a situação inversa, com maior precisão e um número menor de documentos recuperados (GIRARDI, 1995). No caso de uma abordagem baseada no conhecimento, apesar do uso de um conjunto de conceitos, expressos através de uma semântica precisa, diminuir o número de documentos não relevantes retornado, a consulta do usuário pode ainda ser mais ou menos precisa do que a real necessidade de informação do usuário. No contexto de uma ontologia, documentos relevantes podem se referir tanto a super quanto a subconceitos do conceito pesquisado, uma vez que os conceitos são organizados hierarquicamente em uma ontologia na qual um conceito é uma especialização de um outro e herda as suas propriedades e características.

O modelo proposto define os componentes para lidar com essa característica identificada no processo de busca dos usuários. O primeiro componente é a função que identifica os itens de informação considerados relevantes para a consulta. Essa função define como será feito o casamento (*matching*) entre os termos da consulta e dos itens de informação. O segundo componente é a função que ordena os itens de informação segundo a proximidade semântica entre a representação interna da consulta e dos itens de informação.

4.4.1 Casamento (*Matching*)

Muitos modelos de recuperação, como o do espaço vetorial (BAEZA-YATES, RIBEIRO-NETO, 1998), por exemplo, definem apenas uma função de similaridade que ordena toda a coleção em termos de uma função de relevância, por exemplo, ainda no modelo vetorial, a distância entre os termos no espaço vetorial. Em um modelo baseado em conhecimento, determinar a relevância de todos os documentos pode ser um processo custoso. Por isso, o modelo proposto define uma

função para a recuperação dos itens de informação baseada apenas na ocorrência de conceitos tidos como relevantes para, em seguida, determinar o grau de relevância de cada item recuperado.

O modelo booleano (BAEZA-YATES, RIBEIRO-NETO, 1998) apresenta a fórmula mais simples de recuperação para os modelos tradicionais, pois recupera os documentos segundo a presença dos termos da consulta nos itens de informação. A função de recuperação aqui descrita utiliza uma idéia análoga ao recuperar os itens de informação segundo a presença dos conceitos extraídos da consulta. A diferença entre as abordagens é que a consulta aqui é expandida de modo a diminuir a distância cognitiva entre o conceito pesquisado pelo usuário e o conceito presente no item de informação. Conforme dito antes, o conceito pesquisado pode ser mais, ou menos genérico que os conceitos encontrados na busca.

Para obter uma definição formal do que caracteriza a expansão da consulta em termos de conceitos mais, ou menos genéricos em relação a outros, temos a definição do conjunto de todos os super e subconceitos de um determinado conceito C_i, chamado de "*Semantic Cotopy*" (SC) de C_i (MAEDCHE, STAAB, 2002), descrito como:

Definição 20: "*Semantic Cotopy*" (SC) de C_i.

$$SC(c_i, H^C) = \{c_j \in C | H^C(c_j, c_i) \vee H^C(c_i, c_j)\}$$

onde c_i e c_j são conceitos da hierarquia H^C, $H^C(c_j, c_i)$ o conjunto de subconceitos de c_i e $H^C(c_i, c_j)$ o conjunto de superconceitos de c_i.

Tomando como exemplo o conceito $Author$ que está presente na ontologia vista na Figura 18, que define uma hierarquia H^C, podemos exemplificar a formação da "*Semantic Cotopy*" de um conceito, como apresentado na Tabela 9:

Tabela 9 Exemplo da "*Semantic Cotopy*" de um conceito

Conceito: $Author$	
Superconceitos	$Root; Person; Author$
Subconceitos	$Author; Reporter; Editor; Columnist$
$SC(Author, H^C) = \{Root, Person, Author, Reporter, Editor, Columnist\}$	

A "*Semantic Cotopy*", o conjunto formado pelos super e subconceitos de C_i, representa o conteúdo semântico do conceito C_i. Portanto, a função de recuperação selecionará conceitos presentes na representação do item de informação com conteúdo semântico similar ao dos conceitos presentes na consulta. Se existir um conceito na representação do item de informação que esteja presente na "*Semantic Cotopy*" de algum dos conceitos da consulta então o item de informação é considerado relevante e recuperado.

A função de recuperação utiliza os conceitos dos conjuntos de conceitos representativos da representação interna do documento, conforme a Definição 18, para fazer o casamento entre itens de informação e consulta conforme segue:

Definição 21: Fórmula para recuperação dos itens de informação do modelo proposto.

$$RF(d,q) := \begin{cases} 1, se\ \exists_{c_i,c_j} | \left(c_i \in CR_q\right) \wedge \left(c_j \in CR_d\right) \wedge \left(c_j \in SC(c_i, H^C)\right) \\ 0, caso\ contrário \end{cases}$$

onde CR_q e CR_d são conjuntos com os conceitos representativos, respectivamente, encontradas na consulta e no item de informação.

Considere como exemplo a seguinte consulta *"q: New York authors*", cuja representação interna é:

$$q = \{(\{Author\}, \{\}), (\{\}, \{\}), (\{City\}, \{New\ York\})\}$$

Ao executar a função de recuperação para a consulta q e a sentença vista na Tabela 8, chamada d, a busca será feita por conceitos de d (i.e., $Columnist$, $City$ e $Article$) que estejam presentes na "*semantic cotopy*" dos conceitos de q (i.e. $Author$ e $City$). A "*semantic cotopy"* de cada conceito em q são dadas a seguir:

- $SC(Author, H^C) = \{Root, Person, Author, Reporter, Editor, Columnist\}$;
- $SC(City, H^C) = \{Root, Location, City\}$.

Os conceitos $Columnist$ e $City$ aparecem tanto na "*semantic cotopy*" dos conceitos de q quanto na representação interna de d, portanto, o resultado da execução da função $RF(q,d)$ será 1.

A recuperação dos itens de informação descrita realiza uma tarefa similar à recuperação de dados. Um modelo para a recuperação de informação deve conter uma função para a ordenação dos resultados a ser definida no tópico seguinte.

4.4.2 Análise de Similaridade

Em um modelo de recuperação de informação, a análise de similaridade mede a relevância do documento recuperado em relação à consulta. O processo de recuperação guiado por conceitos já preserva o contexto de busca e assegura que os itens recuperados terão sempre alguma relevância. Assim sendo, o modelo de similaridade descrito abaixo tem por objetivo definir uma ordenação para os documentos recuperados. Esta ordenação é representada pela função $R(d,q)$ conforme definido na descrição do modelo genérico de recuperação.

No modelo aqui proposto, a representação interna de um item de informação d é um conjunto de pares de acordo com a Definição 18. Para determinar a relevância de um item de informação em relação a uma consulta, representada de modo similar ao item de informação de acordo com a Definição 19, um critério de comparação entre os pares de ambas as representações se faz necessário. A análise de similaridade descrita abaixo usa uma abordagem similar à apresentada em (DRUMOND, GIRARDI, SILVA, 2008) (GIRARDI, 1995). Nessas abordagens, possíveis grupos de interesse do usuário em um domínio são representados como casos semânticos. A função para o cálculo de similaridade realiza o casamento entre os pares do item de informação com aqueles existentes na consulta que referenciem o mesmo caso semântico, ou seja, que referenciem o mesmo índice i em suas representações. Os elementos dos conjuntos CR_i e I_i dos pares casados pela função são comparados através de uma medida de similaridade semântica. O valor obtido para a relevância de um documento é a soma, normalizada, da similaridade obtida entre todos os pares casados. Adicionalmente, é possível também associar um peso ω para determinar a importância relativa de cada caso semântico pesquisado. Esse peso, previsto na função definida no modelo,

pode ser definido também por um especialista do domínio que identifica os casos semânticos de maior importância na ontologia. Cada instância do modelo pode também implementar outra fórmula para a definição destes pesos de acordo com outro critério. A fórmula de similaridade é dada por:

Definição 22: Fórmula para a análise de similaridade entre uma consulta e um item de informação do modelo proposto.

$$R(d,q) = \frac{1}{\sum_{i=1}^{m} \omega_i} * \sum_{i=1}^{m} \omega_i * \frac{Sim\left(CR_{i_d}, CR_{i_q}\right) + Sim\left(I_{i_d}, I_{i_q}\right)}{N_{CR_{i_q}} + N_{I_{i_q}}}$$

onde:

- i é o índice para um dos casos semânticos do domínio;
- m é o total de casos semânticos do domínio;
- CR_{i_d} e I_{i_d} são conjuntos de conceitos e instâncias da representação interna do item de informação em relação ao caso semântico i;
- CR_{i_q} e I_{i_q} são conjuntos de conceitos e instâncias da representação interna da consulta em relação ao caso semântico i;
- $N_{CR_{i_q}}$ e $N_{I_{i_q}}$ são, respectivamente, o número de conceitos e instâncias extraídos da consulta.

Na função definida acima, caso ambos os conjuntos do par da consulta relativo ao caso semântico de índice i sejam vazios, a iteração terá o valor zero.

A similaridade entre os conjuntos de conceitos previstos na equação da Definição 22 é dada por:

Definição 23: Medida de similaridade entre conceitos.

$$Sim\left(CR_d, CR_q\right) = \sum_{j \in CR_q} \max Sim_c\left(c_{q_j}, c_{d_k}\right)$$

onde:

- c_{q_j} é o elemento j do conjunto CR_q;
- c_{d_k} é o elemento k do conjunto CR_d;
- Sim_c é a medida de similaridade entre os dois conceitos;
- max obtém o valor da similaridade máxima entre o elemento de índice j, comparado com todos os elementos de CR_d.

.

Uma equação similar à apresentada na Definição 23 é utilizada para calcular a similaridade entre os conjuntos de instâncias previstas na Definição 22:

Definição 24: Medida de similaridade entre instâncias.

$$Sim(I_d, I_q) = \sum_{j \in I_q} \max Sim_i \left(i_{q_j}, i_{d_k}\right)$$

onde:

- i_{q_j} é o elemento j do conjunto I_q;
- i_{d_k} é o elemento k do conjunto I_d;
- Sim_i é a medida de similaridade entre as duas instâncias;
- max obtém o valor da similaridade máxima entre o elemento de índice *j*, comparado com todos os elementos de I_d.

A aplicação da Definição 22 será exemplificada com o uso da sentença e da consulta de exemplo mostradas anteriormente. A função de similaridade estabelece a comparação entre os pares relativos ao mesmo caso semântico no item de informação e na consulta. Essa comparação é feita distintamente para cada conjunto do par. A Tabela 10 mostra como os pares de mesmo índice estão alinhados nesse exemplo.

Tabela 10 Exemplo do alinhamento de pares entre um item de informação e uma consulta

Caso Semântico	**1**	**2**	**3**
Consulta	$(\{Author\}, \{\})$	$(\{\}, \{\})$	$(\{City\}, \{New\ York\})$
Item de Informação	$(\{Columnist\}, \{Thomas\ Friedman\})$	$(\{Article\}, \{\})$	$(\{City\}, \{New\ York\})$

Aplicando a equação da Definição 22, temos que similaridade entre a consulta q e o item de informação d no exemplo é dada por:

$$
\begin{aligned}
&Sim(d,q) \\
&= \left(\frac{1}{\omega_1 + \omega_2 + \omega_3}\right) \\
&* \left(\begin{array}{c} \omega_1 * \left(\dfrac{Sim(\{Columnist\},\{Author\}) + Sim(\{\quad\},\{Thomas\ Friedman\})}{1+0}\right) + \\ \omega_3 * \left(\dfrac{Sim(\{City\},\{City\}) + Sim(\{New\ York\},\{New\ York\})}{1+1}\right) \end{array} \right)
\end{aligned}
$$

O modelo de similaridade para os documentos prevê uma medida de similaridade entre conceitos/instâncias presentes na representação interna da consulta e do documento. Essa medida de similaridade entre conceitos não é descrita ou sugerida no modelo. Cada instância do modelo será um novo sistema que é responsável por selecionar e implementar tal fórmula. Várias medidas para este propósito já foram sugeridas na literatura (D'AMATO, FANIZZI, 2005) (DRUMOND, GIRARDI, SILVA, 2008) (JANOWICZ, 2006) (LEACOCK, CHODOROW, 1998) (LIN, 1998) (RESNICK, 1999) (RODRIGUEZ, EGENHOFER, 2004) (SCHADBOLT, HALL, BERNERS-LEE, 2006) e podem ser adaptadas por uma instância do modelo de recuperação para realizar a comparação de conceitos. Por esse motivo, o valor final para o cálculo da similaridade entre os documentos mostrados no exemplo acima, que seria obtido segundo o modelo, não é apontado.

4.5 PROCESSO DE RECUPERAÇÃO

Os componentes definidos pelo modelo são necessários para compor todas as etapas de um processo de recuperação de informação (Figura 19). Um processo de recuperação de informação possui um conjunto de etapas que, embora não sendo padronizadas, possui características comuns na maioria dos sistemas. Entre elas, podemos destacar a criação de uma coleção de itens de informação, indexação, formulação da consulta, operação textuais, a criação da representação interna, o casamento e a análise de similaridade entre a consulta e os itens de informação.

