WANDERSON RIGO

SEMÂNTICA E VISUALIZAÇÃO PARA ANOTAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO

FLORIANÓPOLIS

2011

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

SEMÂNTICA E VISUALIZAÇÃO PARA ANOTAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO

Dissertação submetida ao Programa de Pós-graduação em Ciências da Computação da Universidade Federal de Santa Catarina para a obtenção do grau de Mestre em Ciências da Computação

WANDERSON RIGO

Florianópolis, Outubro de 2011

SEMÂNTICA E VISUALIZAÇÃO PARA ANOTAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO

WANDERSON RIGO

Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Ciências da Computação, e aprovada em sua forma final pelo Programa de Pós Graduação em Ciências da Computação da Universidade Federal de Santa Catarina.

_____

Prof. Renato Fileto, Dr
Orientador

_____

Prof. Mario Antonio Ribeiro Dantas, Dr
Coordenador do Programa de Pós-Graduação em Ciências da Computação

Banca Examinadora:

_____

Renato Fileto, Dr.
Presidente

_____

Andre Santanchè, Dr.

_____

Christiane Gresse von Wangenheim, Dra.

_____

Ricardo José Rabelo, Dr.

_____

Roberto Pacheco, Dr.

"Vocês homens são muito melhores do que a cultura moderna os faça acreditar, vocês são muito melhores do quanto vocês mesmos acreditam ser. Então não tenham medo de ser o que vocês são: criaturas divinas."

Papa João Paulo II.

DEDICATÓRIA

Aos meus pais, Volmir e Marilene, que mesmo carentes de estudo, porém abundantes em valores e saber, me incentivaram a trilhar este caminho. As minhas irmãs Márcia e Morgana, pelas primeiras lições escolares e pelos exemplos de determinação e luta. Ao meu irmão gêmeo Walmir, meu companheiro de viagem, por ter me acompanhado desde o ventre materno até esta formação.

AGRADECIMENTOS

A Deus pela oportunidade de estar aqui desempenhando um papel que mostra que estou no caminho certo e que abre portas para que eu possa ajudar outros a se aperfeiçoarem.

A minha Família, pelo zelo, carinho e atenção. Mesmo longe se fazem presentes nos valores que me passaram e que me acompanham sempre.

Aos Amigos, que mesmo distante me recebem sempre com muito carinho...a distância só faz transparecer mais o valor de cada um e o quão são importante para minha vida.

Aos que souberam entender minha ausência quando este trabalho exigiu minha dedicação exclusiva.

Ao meu orientador, Professor Dr. Renato Fileto, pelo rigor e formalismo exigido, pela dedicação e inúmeras horas dispensadas, pelos conselhos, pelos exemplos construídos e aperfeiçoados ao "pegar junto" e por colocar a linha de chegada sempre mais a frente, estimulando a superação dos limites.

Ao Vilmar César Pereira Júnior, pela oportunidade de aprendizado proporcionado ao ser seu co-orientador. Ainda pela sua ajuda, dedicação e comprometimento, além dos inúmeros momentos de descontração.

Aos membros do Grupo de Tecnologia da Informação do projeto UnA-SUS da Universidade Federal de Santa Catarina, pelo apoio e troca de idéias.

Ao Grupo Specto, pela disciplina e experiência acumulada durante anos de trabalho, insumo essencial para a realização de etapas deste trabalho. Também pelo apoio e oportunidade de desenvolvimento interpessoal e técnico.

A CAPES e ao Ministério da Saúde (programa UnA-SUS) por amparem esta pesquisa e a todos os que contribuíram para o desenvolvimento e a avaliação do sistema descrito neste trabalho.

Aos desenvolvedores de *software* livre, que disponibilizam gratuitamente ferramentas úteis e funcionais.

Ao meu computador, meu "braço direito" que executou centenas de pesados testes e se portou muito bem durante estes quase 3 anos.