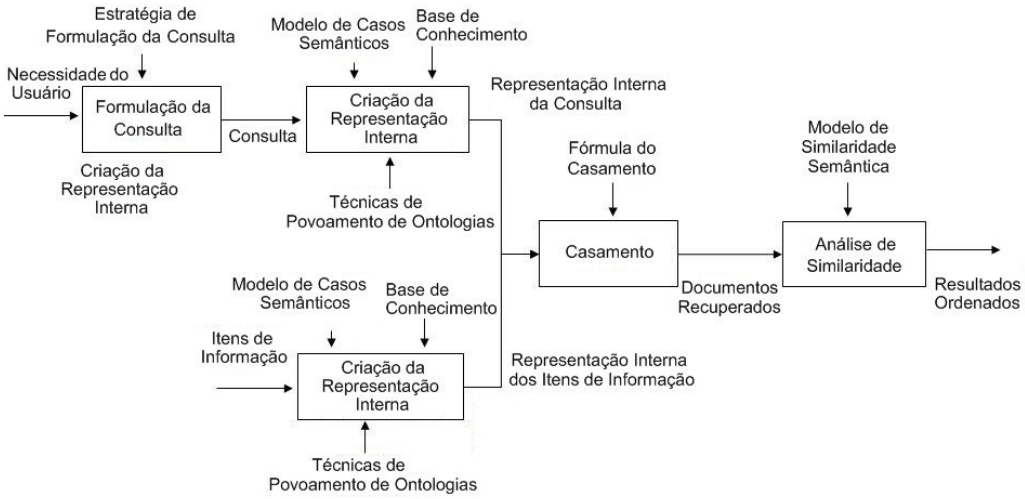

Figura 19 Visão geral do processo de recuperação de informação proposto

O processo de recuperação começa com a estratégia para a formulação da consulta, que é uma etapa dependente do usuário, que decide como expressar a sua necessidade de informação. A etapa seguinte é a formulação da consulta pelo usuário que pode ser especificada de diversas formas como: texto livre, de modo similar aos modelos de recuperação de informação tradicionais, através de uma interface personalizada para o domínio ou mesmo por alguma linguagem de consulta. Cada formato deve ser capaz de produzir um conjunto de termos a serem representados de acordo com a estratégia de casos semânticos. A criação da representação interna da consulta origina o conjunto Q previsto no modelo.

A inclusão de novos itens de informação na base de conhecimento é feita através da instanciação desses itens na etapa de criação da representação interna, de modo similar à criação da representação interna da consulta. O conjunto D previsto no modelo é formado nessa etapa que é responsável pela identificação de conceitos e instâncias da base de conhecimento que aparecem nos itens de informação.

O processo de instanciação dos documentos, tanto a consulta quanto os itens de informação, conta com o suporte de técnicas para o povoamento de ontologias. O povoamento de ontologias é uma atividade de aquisição de conhecimento que tem por objetivo instanciar conceitos de uma base de conhecimento a partir de fontes de dados desestruturadas ou semi-estruturadas (CIMIANO, HANDSCHUH, STAAB, 2004).

Outras operações OP do modelo são também utilizadas para dar suporte ao processo. Os conceitos são extraídos diretamente do texto ou obtidos através do cálculo do conceito mais específico (*most specific concept*) das instâncias extraídas. O esquema de representação deve estar especificado através de uma ontologia armazenada na base de conhecimento e ajustada ao domínio. Os conceitos e instâncias obtidos nesta etapa são separados de acordo com uma estratégia baseada no modelo de casos semânticos. Essa estratégia utiliza a definição do conjunto SS dos casos semânticos do domínio prevista no modelo e as operações definidas para separar os elementos em pares da representação interna de acordo com cada caso semântico. As representações internas dos itens de informação são armazenadas na base de conhecimento do domínio. O resultado desta etapa, portanto, são instâncias dos documentos, criadas segundo o modelo de representação, para posterior processamento na fase de recuperação dos itens de informação que irão satisfazer uma dada consulta.

Em seguida, a fase de recuperação dos itens seleciona os itens de informação de mesmo conteúdo semântico de uma dada consulta. Nessa fase, o componente do modelo $RF(d,q)$ é responsável por selecionar os itens de informação relevantes, utilizando a base de conhecimento como referência e o critério de seleção definido pela função, chamada na figura de "*Fórmula do Casamento*".

Finalmente, os itens recuperados são então ordenados, usando a função $R(d,q)$ do modelo, de acordo com a análise de similaridade semântica entre as representações para que seja apresentado o resultado final. O modelo de similaridade semântica a ser aplicado, conforme citado no tópico sobre a análise de similaridade, irá determinar a similaridade entre os termos das representações. Tal modelo pode ser uma instância de alguma medida já descrita na literatura.

Uma desvantagem comum nas abordagens de representação baseada em metadados é a pouca abrangência da base de conhecimento (VALLET, FERNÁNDEZ, CASTELLS, 2005). Novos documentos adicionados à máquina de busca podem conter conceitos que não estão presentes na base de conhecimento. Portanto, existe a necessidade de um processo de manutenção da base de conhecimento. Este processo pode ser manual, executado por um especialista, ou automático, suportado por técnicas de povoamento de ontologias (*ontology*

population) (CIMIANO, HANDSCHUH, STAAB, 2004). Apesar da importância deste processo de manutenção da base de conhecimento para a recuperação de informação baseada em ontologias, a discussão das técnicas de aprendizado de ontologias está fora do escopo deste trabalho.

4.6 CONSIDERAÇÕES FINAIS

Neste capítulo foi descrito um modelo de recuperação de informação para a Web Semântica, construído a partir de componentes semânticos e serviços como ontologias e regras de inferência. Para evitar resultados ambíguos e ruidosos, o modelo utiliza representação semântica, ou seja, conceitos e instâncias, em lugar de palavras-chave para representar os itens de informação. Uma base de conhecimento armazena os conceitos e instâncias do domínio, bem como os itens de informação coletados.

O modelo usa uma estratégia baseada em casos semânticos para organizar conceitos e instâncias na representação interna dos itens de informação. Os casos semânticos representam grupos de interesse do usuário e criam diferentes contextos dentro do domínio que são utilizados pelos processos de recuperação e análise de similaridade para encontrar um valor de relevância para os documentos segundo interesses específicos do usuário.

No capítulo a seguir, será feito um estudo de caso para validar a instanciação do modelo e ter uma avaliação de um sistema construído a partir das idéias apresentadas.

5. ESTUDO DE CASO

Neste capítulo é apresentado um estudo de caso para a avaliação do modelo de recuperação de informação proposto, através da instanciação do modelo, usando a medida de similaridade proposta em (DRUMOND, GIRARDI, SILVA, 2008). O sistema resultante da instanciação do modelo fará a recuperação de itens de informação para um conjunto de consultas da área do Direito Tributário brasileiro.

Para realizar o estudo de caso foi necessária a criação de uma base de conhecimento, prevista no modelo, para a instanciação dos documentos. Esses documentos são leis e decretos da área jurídica que foram modelados na ontologia ONTOJURIS (ARAÚJO et al., 2008) que descreve o conhecimento a cerca do Direito Brasileiro. Mais especificamente, foi utilizada a ONTOTRIB que é uma ontologia que estende a ONTOJURIS no ramo do Direito Tributário (ARAÚJO et al., 2008).

A instanciação do modelo foi realizada utilizando-se a ferramenta Protégé (GENNARI, MUSEN, FERGERSON, 2002) que permite a edição de ontologias para a criação de bases de conhecimento e uma API para a construção de aplicações baseadas no conhecimento. A medida de similaridade utilizou a API JENA[1] que permite a manipulação de ontologias. O sistema foi construído na linguagem JAVA[2], com a implementação dos componentes do modelo necessários e do processo de recuperação.

Para efeito comparativo, foi implementado um segundo sistema a partir da API disponibilizada pela ferramenta LUCENE (HATCHER, GOSPODNETIC, 2005) que possibilita a criação de sistemas de recuperação baseados no modelo vetorial. Tanto os itens de informação, quanto as consultas foram indexados a partir dos documentos textuais, e a ferramenta LUKE[3] serviu de interface para inserção das consultas e apresentação dos itens de informação relevantes.

Os resultados obtidos por ambos os sistemas foram avaliados em função das conhecidas taxas de revocação e precisão. Essas medidas são as mais comumente utilizadas para determinar a efetividade de um sistema de recuperação de informação.

---

[1] JENA - Disponível em http://jena.sourceforge.net/ontology/
[2] JAVA - Disponível em http://www.sun.com/java/
[3] LUKE - http://www.getopt.org/luke/

As próximas partes deste capítulo apresentam a ontologia ONTOTRIB, os casos semânticos do contexto da ONTOTRIB, o sistema implementado com modelo do espaço vetorial (BAEZA-YATES, RIBEIRO-NETO, 1998), a instanciação do modelo, os testes realizados a partir da instanciação dos itens de informação e das consultas, os resultados obtidos e, por fim, as conclusões a respeito do estudo de caso.

5.1 ONTOJURIS E ONTOTRIB

A ONTOJURIS é uma ontologia criada para descrever o conhecimento jurídico brasileiro. As classes da ONTOJURIS (Figura 20), modelam os instrumentos jurídicos normativos, ramos do direito e seus relacionamentos.

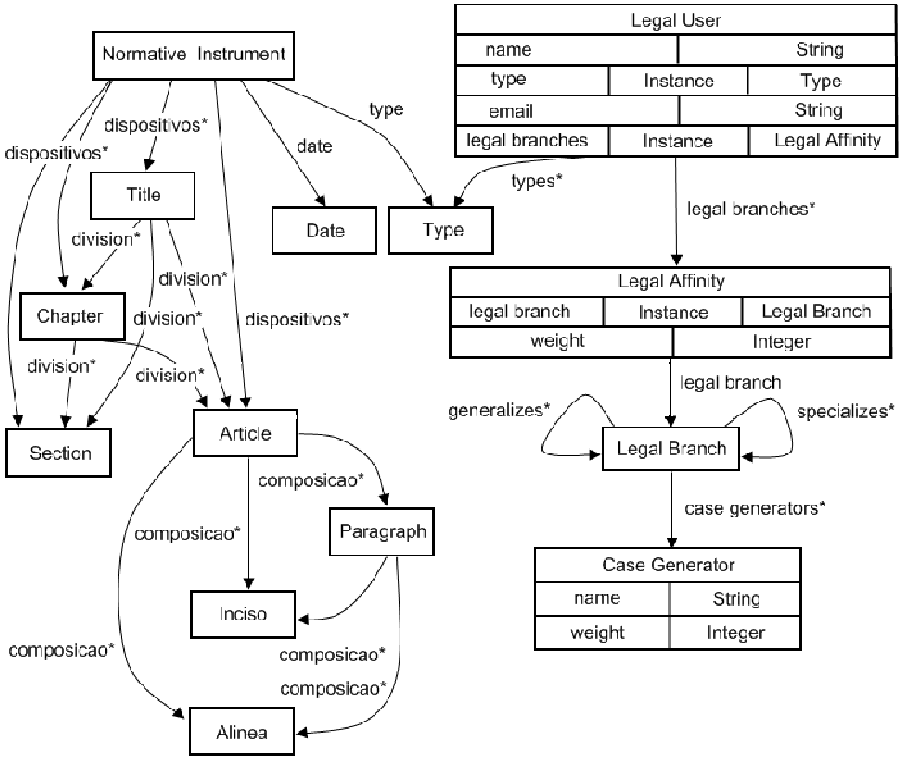

Figura 20 Trecho da ontologia ONTOJURIS

A classe *Normative_Instrument* possui atributos que representam a estrutura de um instrumento normativo enquanto os ramos do direito, representada pela classe *legal_branches*, subdividem o direito em áreas mais específicas, tais como, o Direito Civil, o Direito Tributário, o Direito Trabalhista, entre outros. Os

ramos do direito possuem como atributo os geradores de caso, cuja presença em um instrumento normativo indica a afinidade do instrumento normativo com o ramo em questão (ARAÚJO et al., 2008).

Dentre os conceitos descritos pela ONTOTRIB está o conceito de tributo, mostrado na Figura 21, que segundo a Teoria Pentapartida dos Tributos, do STF, seria gênero, cujas espécies seriam: impostos, taxas, contribuições de melhoria, contribuições especiais e empréstimo compulsório (CAPEZ, MALTINTI, 2006).

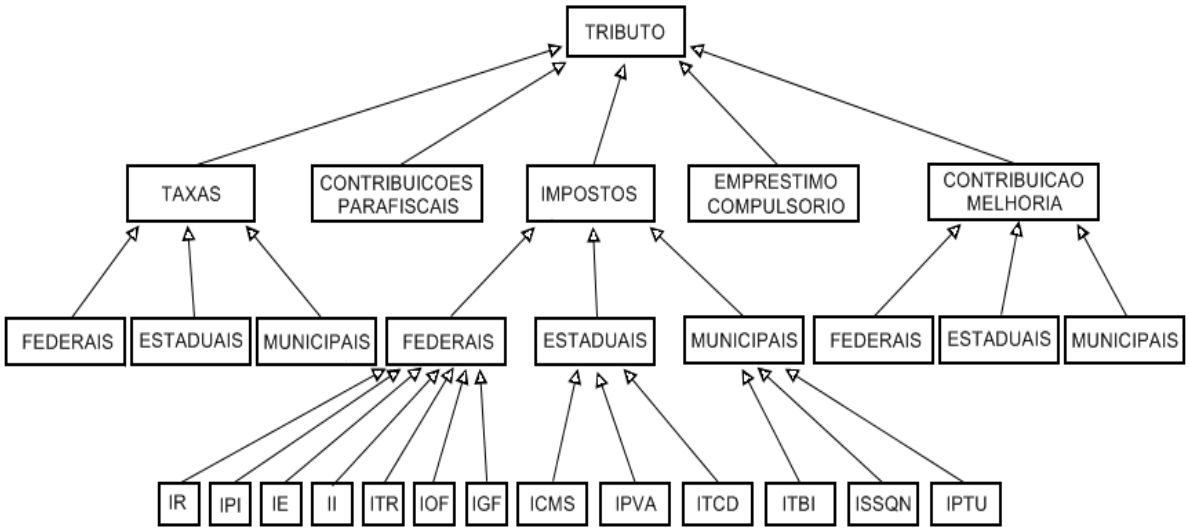

Figura 21 Hierarquia de tributos na ONTOTRIB

Além da classe que descreve um tributo, a ONTOTRIB especializa o instrumento normativo da ONTOJURIS em um instrumento normativo tributário através da adição de novos atributos: aplicação tributária, conceitos tributários e tributos (Tabela 11).

Tabela 11 Atributos exclusivos da subclasse "Instrumento Normativo Tributário"

Atributos	Descrição
Aplicação Judicial	Escopo de aplicação da norma tributária (Federal – Brasil, Estado ou Município específico)
Tributo	Espécie tributária a que se relaciona.
Conceitos Tributários	Sobre qual(is) conceito(s) a norma trata.

A aplicação tributária delimita o escopo de incidência de uma norma tributária que pode ser válida para toda a União ou para um Estado, um Município ou

para o Distrito Federal. Os conceitos tributários (Figura 22) compõem os elementos da relação tributária tais como: hipótese de incidência, base de cálculo, alíquota, imunidade, obrigação tributária, tipos de sujeição, etc.

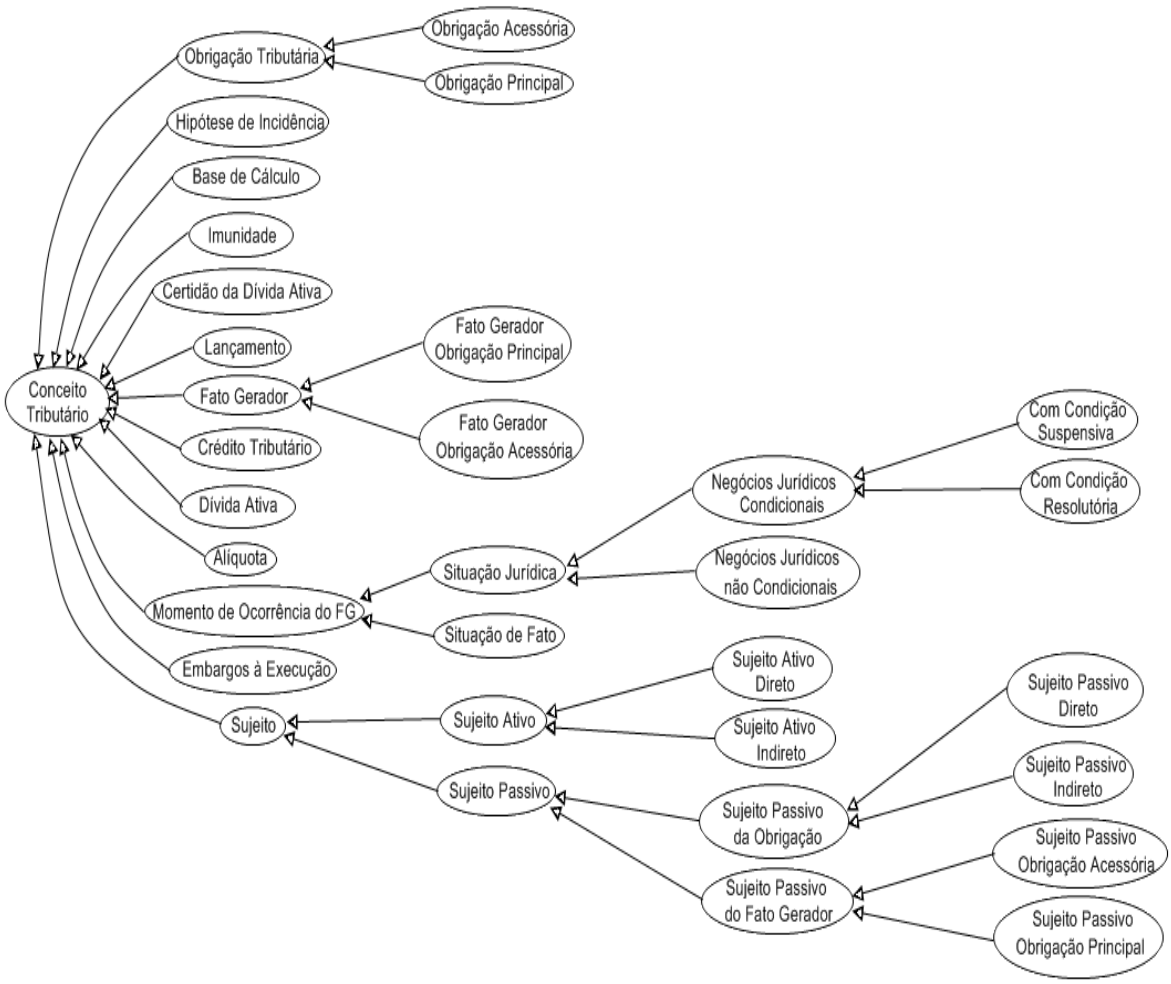

Figura 22 Hierarquia dos Elementos da relação tributária

5.2 CASOS SEMÂNTICOS NA ONTOTRIB

Os casos semânticos definidos para o contexto da ONTOTRIB baseiam-se nas hierarquias dos tributos brasileiros, dos instrumentos normativos tributários, dos conceitos tributários e das aplicações tributárias. Essas hierarquias compreendem os principais conceitos do domínio e, embora relacionadas, representam aspectos distintos da relação tributária. Por exemplo, um usuário poderá expressar o seu interesse em um instrumento normativo ou em um tributo em particular. A Tabela 12 apresenta os casos semânticos selecionados na ONTOTRIB.

Tabela 12 Conceitos-raiz dos casos semânticos da ONTOTRIB

Caso Semântico	Conceito Raiz
1	Tributo
2	Instrumento Normativo Tributário
3	Conceito Tributário
4	Aplicação Tributária

A hierarquia tributária, em termos de tributos, é aquela definida nos artigos 145, 147, 148, 149, 149-A, 153, 154, 155, 156, 194, 195, 239, 240 da Constituição Federal de 1988 e demais artigos associados. A Figura 21 mostra a hierarquia representada na ONTOTRIB, sendo que o conceito $Tributo$ é a raiz desse caso semântico, enquanto os demais conceitos mostrados na figura são os termos desse caso semântico.

O caso semântico $Instrumento\ Normativo\ Tributário$ representa todas as normas que se relacionam com o Direito Tributário, quer tratem, ou não, de tributos.

Os conceitos da relação tributária são aqueles definidos pela Doutrina e Jurisprudência (Figura 22). A raiz do caso semântico é $Conceito\ Tributário$ e os demais conceitos apresentados, que constam nas normas gerais do Direito Tributário, são os termos do caso semântico.

O último caso semântico é $Aplicação\ Tributária$ que diz respeito à esfera na qual os tributos podem ser aplicados, sendo descritos aqui instâncias referentes à união, aos estados, aos municípios e o Distrito Federal.

5.3 LUCENE

Para avaliar a aplicabilidade do modelo e ter uma noção do desempenho possível para um sistema construído a partir dele, foi criado um segundo sistema baseado nos modelos clássicos de recuperação, de modo a permitir uma comparação dos resultados obtidos. Esse tópico apresenta a API Lucene, que foi usada para a construção desse sistema. A estrutura e os principais componentes da API são brevemente descritos abaixo para que se tenha uma visão de todas as técnicas empregadas pelo sistema clássico construído.

O Apache Lucene (HATCHER, GOSPODNETIC, 2005), ou simplesmente Lucene, é uma API que disponibiliza recursos para a indexação e busca de documentos, escrito na linguagem de programação Java, através dos quais é possível construir sistemas de recuperação de informação (Figura 23). O Lucene é um software de código aberto, da empresa Apache Software Foundation, contendo apenas o núcleo do "motor" de busca, que fornece meios para criar a representação interna da consulta e dos itens de informação, indexação, análise de similaridade, retorno ao usuário dos elementos de informação recuperados, ordenados segundo uma medida de similaridade. O Lucene utiliza técnicas morfolexicais, não suportando a análise semântica, e processa documentos textuais.

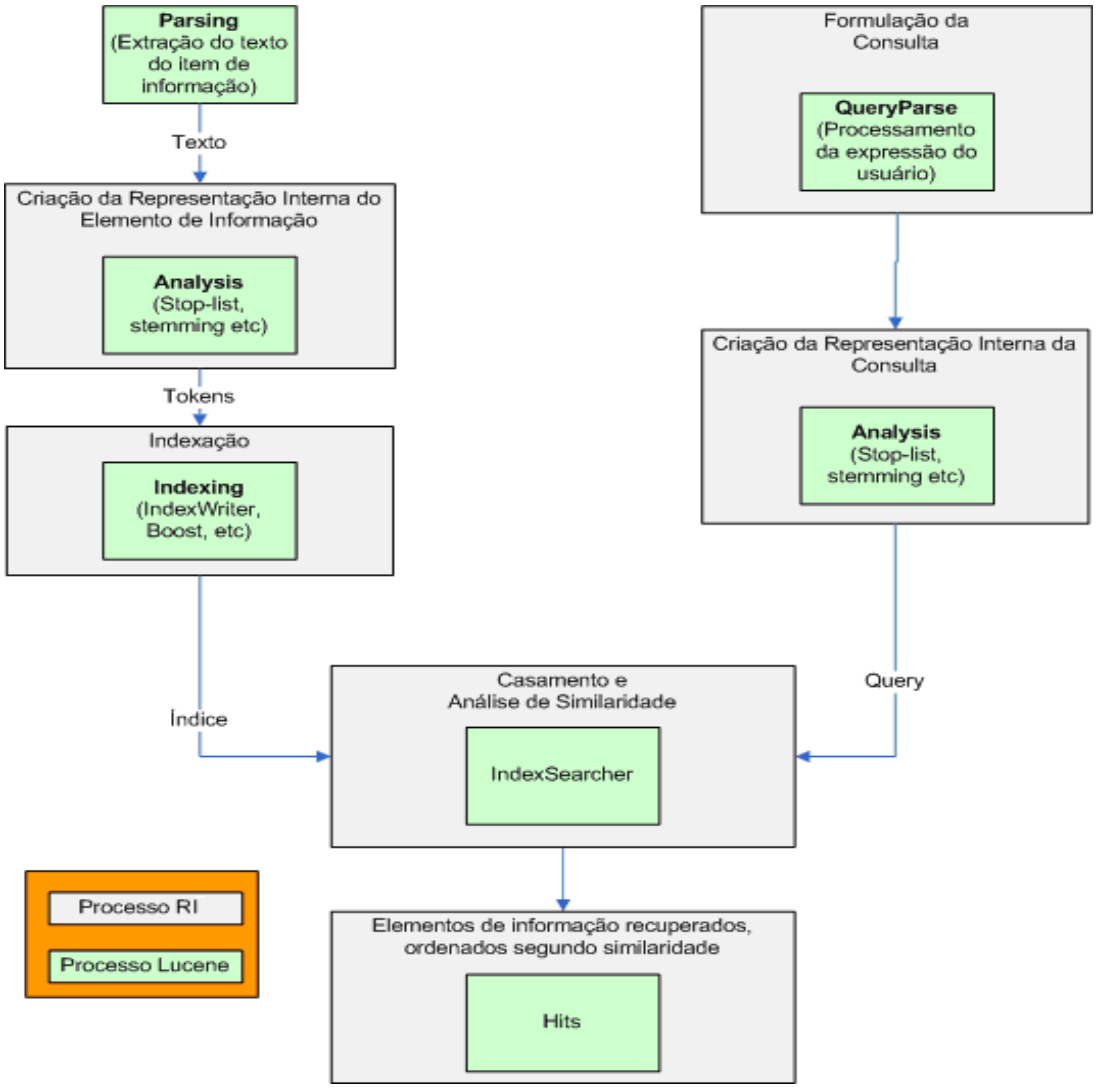

Figura 23 Processo de recuperação implementado a partir da API Lucene

5.3.1 Representação Interna

A API Lucene implementa a maior parte dos métodos necessários ao tratamento dos itens de informação comuns aos modelos clássicos de recuperação. A criação da representação interna utiliza analisadores para a criação de termos a partir do texto fornecido. Os analisadores aplicam métodos baseados na freqüência e lingüísticos ao texto, tais como:

- Retirada de caracteres de pontuação;
- Conversão do texto para caracteres minúsculos (*lowercase*);
- Retirada de afixos (*lemmatization*);
- Redução dos termos à sua forma primária ou raiz (*stemming*);
- Retirada dos termos sem significado próprio (*stop-list*);
- Possibilita a utilização de dicionários de sinônimos.