A minha gata Mima, por me fazer companhia, mesmo que manhosa e sonolenta, e por incansavelmente requisitar minha atenção durante o desenvolvimeto deste trabalho.

Resumo da Dissertação apresentada à UFSC como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências da Computação.

SEMÂNTICA E VISUALIZAÇÃO PARA ANOTAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO

Wanderson Rigo

Outubro / 2011

Orientador: Prof. Renato Fileto, Dr.
Área de Concentração: Sistemas de Recuperação de Informação.
Palavras-chave: Interfaces Gráficas do Usuário (IGU), Visualização de Conhecimento, Anotação de Conteúdo, Recuperação de Informação, Ativação por Espalhamento, Busca Associativa, Busca Semântica.
Número de Páginas: 166

Sistemas de Recuperação de Informação (SRI) tradicionalmente se apóiam na correspondência léxica entre as palavras-chave usadas na consulta formulada pelo usuário e palavras encontradas nos metadados usados para descrever objetos informacionais (artigos, objetos multimídia, etc.) ou no próprio conteúdo desses objetos informacionais (OI). O objetivo é recuperar os objetos que satisfazem as consultas, com rapidez, boa precisão e boa cobertura. Porém tais sistemas são afetados por fenômenos lingüísticos e limitações semânticas. Nossa abordagem usa conhecimento específico de domínios, presente em Vocabulários Controlados (VC) e adaptado ao formalismo de ontologias tanto na anotação dos objetos (i.e., preenchimento de certos campos de metadados) quanto para o processamento das buscas, visando minimizar tais problemas. Para amparar usuários na anotação de OIs nosso trabalho utiliza interfaces gráficas que direcionam a escolha dos valores de metadados a termos de VCs. A recuperação dos OIs na nossa abordagem leva em conta as relações semânticas formalizadas em ontologias e as estabelecidas para anotar os OIs com termos de VCs, as quais formam uma Rede Semântica (RS). Tal estrutura permite expandir semanticamente as buscas a partir dos termos usados como palavras-chave na especificação das consultas valendo-se da técnica de *Spreading Activation (SA)*. Testes de usabilidade realizados com usuários em um estudo de caso na área da saúde permitiram identificar as interfaces baseadas em conhecimento por eles consideradas mais apropriadas para anotação. Testes com diferentes configurações

de parâmetros e testes de carga com o SA apontaram valores de parâmetros adequados para otimizar o SA e mostraram a sua viabilidade em termos de desempenho, com um VC contendo dezenas de milhares de termos e coleções com tamanho crescente de objetos anotados.

Abstract of Dissertation presented to UFSC as a partial fulfillment of the requirements for the degree of Master in Computer Science.

SEMANTIC AND VISUALIZATION FOR ANNOTATION AND RETRIEVAL OF INFORMATION

Wanderson Rigo

Outubro / 2011

Advisor: Prof. Renato Fileto, Dr.
Area of Concentration: Retrievel Information System.
Keywords: Graphical User Interfaces (GUI), Knowledge Visualization, Content Tagging, Information Retrieval, Spreading Activation, Associative Search, Semantic Search.
Number of pages: 166

Information Retrieval Systems (IRS) typically rely on lexical matching between the keywords used in the query formulated by the user and words found in the metadata used to describe the information objects (articles, multimedia objects, etc.) or on the content of these objects (IO). The IRS goal is retrieve objects that satisfy queries with speed, good precision and good coverage. However such systems are affected by linguistic phenomena and semantic limitations. Our approach uses domain specific knowledge, present in controlled vocabulary (CV) and adapted to ontologies formalism as in the annotation of objects (ie, filling of certain metadata fields) as in the search processing in order to minimize such problems. To support users while annotate IOs this work uses graphical interfaces to guide the users when choosing metadata values from terms of CV. As for the retrieval of IOs our approach takes into account the semantics relationships formalized in ontologies and ones established between the terms used to annotate the OIs with CVs terms, which form a Semantic Network (SN). This structure allows semantically expanding the search from the terms used as keywords in the query specification using the Spreading Activation technique. Usability tests performed with users on a case study in health area enabled the identification of those knowledge-based user interface considered most appropriate for annotation. Tests with different parameter settings and load tests conducted with the SA indicated parameter values appropriate to optimize the SA and showed its feasibility in terms of performance, with a CV containing dozens of thousands of terms and collections with increasing size of annotated objects.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