Um índice é criado a partir dos termos obtidos na fase de representação interna do documento. A estrutura de dados é uma lista invertida, ou seja, a pesquisa é feita através dos termos para assim encontrar os documentos desejados.

5.3.2 Formulação da Consulta

A consulta é formulada pelo usuário em formato texto. A consulta então passa pelo mesmo procedimento de análise para a criação de sua representação interna.

O Lucene trabalha com formatos pré-definidos para a consulta ao índice:

- Consulta por termos simples;
- Consulta por frases;
- Consulta por faixa de valores de datas ou valores numéricos;
- Consulta utilizando caracteres coringa;
- Consulta por termos com grafia similar;
- Consulta com operadores booleanos.

O Lucene também disponibiliza uma classe capaz de mapear a consulta expressa em linguagem natural para os formatos descritos acima e suportados pela

API. Na expressão da consulta o usuário pode indicar termos com um maior peso (*Boost*) indicando assim a sua maior relevância.

5.3.3 Casamento e Análise de Similaridade

O Lucene combina o modelo booleano e o modelo vetorial no processo de análise de similaridade. A recuperação dos elementos de informação do índice utiliza o casamento exato de padrões, ou seja, os termos contidos na consulta devem corresponder aos termos do índice. Uma vez recuperados, através de expressões booleanas, os documentos são classificados, segundo uma fórmula derivada do produto entre os vetores dos termos representando a consulta e os itens de informação, de acordo com:

- A freqüência dos termos no documento;
- O inverso da freqüência dos termos em todos os documentos do índice;
- Os pesos informados para os documentos e para os campos dos documentos do índice e da consulta;
- Normalização dos valores considerando o tamanho do corpo do documento.

5.4 EXPERIMENTOS

Os experimentos foram desenvolvidos com sistemas criados a partir da instanciação do modelo de recuperação proposto neste trabalho e a partir da API Lucene. Foram selecionados itens de informação da área do Direito Tributário, formuladas consultas para serem submetidas ao sistema e instanciada uma medida de similaridade semântica adaptada ao componente de análise de similaridade do modelo proposto. Os testes foram conduzidos com os seguintes objetivos:

- Criação de rotinas capazes de gerar a representação interna dos itens de informação e para a consulta;
- Validação das fórmulas propostas para a análise de similaridade;
- Avaliação das taxas de recuperação e precisão obtidas por um sistema implementado a partir do modelo.

5.4.1 Seleção dos Itens de Informação

A seleção dos itens de informação foi realizada a partir do estudo do sistema tributário brasileiro, obtido em (MACHADO, 2004), sendo escolhidos os instrumentos normativos mais representativos para diversos impostos previstos na constituição e no código tributário nacional. Foram adicionados alguns instrumentos normativos das esferas estaduais e municipais, com características similares aos instrumentos iniciais, para aumentar o tamanho da coleção. Instrumentos normativos que não versam sobre o Direito Tributário foram também incluídos para analisar a incidência de documentos não relevantes na efetividade dos sistemas utilizados nos experimentos. A lista desses instrumentos é mostrada na Tabela 13.

Tabela 13 Instrumentos Normativos usados no estudo de caso

Índice	Instrumento Normativo	Tributo
IN_1	Decreto-Lei nº 1.578	Dispõe sobre o imposto de exportação
IN_2	Lei Complementar nº 116	Dispõe sobre o ISSQN
IN_3	Lei nº 11.527 CE	Institui o ITCD no estado do Ceará
IN_4	Lei nº 14.937 MG	Dispõe sobre o IPVA em Minas Gerais
IN_5	Lei nº 9.393	Dispõe sobre o ITR
IN_6	Decreto-Lei nº 406	Estabelece normas gerais para o ISSQN e ICMS
IN_7	Decreto-Lei nº 37	Dispõe sobre o Imposto de Importação
IN_8	Lei nº 13.417 CE	Dispõe sobre o ITCD no estado do Ceará
IN_9	Lei nº 12.397 CE	Altera dispositivos do IPVA no estado do Ceará
IN_{10}	Lei nº 12.023 CE	Dispõe sobre o IPVA no estado do Ceará
IN_{11}	Lei nº 12.391 Campinas	Dispõe sobre o ITBI no município de Campinas
IN_{12}	Lei nº 8.216 PR	Dispõe sobre o IPVA no estado do Paraná
IN_{13}	Lei nº 43.981 MG	Dispõe sobre o ITCD em Minas Gerais
IN_{14}	Lei nº 9.430	Dispõe sobre a legislação tributária federal
IN_{15}	Lei nº 8.927 PR	Dispõe sobre o ITCD no estado do Paraná
IN_{16}	Lei nº 3.944 Curitiba	Dispõe sobre o ISSQN no município de Curitiba
IN_{17}	Lei nº 9.428	Dispõe sobre o orçamento fiscal da União
IN_{18}	Lei nº 11.497	Dispõe sobre a organização da Presidência da República
IN_{19}	Lei nº 11.476	Dispõe sobre a profissão de enólogo
IN_{20}	Lei nº 11.428	Dispõe sobre a proteção do Bioma Mata Atlântica

A instanciação desses instrumentos normativos na ONTOTRIB foi realizada de modo semi-automático. Foi desenvolvida uma rotina para extrair os dados de identificação e os dispositivos de cada instrumento e incluí-los na base de conhecimento de modo automático. Para gerar instâncias dos conceitos tributários, identificar os tributos referenciados e a aplicação jurídica do instrumento foi desenvolvida uma interface com o usuário que permite a ele examinar os dispositivos e criar as instâncias necessárias na base de conhecimento (Anexo I). A geração dos conceitos foi feita com a ajuda de um especialista no domínio tributário, sendo que o uso da rotina desenvolvida conseguiu reduzir o tempo de inclusão de um instrumento normativo em até 70% quando comparada à instanciação manual, utilizando-se um formulário padrão da ferramenta Protégé (GENNARI, MUSEN, FERGERSON, 2002).

A base de conhecimento criada na ferramenta Protégé permite instanciar instrumentos normativos de acordo com a estrutura desses instrumentos definida pela ONTOTRIB (Figura 24).

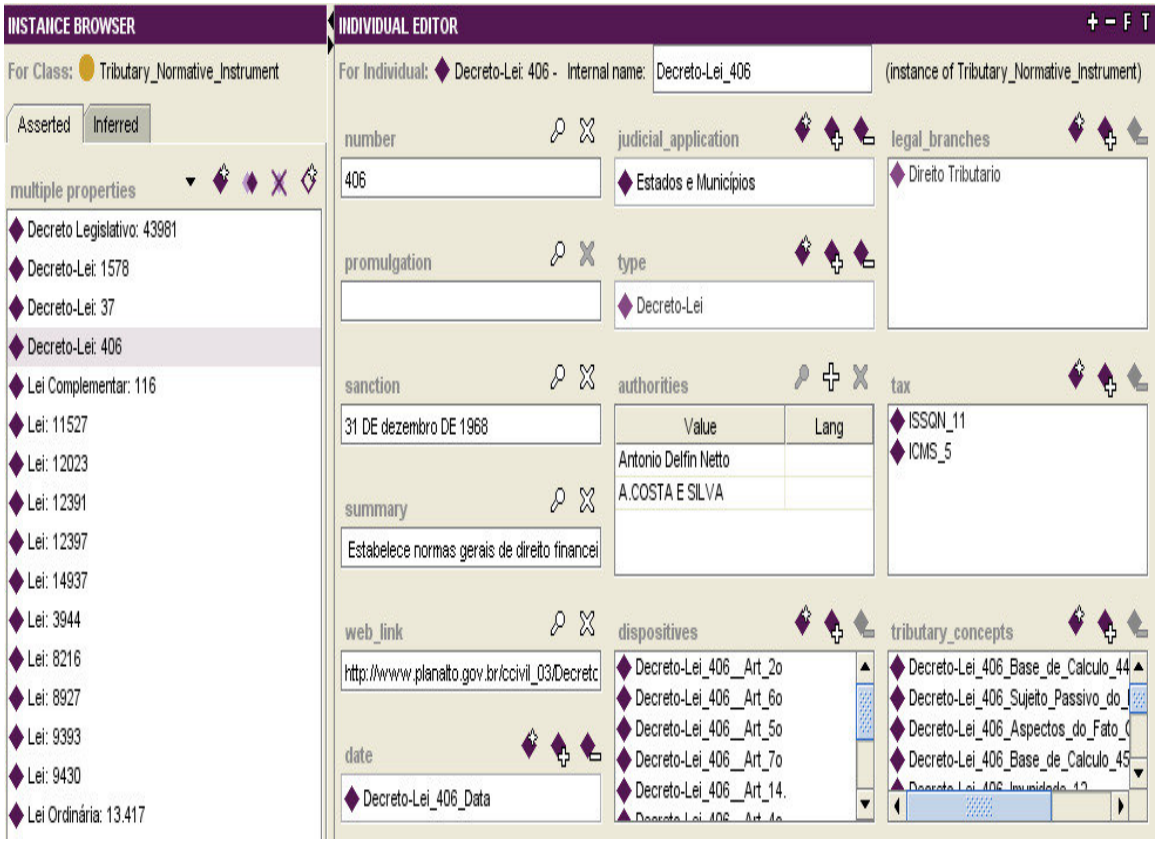

Figura 24 Estrutura do Decreto-Lei 406

A representação interna dos itens de informação segundo o modelo proposto instancia os instrumentos normativos agrupando os elementos da base de conhecimento em pares de conjuntos com elementos pertencentes a diferentes casos semânticos (Figura 25).

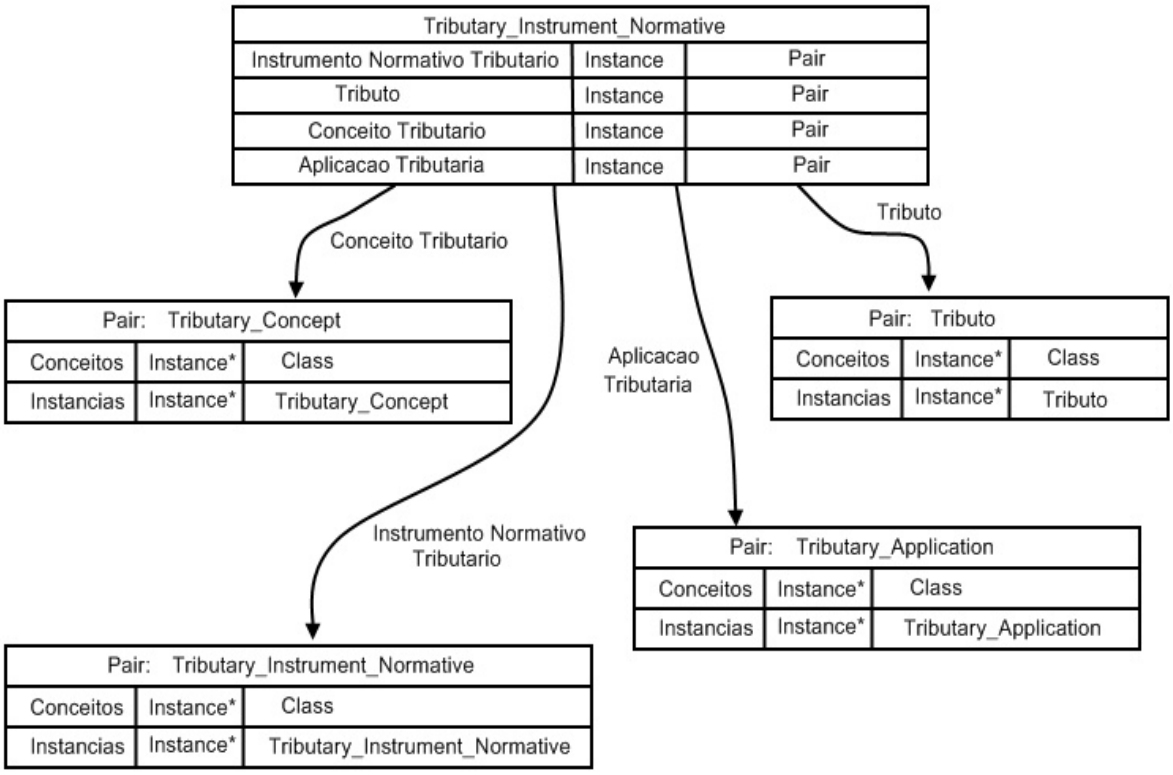

Figura 25 Estrutura da representação interna do sistema baseado no modelo proposto

A criação da representação interna dos itens de informação foi realizada por uma segunda rotina que, alimentada com os casos semânticos da Tabela 12, organiza os conceitos e instâncias dos itens de informação de acordo com o proposto pelo modelo apresentado neste trabalho. A representação interna criada para um instrumento normativo por essa segunda rotina é exemplificada na Figura 26.

O sistema desenvolvido sobre a API LUCENE utilizou uma base de itens de informação formada pelos mesmos instrumentos normativos mostrados na Tabela 13, porém representados através de vetores de palavras-chaves extraídas do texto dos instrumentos.

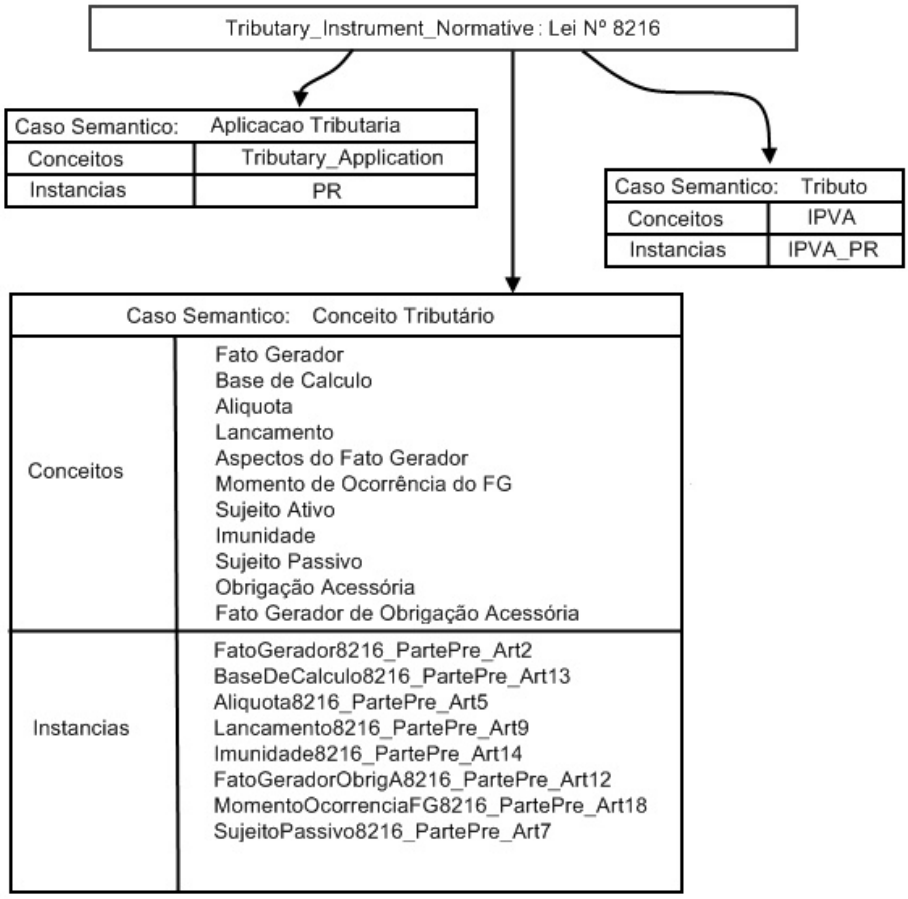

Figura 26 Parte da representação interna da Lei nº 8216

5.4.2 Consultas

Foram criadas 8 consultas para uma avaliação do sistema construído. Os termos usados nas consultas referenciam diferentes instrumentos normativos da base de conhecimento e as consultas utilizaram tanto palavras-chave, quanto perguntas mais específicas inspiradas em questões usadas em concursos públicos da área do Direito. Todas as consultas foram instanciadas manualmente, com a identificação dos seus conceitos e instâncias sendo realizada pelo especialista do domínio, sendo criadas representações internas através da interface da ferramenta Protégé, conforme o modelo proposto, similares às dos itens de informação. O texto das consultas submetidas aos sistemas, bem como os instrumentos normativos considerados mais relevantes para cada consulta, são mostrados na Tabela 14.

A representação interna criada para uma consulta é exemplificada na Figura 27, mostrando os conceitos e instâncias que representam os elementos de interesse do usuário.

Tabela 14 Consultas submetidas aos sistemas de recuperação

Índice	Consulta	Instrumentos Relevantes
Q_1	Impostos estaduais que incidem sobre a propriedade de um bem imóvel	$IN_3, IN_5, IN_8, IN_{11}, IN_{13}, IN_{15}$
Q_2	Imunidade IPVA	$IN_4, IN_9, IN_{10}, IN_{12}$
Q_3	Quais alíquotas são aplicáveis ao IPVA no Ceará?	IN_{10}, IN_9
Q_4	ITCD Paraná	IN_{15}
Q_5	Como é fixada a alíquota do ISSQN?	IN_2, IN_6, IN_{16}
Q_6	Bens móveis compõem a base de cálculo do ITCD no Ceará?	IN_8, IN_3
Q_7	Fato gerador dos impostos cobrados no estado de Ceará	$IN_3, IN_8, IN_9, IN_{10}$
Q_8	Imposto sobre serviço nos municípios de Minas Gerais	IN_2, IN_6

As consultas (Tabela 14) foram também submetidas ao sistema desenvolvido sobre a API LUCENE, porém foram representadas através de vetores de palavras-chaves extraídas do texto dos instrumentos.

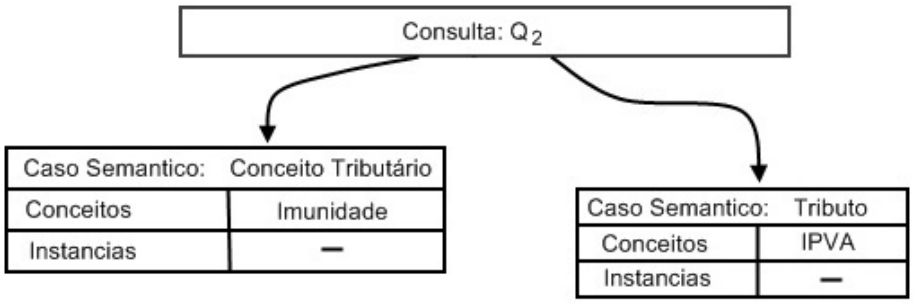

Figura 27 Representação interna da consulta Q_2 da Tabela 14

5.4.3 Medida de Similaridade

A medida de similaridade entre dois conceitos c e d que pertençam a uma mesma hierarquia, característica encontrada entre os conceitos de um mesmo caso semântico, é definida em (DRUMOND, GIRARDI, SILVA, 2008), como sendo:

Definição 25: Medida de similaridade proposta em (DRUMOND, GIRARDI, SILVA, 2008).

$$sim(C, D) = \frac{2 \cdot |C_{\sqsubseteq} \cap D_{\sqsubseteq}|}{|C_{\sqsubseteq}| + |D_{\sqsubseteq}|}$$

onde:

- $c_{\sqsubseteq}$ é o conjunto de todos os conceitos e, tal que $e \sqsupseteq c$;
- $d_{\sqsubseteq}$ é o conjunto de todos os conceitos e, tal que $e \sqsupseteq d$;
- $|c_{\sqsubseteq}|$ é o número de elementos do conjunto $c_{\sqsubseteq}$;
- $|d_{\sqsubseteq}|$ é o número de elementos do conjunto $d_{\sqsubseteq}$;
- $|c_{\sqsubseteq} \cap d_{\sqsubseteq}|$ é o número de elementos na interseção dos conjuntos $c_{\sqsubseteq}$ e $d_{\sqsubseteq}$;

em que a interseção dos conjuntos $c_{\sqsubseteq}$ e $d_{\sqsubseteq}$ no numerador da equação da Definição 25 representa as características comuns e compartilhadas pelos conceitos e o denominador representa a soma das características de cada conceito em particular. Tomando a hierarquia mostrada na Figura 28, podemos exemplificar a aplicação da medida de similaridade entre conceitos.

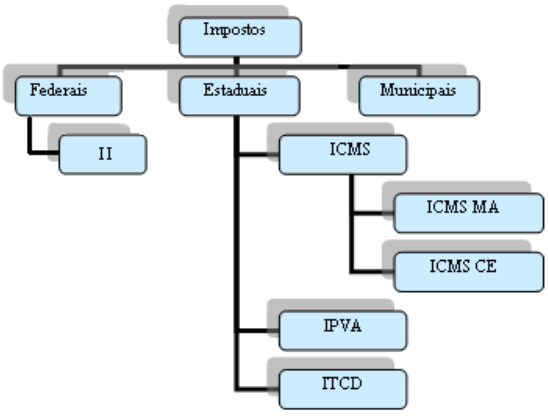

Figura 28 Parte da Hierarquia de tributos da ONTOTRIB

O cálculo é realizado conforme se segue:

$$ICMS\ CE_{\sqsubseteq} = \{ICMS\ CE, ICMS, Estaduais, Impostos\}$$
$$ITCD_{\sqsubseteq} = \{ITCD, Estaduais, Impostos\}$$
$$ICMS\ CE_{\sqsubseteq} \cap ITCD_{\sqsubseteq} = \{Estaduais, Impostos\}$$

$$sim(ICMS\ CE_{\sqsubseteq}, ITCD_{\sqsubseteq}) = \frac{2 * |ICMS\ CE_{\sqsubseteq} \cap ITCD_{\sqsubseteq}|}{|ICMS\ CE_{\sqsubseteq}| + |ITCD_{\sqsubseteq}|} = \frac{2 * 2}{4 + 3} = 0{,}5714$$

Para calcular a distância semântica entre duas instâncias i_1 e i_2, utilizando-se a mesma medida de similaridade, considerou-se que uma instância é uma especialização do seu conceito mais específico. Assim, instâncias idênticas são consideradas como conceitos idênticos e obtêm o valor máximo de similaridade por estarem na mesma posição da hierarquia. Por outro lado, instâncias distintas de um mesmo conceito terão sempre o mesmo valor de similaridade.

O sistema construído a partir da instanciação do modelo proposto utilizou essa medida de similaridade na etapa da análise de similaridade instanciando as equações Sim_c e Sim_i previstas nas definições 19 e 20. A especificação dessas equações permite a aplicação da equação da Definição 22 para calcular a similaridade entre uma consulta e um item de informação. Para exemplificar essa etapa do processo de recuperação de informação e a aplicação dos componentes do modelo no sistema resultante de sua instanciação é apresentado o cálculo de similaridade entre o item de informação IN_{12} (Tabela 13) e a consulta Q_2 (Tabela 14), considerando $\omega = 1$. Como a consulta e o item de informação possuem 2 casos semânticos em comum, $Tributo$ e $Conceito\ Tributário$, a similaridade entre eles é dada por:

$$Sim(IN_{12}, Q_2)$$
$$= \left(\frac{1}{1+1+1+1}\right)$$
$$* \left(\left(1 * \frac{sim_c(\{IPVA\}, \{IPVA\})}{1+0}\right)\right.$$
$$+ \left(1\right.$$
$$\left. * \frac{sim_c(\{Fato\ Gerador,\ \ Base\ de\ Calculo, \ldots, Fato\ Gerador\ de\ Obrigacao\ Acessoria\}, \{Imunidade\})}{1+0}\right)\Bigg)$$

A aplicação da medida de similaridade para a comparação dos pares do caso semântico $Tributo$ resulta no valor máximo, uma vez que os conceitos são idênticos: $sim_c(\{IPVA\}, \{IPVA\}) = 1$. Os pares do caso semântico $Conceito\ Tributário$ são comparados um a um segundo a Definição 23, encontrando-se, para cada

conceito da consulta, o maior valor de similaridade entre os conceitos da representação do item de informação, ou seja:

$$max\begin{pmatrix} sim_c(Fato\ Gerador, Imunidade), sim_c(Base\ de\ Calculo, Imunidade), \dots, \\ sim_c(Fato\ Gerador\ de\ Obrigacao\ Acessoria, Imunidade) \end{pmatrix}$$

A Tabela 15 lista alguns dos valores encontrados para a medida de similaridade entre os conceitos das representações internas do item de informação e da consulta escolhidos. Como o conceito $Imunidade$ é o único conceito da representação interna da consulta para o caso semântico $Conceito\ Tributário$, e também faz parte da representação do instrumento normativo IN_{12}, obtêm-se o valor 1 para a similaridade entre os pares desse caso semântico no exemplo.