API	Application Programming Interface.
AS	Anotações Semânticas.
AVEA	Ambiente Virtual de Ensino e Aprendizagem.
BIREME	Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde.
BVS	Biblioteca Virtual em Saúde.
DC	Dublin Core.
DeCS	Descritores em Ciências da Saúde.
DG	Design Gráfico.
DH	Diagramas Hierárquicos.
DI	Design Instrucional.
DRI	Digital Repository Interface.
e-PMG	Padrão de Metadados do Governo Eletrônico.
EaD	Ensino a Distância.
HP	Hewlett-Packard.
IHC	Interfaces Humano-Computador.
IHMC	Florida Institute for Human and Machine Cognition.
LMS	Learning Management System.
LOM	Learning Objects Metadata.
MC	Mapas Conceituais.
MeSH	Medical Subject Headings.
MH	Mapas Hiperbólicos.
MIT	Massachusetts Institute of Technology.
NLM	United States National Library of Medicine.
OA	Objeto de Aprendizagem.
OWL	Web Ontology Language.
RA	Rede Associativa.
RDF	Resource Description Framework.
RI	Recuperação de Informação.
RS	Rede Semântica.
SA	Spreading Activation.
SAN	Spreading Activation Network.
SCORM	Sharable Content Object Reference Model.
SGA	Sistema de Gestão da Aprendizagem.
SRI	Sistemas de Recuperação de Informação.
SUS	Sistema Único de Saúde.
UML	Unified Modeling Language.

UnA-SUS	Universidade Aberta do SUS.
URI	Uniform Resource Identifier.
VC	Vocabulário Controlado.

1 INTRODUÇÃO

Este capítulo contextualiza o trabalho e descreve o problema tratado, relata quais são os objetivos, as estratégias, as justificativas, as limitações, a metodologia empregada e os resultados esperados ao término deste trabalho.

O crescimento do número de publicações em formato digital motivou o desenvolvimento de sistemas computacionais que visam facilitar o acesso do usuário à informação. Neste contexto, surgiram os Sistemas de Recuperação de Informação (SRI), os quais provêem um ambiente de alta disponibilidade para armazenar grandes quantidades de objetos de informação (artigos, objetos multimídia, etc.) em formato digital e facilidades para permitir a recuperação de objetos com rapidez e precisão (CRESTANI, 1997) (SOUZA, 2006).

Geralmente os objetos informacionais (OI) são descritos através do preenchimento de campos de metadados como título, autor, formato, descrição, assunto, palavra-chave, etc. de forma a fornecer subsídios para que a posterior recuperação de tais OIs tenha êxito. A recuperação de objetos em SRI muitas vezes é realizada por meio da navegação: usuários exploraram exaustivamente o conteúdo de tais sistemas visando satisfazer suas necessidades de informação. Porém, à medida que cresce o volume de objetos disponíveis, esta técnica mostra-se inapropriada (DIAS; CARVALHO, 2007). Outros mecanismos de busca são necessários para se lidar com grandes coleções de objetos e diversas pesquisas visam o desenvolvimento de mecanismos mais eficazes para a descrição e recuperação de OIs (DIAS; CARVALHO, 2007)(SILVA, 2007)(FALOUTSOS; OARD, 1995)

Dentre as técnicas clássicas de especificação de consultas destacam-se seleção de valores de metadados e especificação de palavras-chave. Palavras-chave são amplamente usadas pelos atuais motores de busca da Web e em SRI, embora os usuários podem ter dificuldade para especificar as palavras-chaves que resultem na recuperação da informação desejada, principalmente em sistemas que não consideram a semântica das palavras. Isso acontece porque as palavras utilizadas na descrição dos objetos podem diferir lexicamente daquelas utilizadas (normalmente por outros usuários) na especificação de consultas visando a recuperação de tais objetos.

1.1 Contextualização

O preenchimento de campos de metadados visa identificar e individualizar os objetos digitais (AFONSO, 2010). Campos de metadados como palavra-chave (*keyword*) e assunto, os quais podem inclusive ser multivalorados, são essenciais para suportar a descrição de um OI. Além de texto livre, tais campos pode assumir valores definidos em certas estrutura de conhecimento, como Ontologias e Vocabulários Controlados (VC), caracterizando assim uma **anotação semântica**. Tais campos de metadados são preenchidos durante a etapa de catalogação dos objetos. A **Catalogação** visa a construção de um catálogo que sumariza os objetos descritos levando-se em conta os valores de metadados que os descrevem (LEVY, 1995).

1.2 O problema abordado

Em geral, a necessidade de se encontrar informações ou objetos é motivada por um tópico ou assunto específico (BAEZA-YATES; RIBEIRO-NETO, 1999). Por exemplo, como localizar objetos que versem sobre 'Acidente Vascular Cerebral'? Valores assumidos pelos metadados que descrevem tais objetos ou o próprio conteúdo desses objetos fornecem subsídios para guiar os métodos de recuperação (AFONSO, 2010).

A técnica de recuperação popularizada pelo atuais motores de busca da Web (e.g., Google [1], Bing [2], Altavista [3]) exige que o usuário traduza a sua necessidade de informação em uma consulta (SHAH et al., 2002). Apesar da inerente dificuldade de se formular a consulta, geralmente esse esforço resulta em um conjunto de palavras-chave que resumem a intenção do usuário (BAEZA-YATES; RIBEIRO-NETO, 1999). **Palavras-chave** são expressões lingüísticas, constituídas de uma ou mais palavras que juntas referenciam algum assunto (adaptado de (MENG; CHU, 1999)).

Como a meta perseguida pelos SRIs é recuperar os objetos relevantes à consulta formulada pelo usuário (BAEZA-YATES; RIBEIRO-NETO, 1999), tais sistemas procuram fazer a correspondência entre o conjunto de palavras-chave que resumem a intenção do usuário e os valores de metadados que descrevem os objetos armazenados no SRI. Porém, podem acontecer vários problemas na tentativa de recuperar objetos via palavras-chaves livres:

[1] http://www.google.com
[2] http://www.bing.com
[3] http://www.altavista.com

- O usuário pode não conseguir traduzir sua necessidade de informação em uma consulta.
- As palavras-chave que o usuário usa na consulta não correspondem aos valores dos metadados que anotam os objetos que melhor satisfazem as necessidades deste usuário.
- Uma palavra-chave pode possuir diversos sinônimos. Qual usar na busca? Vai funcionar? Como saber?
- Uma palavra-chave pode denotar diferentes coisas, ou seja, é ambígua (homonímia). Como diferenciar na busca? Que denotação considerar?
- Nenhum objeto associado a palavra-chave existe no sistema. Algum similar pode satisfazer a necessidade de informação?

SRI que operam buscas baseadas em correspondência léxica também são afetados por problemas decorrentes do uso de linguagem natural, notadamente homonímia e sinonímia. A primeira ocorre quando uma dada palavra ou frase tem significados diferentes (SVENONIUS, 1986a). Exemplificando: a um usuário que efetuou uma consulta coma a palavra-chave "Java", visando recuperar objetos referentes à linguagem de programação Java, poderiam ser retornados resultados sobre a Ilha de Java ou o Mar de Java. Esta situação de recuperação é caracterizada por exibir uma precisão pobre.