Tabela 15 Valores de similaridade entre conceitos das representações internas

Aplicação da Medida de Similaridade
$sim_c(Momento\ de\ Ocorrencia\ do\ FG, Imunidade) = 0{,}67$
$sim_c(Base\ de\ Calculo, Imunidade) = 0{,}67$
$sim_c(Sujeito\ Ativo\ Direto, Imunidade) = 0{,}50$
$sim_c(Sujeito\ Passivo, Imunidade) = 0{,}57$
$sim_c(Imunidade, Imunidade) = 1{,}00$

Por fim, a similaridade final entre o item de informação IN_{12} e a consulta Q_2 é dado pelo valor:

$$Sim(IN_{12}, Q_2) = \left(\frac{1}{1+1+1+1}\right) * \left(\left(1 * \frac{1}{1+0}\right) + \left(1 * \frac{1}{1+0}\right)\right) = \frac{1}{4} * 2 = 0{,}50$$

O exemplo demonstrou como os valores de similaridade entre os documentos são obtidos pelo sistema criado a partir do modelo proposto através da comparação de um instrumento normativo com a consulta mais simples dentre as propostas no trabalho para a avaliação dos sistemas. A avaliação dos resultados,

bem como uma análise dos valores obtidos no experimento são discutidos nas próximas duas seções.

5.5 RESULTADOS

A avaliação preliminar realizada teve como objetivo verificar a efetividade dos sistemas construídos através da comparação das taxas de revocação e precisão obtidas para as consultas formuladas (Tabela 14). A Tabela 16 lista os valores de precisão para diferentes percentuais de revocação baseados nos resultados obtidos pelo sistema construído a partir do modelo proposto. Todos os casos semânticos foram considerados como tendo pesos idênticos no sistema, sem a utilização do peso ω previsto no modelo, ou seja, $\omega = 1$.

Tabela 16 Valores de precisão do sistema instanciado a partir do modelo proposto

Consulta	20%	40%	60%	80%	100%	% Recall
Q_1	0,67	0,67	0,6	0,46	0,38	1
Q_2	1	0,67	0,4	0,31	0,25	1
Q_3	0,67	0,33	0,2	0,15	0,13	1
Q_4	0,33	0,17	0,1	0,08	0,06	1
Q_5	1	0,5	0,3	0,23	0,19	1
Q_6	0,67	0,33	0,2	0,15	0,13	1
Q_7	1	0,5	0,3	0,23	0,25	1
Q_8	0,67	0,33	0,2	0,15	0,13	1
Média	0,75125	0,4375	0,2875	0,22	0,19	1

Os resultados obtidos pelo sistema construído a partir da API Lucene, que se baseia no modelo do espaço vetorial são mostrados na Tabela 17.

Tabela 17 Valores de precisão do sistema baseado na API Lucene

Consulta	20%	40%	60%	80%	100%	% Recall
Q_1	0,5	0,71	0,55	0,43	0,33	1
Q_2	1	0,8	0,57	0,4	0,33	1
Q_3	0,33	0,33	0,22	0,17	0,13	1
Q_4	0	0	0	0	0,2	1
Q_5	0,5	0,2	0,14	0,1	0,08	0,33
Q_6	0,33	0,29	0,2	0,14	0,12	1
Q_7	0,33	0,43	0,4	0,29	0,24	1
Q_8	0,25	0,29	0,18	0,14	0,11	1
Média	0,405	0,38125	0,2825	0,20875	0,1925	0,91625

Ambos os sistemas considerados apresentaram percentuais razoáveis de recuperação, ou seja, conseguiram, na maioria dos casos, recuperar todos os itens relevantes para percentuais baixos de precisão. A precisão encontrada foi maior no sistema baseado no modelo proposto em 6 das 8 consultas submetidas no teste, sendo que, os valores encontrados para as duas consultas em que o sistema proposto se saiu pior foram próximos dos valores encontrados no sistema baseado na API Lucene. O gráfico de revocação x precisão criado a partir da média de desempenho dos sistemas é mostrado na Figura 29.

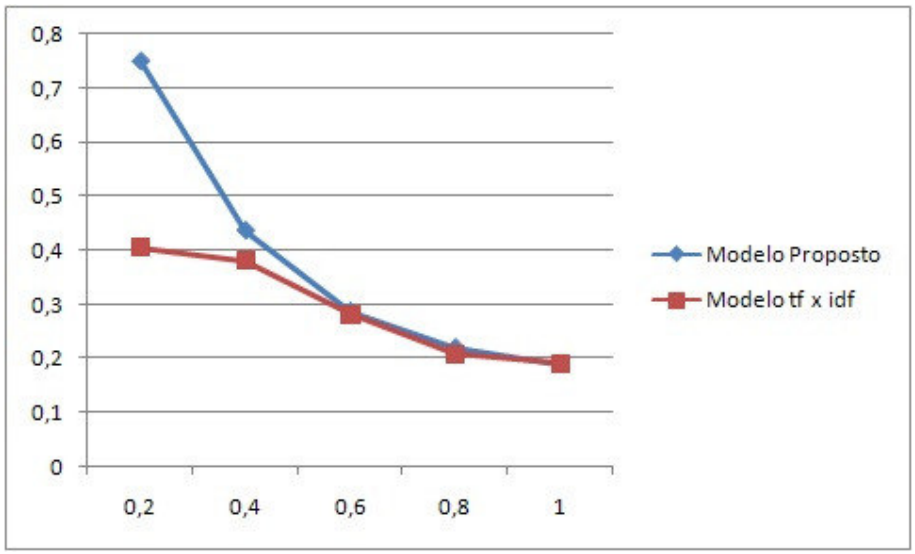

Figura 29 Gráfico revocação x precisão dos sistemas comparados

A diferença entre os valores expostos nos gráficos é mais acentuada nas taxas de revocação entre 20% e 40%, em que a precisão do sistema baseado no modelo proposto alcançou valores superiores aos do sistema baseado na API Lucene. A queda mais significativa entre os valores de precisão a 20% e 40% da taxa de revocação, mostrada no gráfico mostrado na Figura 29, é conseqüência do número reduzido de instrumentos normativos relevantes para cada consulta em relação ao número total de instrumentos utilizados. Por exemplo, o sistema baseado no modelo proposto recuperou 15 documentos para a consulta de índice Q6, que possui apenas 2 itens de informação considerados relevantes na coleção. Avaliando-se o percentual de precisão (Tabela 16) para essa consulta à uma taxa de revocação de 20%, o que representa os 3 primeiros itens recuperados, é encontrado o valor de 0,67, pois foram encontrados 2 itens de informação relevantes dentre os 3

itens de informação recuperados. Avaliando-se o percentual de precisão para essa consulta à uma taxa de revocação de 40%, o que representa os 6 primeiros itens recuperados, a quantidade de itens relevantes permanece a mesma, apenas 2 itens, enquanto a quantidade de itens recuperados dobrou, gerando uma queda no percentual de precisão obtido. Ao consideramos taxas de revocação superiores à 40%, os valores de precisão dos sistemas são próximos, pois a quantidade de documentos recuperados pelos sistemas é praticamente a mesma e todos os itens relevantes já constam no conjunto de itens recuperados.

5.6 CONSIDERAÇÕES FINAIS

Esse capítulo apresentou um estudo de caso com o modelo de recuperação proposto na área do Direito Tributário através da implementação de um sistema de recuperação baseado nesse modelo. O estudo cumpriu os objetivos de criar os mecanismos de instanciação dos documentos a partir dos casos semânticos segundo a representação interna proposta no modelo, gerar uma ordenação para os itens de informação recuperados e obteve bons índices de precisão nos resultados.

Alguns aspectos não abordados diretamente na definição do modelo e utilizados na construção do sistema baseado no modelo proposto também contribuíram para a obtenção de altos índices de precisão em comparação com o sistema baseado no modelo do espaço vetorial. A escolha dos casos semânticos apropriados e a correta instanciação das consultas e dos itens de informação contribuíram para uma maior efetividade do sistema baseado no modelo proposto.

Os casos semânticos escolhidos estão relacionados a conceitos freqüentes e comuns aos instrumentos normativos jurídico-tributários. O caso semântico $Instrumento\ Normativo\ Tributário$ não teve conceitos ou instâncias identificados nas consultas e não contribuiu na geração dos valores de similaridade. Os casos semânticos $Tributo$ e $Aplicação\ Tributária$ por sua vez, possuem conceitos e instâncias comuns a diversos itens de informação e a diversas consultas, sendo mais importantes no cálculo de similaridade. Por fim, o caso semântico $Conceito\ Tributário$, identifica conceitos presentes nos dispositivos tributários e gera valores para a análise de similaridade baseado principalmente na

ocorrência dos conceitos uma vez que os dispositivos são específicos de cada instrumento instanciado.

Uma melhoria importante no estudo de caso seria a instanciação de elementos da relação jurídico-tributária mais específicos presentes nos textos. Por exemplo, em um instrumento normativo que trate do imposto sobre a propriedade de veículos (IPVA), é importante identificar os tipos de veículo sujeitos à cobrança do imposto, a classificação dos veículos, as penalidades previstas para o não pagamento do imposto, entre outros. Com isso, seriam identificadas características comuns aos itens de informação que tratam do IPVA e o caso semântico $Conceito\ Tributário$ poderia exercer um papel mais importante comparando uma quantidade maior de instâncias. Um exemplo de como a instanciação de mais elementos do domínio poderia melhorar o desempenho do sistema é vista ao analisarmos os resultados obtidos para a consulta Q_1 da Tabela 14, em que $bem\ imóvel$ é um termo identificado pelo sistema de busca por palavras-chave, enquanto não é considerado pelo sistema baseado no modelo proposto, gerando assim uma menor precisão. Essa limitação é, em parte, fruto da generalidade da ONTOTRIB que é restrita aos elementos do domínio tributário puro. Além disso, capturar esse tipo de informação depende de técnicas de instanciação automática e efetiva de ontologias capazes de instanciar os elementos textuais de forma eficiente.

O povoamento de ontologias é uma atividade importante para a aquisição de conhecimento e a construção de sistemas baseados no conhecimento, à exemplo do sistema baseado no modelo proposto exposto neste estudo de caso. Esta atividade é definida em (BONTCHEVA, CUNNINGHAM, 2003) como sendo a extração de informação dirigida por ontologia, e consiste de métodos para a identificação de conceitos, instâncias e relacionamentos de uma base de conhecimento em fontes de informação desestruturadas de forma semi-automática ou automática. A extração de informação inclui métodos para reconhecer nomes próprios, nomes de lugares, datas e outros elementos em textos, utilizando análise lingüística e técnicas comuns à mineração de texto para a classificação e agrupamento dos elementos identificados (BUITELAAR, DECLERK, 2003). O povoamento de ontologias, por sua vez, visa atender ontologias variadas, com grande número de classes e com relações complexas. Por isso, incorpora, além dos métodos oriundos da extração de informação, métodos estatísticos, o uso de

padrões e regras de raciocínio para identificar nos textos os elementos da base de conhecimento. Devido à sua abrangência, o detalhamento e desenvolvimento destes métodos para o povoamento de ontologias está fora do escopo deste trabalho, sendo objeto de pesquisa de diversos autores (ALANI et al., 2003) (BUITELAAR, DECLERK, 2003) (CIMIANO, 2006) (SUCHANEK, IFRIM, WEIKUM, 2006) (TANEV, MAGNINI, 2006) e também do grupo GESEC.

O estudo apresentado serviu para exemplificar a aplicação do modelo proposto através de sua instanciação em um sistema de recuperação para um domínio específico. Neste estudo, as atividades de povoamento de ontologia foram realizadas de forma semi-automática, sendo que, as consultas foram instanciadas manualmente, com a identificação das instâncias sendo feita por um especialista. A análise dos resultados obtidos pelo sistema avaliou os seus índices de precisão, pois são mais representativos para coleções de tamanho reduzido como a utilizada neste estudo, com apenas 20 itens de informação. Os resultados se mostraram promissores uma vez que alcançaram maior efetividade do que o sistema baseado no modelo do espaço vetorial. Isto mostra a aplicabilidade do modelo proposto e a possibilidade de incrementar a precisão com o uso dos casos semânticos associados às técnicas de representação semântica.

6. CONCLUSÕES

Este trabalho contribui com a especificação de um modelo de recuperação de informação que utiliza as estruturas de representação baseadas no conhecimento aplicáveis à Web Semântica.

O modelo concebido trabalha com o conceito de casos semânticos representando grupos de interesse do usuário. O contexto definido pelos casos semânticos provê uma maneira de especificar e quantificar o interesse do usuário por partes da ontologia do domínio. Os elementos de uma base de conhecimento são mapeados em função dos casos semânticos e utilizados para gerar a representação interna dos itens de informação e da consulta, e também na análise de similaridade especificada.

O modelo proposto foi avaliado através de um estudo de caso na área do Direito Tributário, com o desenvolvimento de um sistema de recuperação de informação baseado nesse modelo. A base de conhecimento utilizou uma ontologia desenvolvida no grupo GESEC, a ONTOTRIB, que descreve os elementos do domínio, capaz de representar instrumentos normativos jurídico-tributários.

6.1 RESULTADOS E CONTRIBUIÇÕES DA PESQUISA

As principais contribuições desta pesquisa foram:

- Análise dos principais sistemas de recuperação de informação desenvolvidos com estruturas de representação da Web Semântica para suportar as fases de um processo de recuperação e para superar as limitações encontradas nos sistemas tradicionais de recuperação de informação por palavras-chave baseados no modelo do espaço vetorial. Essa análise apontou que a representação interna dos elementos de informação e da consulta baseada em conceitos e instâncias de uma base de conhecimento é uma técnica importante para aumentar a precisão dos sistemas construídos. Não havendo modelos de recuperação específicos para a recuperação de informação baseada no conhecimento, vários desses sistemas utilizam variações do modelo do espaço vetorial. A avaliação desses sistemas

conduzida por seus autores evidencia que a instanciação de ontologias é um problema recorrente também nesses sistemas, dificultando uma avaliação efetiva dos mesmos.

- Definição dos casos semânticos para representar grupos de interesse do usuário em uma ontologia de domínio. Os casos semânticos exploram relações de hierarquia em que um conceito inclui ou é incluído por outros conceitos. Tal característica é apropriada para a construção de um modelo de recuperação que utilize medidas de similaridade semânticas, em geral, baseadas na distância entre os conceitos em uma hierarquia.
- Definição de um modelo de recuperação utilizando estruturas de representação baseadas no conhecimento. O modelo de recuperação especifica os componentes em termos dos casos semânticos definidos para o domínio. O uso dos casos semânticos privilegia as relações entre conceitos e instâncias de um mesmo grupo de interesse do usuário, de modo a aumentar a precisão dos resultados, evitando comparações entre elementos díspares.
- Desenvolvimento de uma ferramenta para a instanciação de instrumentos jurídicos normativos a ser utilizada no grupo GESEC. A instanciação manual destes instrumentos é uma tarefa árdua e propensa a erros. O uso de uma ferramenta auxilia o usuário nessa tarefa, melhorando a legibilidade ao selecionar os dispositivos válidos, instanciar os dispositivos, identificar os conceitos e relações entre os mesmos.
- Avaliação do modelo de recuperação proposto através de um estudo de caso. O estudo conduzido a partir da instanciação do modelo em um sistema de recuperação aplicado ao domínio jurídico tributário apontou um aumento da precisão em relação a um sistema baseado em palavras-chave para um conjunto-teste de consultas. Apesar disso, novos experimentos devem ser conduzidos para uma melhor avaliação do modelo. A escalabilidade do modelo é difícil de ser avaliada com uma quantidade limitada de itens na coleção de documentos e o uso de técnicas de aprendizagem (DRUMOND, GIRARDI, 2008) e

povoamento de ontologias ajudarão a superar o reconhecido gargalo da instanciação dos itens de informação para as ferramentas desenvolvidas baseadas nas técnicas de representação do conhecimento.

6.2 TRABALHOS FUTUROS

Pela sua complexidade, o escopo desta pesquisa foi limitado em vários aspectos que deverão ser abordados em futuros trabalhos, entre eles:

- A ferramenta de instanciação semi-automática de instrumentos jurídico-normativos desenvolvida neste trabalho permitiu apenas a criação de uma coleção com um número limitado de elementos de informação. Um estudo de caso mais abrangente iria requerer a disponibilidade de ferramentas para a instanciação automática de uma ontologia, tópico que merece ser pesquisado através das técnicas de povoamento de ontologias, objeto de pesquisa recente do grupo GESEC.
- Extensão da ontologia ONTOTRIB com novos conceitos e relacionamentos do domínio tributário. O aumento na variabilidade dos conceitos presentes no domínio contribui para uma melhoria no desempenho da recuperação.
- Utilização do modelo em novos sistemas com diferentes técnicas de instanciação, aplicadas a outros domínios e utilizando outras medidas de similaridade disponíveis na literatura.
- Extensão do modelo, e posterior avaliação, para a área da filtragem de informação que, notadamente, incorpora diversas técnicas da área de recuperação de informação em seus processos.

REFERÊNCIAS

1. ALANI, H., KIM, S., MILLARD, D. E., WEAL, M. J., HALL, W., LEWIS, P. H. and SHADBOLT, N. R. **Automatic Ontology-Based Knowledge Extraction from Web Documents**, IEEE Intelligent Systems, Volume 18, nº 1, pp. 14-21, 2003.

2. ANTONIOU, G. and HARMELEN F. V. **A Semantic Web Primer.** MIT Press, 2004.

3. ARAÚJO, Isabel., DRUMOND, Lucas., MARIANO, Roberval., GIRARDI, Rosário. **ONTOJURIS e ONTOTRIB: ontologias para a modelagem do conhecimento jurídico.** SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL UFF – IACS - Departamento de Ciência da Informação – Niterói 08/2008. Disponível em <http://www.uff.br/ontologia/artigos/314.pdf>. Acessado em 27/10/2008.

4. BAADER, F., CALVANESE, D., MCGUINNESS, D., NARDI, D., and PATEL-SCHNEIDER, P. **The Description Logic Handbook.** Cambridge University Press, 2003.

5. BAEZA-YATES, R. and RIBEIRO-NETO, B. **Modern Information Retrieval.** Addison Wesley, 1998.

6. BECHHOFER, S., van HARMELEN, F., HENDLER, J., HORROCKS, I., MCGUINNESS, D. L., PATEL-SCHNEIDER, P. F., and STEIN, L. A. **OWL Web Ontology Language Reference.** W3C Recommendation 10 February 2004. Disponível em <http://www.w3.org/TR/owl-ref/>. Acessado em 25/06/2008.

7. BECKETT, D. **Resource Description Framework (RDF) Model and Syntax Specification.** W3C Recommendation 10 February 2004. Disponível em <http://www.w3.org/TR/rdf-syntax-grammar/>. Acessado em 25/06/2008.

8. BELKIN, N. J. and CROFT, W. B. **Information Retrieval and Filtering: Two Sides of the Same Coin?** Communications of the ACM, volume 35, nº 12, December, 1992.

9. BENZ, D. and HOTHO, A. **Position paper: Ontology learning from folksonomies.** In Proceedings of LWA ’07, pp. 109–112, 2007.

10. BERNERS-LEE, T., FIELDING, R.T. and MASINTER L. **“Uniform Resource Identifiers (URI): Generic Syntax”**, RFC 2396, 08/1998, Disponível em <http://www.ietf.org/rfc/rfc2396.txt>. Acessado em 25/06/2008.

11. BERNERS-LEE, T., HENDLER, J., LASSILA, O. **The Semantic Web.** Scientific American, volume 284, nº 5, pp. 34-43, 2001.

12. BONINO, D., CORNO, F., and FARINETTI, F. **Dose: a distributed open semantic elaboration platform.** Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, pp. 580, 2003.

13. BONTCHEVA, K., and CUNNINGHAM, H. **The Semantic Web: A New Opportunity and Challenge for Human Language Technology**. In Proceedings of Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference, Sanibel Island, pp. 89-96, 2003.

14. BRAY, T., PAOLI, J., SPERBERG-MCQUEEN, C. M., MALER, E., YERGEAU, F. **Extensible Markup Language (XML).** W3C Recommendation 04 February 2004. Disponível em <http://www.w3.org/TR/REC-xml>. Acessado em 25/06/ 2008.

15. BRICKLEY, D. GUHA, R. V. **RDF Vocabulary Description Language 1.0: RDF Schema.** W3C Recommendation 10 February 2004. Disponível em <http://www.w3.org/TR/rdf-schema/>. Acessado em 25/06/ 2008.

16. BRIN, S., PAGE, L. **The anatomy of a large-scale hypertextual Web search engine.** Computer Networks and ISDN Systems, Volume 30, Issue 1-7, pp. 107-117, 1998.

17. BUITELAAR, P., DECLERCK, T. **Linguistic Annotation for the Semantic Web**. Annotation for the Semantic Web, S. Handschuh, and S. Staab eds., IOS Press, 2003.

18. CAPEZ, F., MALTINTI, E. R. **Direito Tributário (Perguntas e Respostas).** *Pág. 48*. Editora Saraiva, 2006.

19. CHIRITA, P., COSTACHE, S., NEJDL, W., and PAIU, R. **Beagle++: Semantically enhanced searching and ranking on the desktop.** The Semantic Web: Research and Applications, 2006.

20. CIMIANO, P. **Ontology Learning and Population from Text: Algorithms, Evaluation and Applications.** Springer-Verlag New York, Inc, 2006.

21. CIMIANO, P., HANDSCHUH, S., and STAAB, S. **Towards the self–annotating web.** In Proceedings of the 13th WWW Conference, ACM, New York, pp. 462–471, 2004.

22. D'AMATO, C., FANIZZI, N., and ESPOSITO, F. **A semantic similarity measure for expressive description logics.** In Proceedings of Convegno

Italiano di Logica Computazionale (CILC05), Rome, Italy. Ed. A. Pettorossi, 2005.

23. DACONTA, M. C., OBRST, L. G., SMITH K. T. **The Semantic Web: A guide to the future of XML, Web Services, and Knowledge Management.** Wiley Publishing Inc, 2003.

24. DAVIES, J., KROHN, U., and WEEKS, R. **Quizrdf: search technology for the Semantic Web.** Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Volume 4, pp. 40112, 2004.

25. DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T., and HARSHMAN, R. **Indexing by latent semantic analysis.** Journal of the American Society for Information Science, Volume 41, Issue 6, pp. 391–407, 1990.

26. DING, L., FININ, T., JOSHI, A., PENG, Y., PAN, R., REDDIVARI, P. **Search on the Semantic Web**. IEEE Computer, Volume 38, Issue 10, pp. 62-269, 2005.

27. DING, L., KOLARI, P., DING, Z., AVANCHA, S., FININ, T., and JOSHI A. **Using Ontologies in the Semantic Web: A Survey.** University of Maryland Baltimore County, In Springer US, 2007.

28. DRUMOND, L., GIRARDI, R. **Uma Análise das Técnicas e Ferramentas para o Desenvolvimento de Aplicações para a Web Semântica.** Ed. Sociedade Brasileira de Computação. REIC - Revista Eletrônica de Iniciação Científica. Março de 2006.

29. DRUMOND, L., GIRARDI, R. A **Survey of Ontology Learning Procedures**. 3rd Workshop on Ontologies and their Applications (WONTO 2008). Salvador, Brasil. 26 de outubro de 2008.

30. DRUMOND, L., GIRARDI, R., and SILVA, F. **A similarity analysis model for Semantic Web information filtering applications.** In Proceedings of the 20th International Conference on Software Engineering and Knowledge Engineering (SEKE 2008), Ed. Knowledge Systems Institute Graduate School, pp. 638-642. Redwood City, California, USA, 2008.

31. GENNARI, J., MUSEN, M. A., FERGERSON, R. W. et al.. **The Evolution of Protégé: An Environment for Knowledge-Based Systems Development.** Technical Report SMI-2002-0943. 2002.

32. GIRARDI, R. **Classification and Retrieval of Software through their Descriptions in Natural Language.** Ed. Imprimerie de l´Université de Geneve, Geneva, Switzerland, 1995.

33. GIRARDI, R., IBRAHIN, B. **Using English to Retrieve Software.** The Journal of Systems and Software, volume 30, nº 3, pp. 249–270, 1995.

34. GÓMEZ-PÉREZ, A. and BENJAMINS, V.R. **Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods.** In Proceedings of the International Joint Conference on Artificial Intelligence(IJCAI'99), Workshop on Ontologies and Problem-Solving Methods: Lesson learned and Future Trends (KRR5), V.R. Benjamins, et al., Editors, CEUR Publications, Amsterdam, volume 18, pp. 1.1-1.15. Stockolm, Sweden, 1999.

35. GRUBER, T. R. **Toward Principles for the Design of Ontologies used for Knowledge Sharing.** International Journal of Human-Computer Studies, nº 43, pp. 907-928, 1995.

36. GUARINO, N. **Formal ontology in information systems.** In: Proceedings of the 1st international conference on formal ontologies in information systems FOIS 1998, IOS Press, Italy, pp. 3–15, 1998

37. GUHA, R., MCCOOL, R. and MILLER, E. **Semantic Search.** In WWW '03: Proc. of the 12th int. conf. on World Wide Web, 2003.

38. HANDSCHUH, S. and STAAB, S. **Cream: Creating metadata for the Semantic Web.** Computer Networks: The International Journal of Computer and Telecommunications Networking, Volume 42, nº 5, pp. 579–598, 2003.

39. HATCHER, E. and GOSPODNETIC, O. **Lucene in Action.** Manning Publications Co, 2005.

40. HEFLIN, J. and HENDLER, J. **Searching the Web with SHOE.** In Artificial Intelligence for Web Search. Papers from the AAAI Workshop. WS-00-01, 2000.

41. HENDLER, J., STOFFEL, K., TAYLOR, M., RAGER, D., KETTLER, B. **PARKA-DB: A Scalable Knowledge Representation System**. 1996, Disponível em <http://www.cs.umd.edu/projects/plus/Parka/parka-db.html>. Acessado em 25/06/2008.

42. HORROCKS, I., PATEL-SCHNEIDER P. F., MCGUINNESS, D. L., and WELTY, C. A. **OWL: a Description Logic Based Ontology Language for the Semantic Web,** chapter Deborah L. McGuinness and Peter F. Patel-Schneider. From Description Logic Provers to Knowledge Representation Systems. In The Description Logic Handbook: Theory, Implementation and Applications, pp. 458–486. Cambridge University Press, 2nd edition, 2007.

43. JANOWICZ, K. **Sim-DL: Towards a semantic similarity measurement theory for the description logic ALCNR in geographic information retrieval.** In SeBGIS 2006, OTMWorkshops 2006, ser. Lecture Notes in Computer Science, R. Meersman, Z. Tari, P. Herrero, and e. al., Eds, Springer, volume 4278, pp. 1681–1692, 2006.

44. JANOWICZ, K., KELER, C., SCHWARZ, M., WILKES, M., PANOV, I., ESPETER, M., and BAEUMER, B. **Algorithm, implementation and application of the Sim-DL similarity server.** in Second International Conference on GeoSpatial Semantics (GeoS 2007). ser. Lecture Notes in Computer Science, Springer, 2007.

45. KIRYAKOV, A., POPOV, B., TERZIEV, I., MANOV, D., and OGNYANOFF, D. **Semantic annotation, indexing, and retrieval.** Journal of Web Semantics: Science, Services and Agents on the World Wide Web, volume 2, pp. 49–79, 2004.

46. LEACOCK, C. and CHODOROW, M. **Combining local context and wordnet similarity for word sense identification.** In Fellbaum, E. C., editor, WordNet: A Lexical Reference System and its Application, pp. 265–283, Cambridge, MA: MIT, 1998.

47. LEI, Y., UREN, V., and MOTTA, E. **SemSearch: A search engine for the Semantic Web.** In Springer, editor, Proceedings of EKAW 2006 Managing Knowledge in a World of Networks, volume 4248, pp. 238–245, Heidelberg, 2006.

48. LIN, D. **An information–theoretic definition of similarity.** In Proceedings of the International Conference on Machine Learning (ICML) Morgan Kaufman, San Francisco, pp. 296–304, 1998.

49. LOPEZ, V.; PASIN, M.; MOTTA, E. **Aqualog: An ontology-portable question answering system for the Semantic Web.** in: A. Gomez-Perez, J. Euzenat (eds.), ESWC, volume 3532 of Lecture Notes in Computer Science, Springer, 2005.

50. MACHADO, H. B. **Curso de Direito Tributário.** Malheiros Editora Ltda, 2004.

51. MAEDCHE, A. **Ontology Learning for the Semantic Web.** Kluwer Academic Publishing, 2002.

52. MAEDCHE, A. and STAAB, S. **Measuring similarity between ontologies.** In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, pp. 251–263, 2002.

53. MANNING, C. D., RAGHAVAN, P., SCHÜTZE, H. **Introduction to Information Retrieval.** Cambridge University Press, 2008.

54. MANOLA, F., MILLER, E. **RDF Primer.** W3C Recommendation 10 February 2004. Disponível em <http://www.w3.org/TR/rdf-primer/>. Acessado em 25/06/2008.

55. MILLER, G. A., FELLBAUM C., TENGI, R., WAKEFIELD, P., LANGONE, H., HASKELL, B. R. **WordNet - a lexical database for the English language**. 2006. Disponível em <http://wordnet.princeton.edu/>. Acessado em 25/06/2008.

56. RESNICK, P. **Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.** Journal of Artificial Intelligence Research, pp. 95–130, 1999.

57. ROCHA, C., SCHWABE, D. and ARAGAO, M. **A hybrid approach for searching in the Semantic Web.** In Proceedings of the 13th international conference on World Wide Web, pp. 374 – 383, 2004.

58. RODRIGUEZ, A. and EGENHOFER, M. **Comparing geospatial entity classes: an asymmetric and context dependent similarity measure.** International Journal of Geographical Information Science, volume 18, nº 3, pp. 229–256, 2004.

59. SALTON, G. and BUCKLEY C. **Term-Weighting Approaches in Automatic Text Retrieval.** Inf. Process. Manage, volume 24, nº 5, pp. 513-523, Aug, 1988.

60. SALTON G., WONG, A., YANG, C. S. **A vector space model for automatic indexing.** Communications of the ACM, volume 18, nº 11, pp. 613-620, Nov, 1975.

61. SCHEIR, P., PAMMER, V. and LINDSTAEDT, S.N. **Information Retrieval on the Semantic Web - Does it exist.** In Proceedings of the Lernen-Wissen-Adaption, pp. 252-257, 2007.

62. SHADBOLT, N., HALL, W. and BERNERS-LEE, T. **The Semantic Web revisited.** Intelligent Systems, volume 21, pp. 96-101, 2006.

63. SHAH, U., FININ, T. and JOSHI, A. **Information retrieval on the Semantic Web.** In CIKM ’02: Proc. Of the 11th int. conf. on Information and knowledge management, pp. 461 - 468, 2002.

64. SILVA, F. A. S., GIRARDI R. e DRUMOND, L. **An Information Retrieval Model for the Semantic Web.** Proceedings of the 6th International

Conference on Information Technology : New Generations, Ed. IEEE. Las Vegas, Nevada, USA, 2009. *Por aparecer.*

65. SINGHAL, **A. Modern Information Retrieval: A Brief Overview.** Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2001.

66. SOUZA, R. R., ALVARENGA, L. **A Web Semântica e suas contribuições para a ciência da informação**. Ci. Inf., Brasília, volume 33, nº 1, pp. 132-141, jan/abr 2004.

67. SUCHANEK, F. M., IFRIM, G. and WEIKUM, G. **LEILA: Learning to Extract Information by Linguistic Analysis.** In Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, pp. 18-25, Sydney, Australia, July, 2006.

68. TANEV, H. and MAGNINI, B. **Weakly Supervised Approaches for Ontology Population**. In Proceedings of EACL, pp. 3-7, 2006.

69. TANNENBAUM, A. **Metadata Solutions: Using Metamodels, Repositories, XML, and Enterprise Portals to Generate Information on Demand.** Addison Wesley, 2001.

70. VALLET, D., FERNÁNDEZ, M. and CASTELLS, P. **An ontology–based information retrieval model.** Proc. Second European Semantic Web Conf. (ESWC ’05), pp. 455-470, 2005.

71. WEI, W., BARNAGHI, P. M. and BARGIELA A. **Search with Meanings: An Overview of Semantic Search Systems.** International journal of Communications of SIWN, Volume 3, pp. 76-82, 2008.

ANEXO I – VISÃO GERAL DO SISTEMA PARA A INSTANCIAÇÃO DE INSTRUMENTOS NORMATIVOS JURÍDICO-TRIBUTÁRIOS

A necessidade de uma fonte de informação para o estudo de caso apresentado nesse trabalho exigiu a criação de uma ferramenta para facilitar a instanciação de instrumentos normativos jurídico-tributários na ontologia ONTOTRIB. Mais especificamente, a ferramenta em questão foi criada com o propósito de evitar a instanciação manual dos instrumentos normativos, uma tarefa cansativa devido a extensão de alguns instrumentos normativos, demorada, exigindo a identificação de instâncias e a especificação das suas relações na base de conhecimento e propensa a erros.

A ferramenta construída se divide em três rotinas complementares: a primeira é um interpretador (classe *parseLei) (*Figura 30) que faz a leitura dos instrumentos normativos em formato texto e identifica os atributos e dispositivos que fazem parte de sua estrutura.

Figura 30 Conjunto de classes da ferramenta construída

A estrutura dos dispositivos utilizada é a descrita pela Lei Complementar 95 que dispõe sobre a elaboração e redação das leis e também descrita através da

ontologia ONTOJURIS (Figura 20). A tabela 18 detalhada a composição dos instrumentos normativos.

Tabela 18 Composição dos Elementos de um Instrumento Normativo

Elemento	Composição
Alínea	Item
Artigo	Alínea, inciso, parágrafo
Inciso	Alínea
Parágrafo	Alínea, inciso
Item	Não pode ser composto por nenhum outro elemento
Livro	Artigo, titulo
Capitulo	Artigo, seção
Parte	Artigo, parte
Seção	Artigo, subseção
Subseção	Artigo
Titulo	Capitulo, artigo, seção

Os dispositivos identificados pelo interpretador alimentam a segunda rotina que é uma interface com o usuário para a criação das instâncias de cada instrumento normativo-tributário (Figura 31). Através dessa interface o usuário pode interpretar os dispositivos dos instrumentos normativos, avaliar instâncias já existentes na base de conhecimento e especificar os conceitos da ONTOTRIB referentes aos casos semânticos $Instrumento\ Normativo\ Tributário$, $Tributo$, $Conceito\ Tributário$ e $Aplicação\ Tributária$ a serem instanciados. Por exemplo, se a ferramenta exibir o instrumento normativo nº 8.216, artigo 2º, parágrafo 1º, item 1, para o usuário, dispositivo que menciona que o IPVA deve ser pago quando ocorrer o primeiro licenciamento do veículo, o usuário deve selecionar na lista de conceitos que o dispositivo se refere ao momento do fato gerador deste tributo. Dessa forma, será informado à ferramenta que uma instância do conceito "*Momento de Ocorrência do Fato Gerador*" deve ser criada. Da mesma forma, a interface lista todos os tributos existentes na legislação brasileira e as possíveis aplicações tributárias. Por exemplo, a lei nº 8.216 dispõe sobre o IPVA no estado do Paraná, e a ferramenta criará uma instância deste tributo, específica para este estado, se o usuário selecionar estes valores nas listas de "Aplicação Tributária" e "Tributo" e clicar no botão de ícone "+". A última rotina é responsável por incluir as instâncias definidas pelo usuário na base de conhecimento criada na ferramenta Protégé.

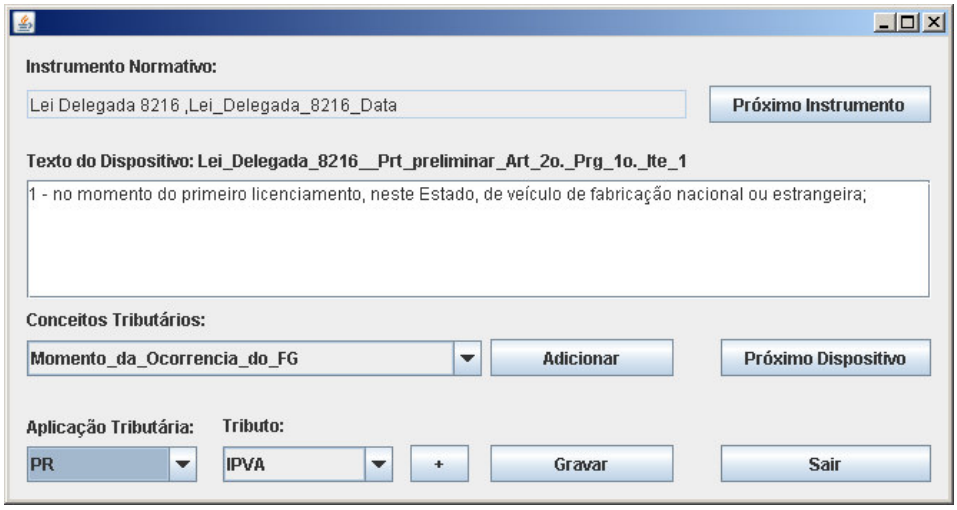

Figura 31 Tela de interface da ferramenta construída

ANEXO II – ARTIGOS ACEITOS E PUBLICADOS BASEADOS NESTE TRABALHO

- SILVA, F. A. S., Girardi R. e Drumond, L. **An Information Retrieval Model for the Semantic Web.** Proceedings of the 6th International Conference on Information Technology : New Generations, Ed. IEEE. Las Vegas, Nevada, USA, 2009.

- SILVA, F. A. S., Girardi R. e Drumond, L. **A Knowledge-Based Retrieval Model.** In Proceedings of the 21th International Conference on Software Engineering and Knowledge Engineering (SEKE 2009). Hyatt Harborside at Logan Int'l Airport, Boston, USA, 2009. (*Por aparecer*)