Já a sinonímia ocorre quando duas palavras compartilham o mesmo significado, ou seja, existe mais de uma palavra para denotar o mesmo referente (SVENONIUS, 1986a). Por exemplo, Florianópolis é conhecida também por Ilha da Magia, Desterro, Ilha de Santa Catarina, Nossa Senhora do Desterro, etc. A um usuário que conhece apenas algumas dessas designações pode ser que não sejam retornados objetos anotados com outras. Esta situação de recuperação é caracterizada por exibir uma revocação pobre.

Além desses problemas, erros (e.g., grafia, interpretação) durante a descrição de um objeto podem inviabilizar a sua recuperação. Esses fatores afetam negativamente a revocação (*recall*) e a precisão (*precision*) (MANGOLD, 2007).

Resumindo, métodos de recuperação baseados apenas em sintaxe são afetados por fenômenos lingüísticos, apresentam limitações e não atendem amplamente às necessidade de informação dos usuários. Tendo em vista estas questões, usuários necessitam de métodos inteligentes e eficientes para acessar objetos e a descrição e a organização da informação são fundamentais para se alcançar este propósito (CRESTANI; RIJSBERGEN, 1993). A iniciativa que

mais tem se destacado é a Web Semântica (BERNERS-LEE; LASSILA, 2001), cujo principal objetivo é a compreensão semântica da informação, tanto na ótica dos humanos como das máquinas. Um esforço importante tem sido empreendido para dotar SRI de características semânticas, visando principalmente descrever, interrelacionar e compreender os objetos armazenados através de metadados e de ontologias (GONçALVES, 2007).

Dado este panorama, que estratégia seria mais adequada para **promover a anotação** de OIs e a posterior **recuperação** de tais OIs de modo que os problemas aqui levantados sejam resolvidos ou minimizados? Este é o desafio perseguido por este trabalho.

1.3 Motivação

Este trabalho visa fomentar a recuperação e reuso de OIs através da catalogação apoiada por Interfaces Humano-Computador (IHC) que facilitem e promovam a anotação dos objetos com termos providos por estruturas de conhecimento como Ontologias e VCs. Como muitos desses objetos demandam esforço e são caros para produzir, seu reuso evita gastos financeiros e atrasos decorrentes de produção replicada. Esta implicação torna-se mais evidente em aplicações como a formação continuada e a atualização de profissionais.

O reuso de OIs na elaboração de cursos, treinamentos e na qualificação de profissionais da área da saúde, por exemplo, permite maior agilidade no atendimento a demandas emergenciais (e.g., pandemia de "influenza A" ou "gripe suína" que assolou o Brasil em 2009) [4].

1.4 Objetivos

O objetivo geral deste trabalho é desenvolver métodos para visualização de conhecimento e assim apoiar a anotação semântica de OIs e a especificação de consultas que visam recuperar tais OIs via buscas associativas atuando sobre representações deste conhecimento e suas associações semânticas com os OIs anotados. Especificamente:

1. Pesquisar, avaliar e selecionar técnicas e ferramentas de visualização de informação em IHCs.
2. Valendo-se das técnicas e ferramentas de visualização avaliadas, propor IHCs para apoiar a anotação de OIs, a navegação em repositórios desses

[4] http://portal.saude.gov.br/portal/arquivos/pdf/influenza_protocolo_atencao_basica_novo2.pdf

objetos e a especificação de consultas usando palavras-chaves de VCs e suas relações semânticas.

3. Desenvolver, implementar e testar artefatos de *software*, independentes de domínio e parametrizáveis para apoiar a anotação de objetos, a especificação e o processamento de consultas, visando a recuperação de informação eficiente e eficaz em grandes repositórios de OIs.

4. Avaliar a aceitação das IHCs para apoiar a anotação de OIs e o desempenho dos métodos de processamento de consultas e recuperação de OIs com diferentes configurações de seus parâmetros e tamanho das coleções de OIs, em um estudo de caso real na área de saúde.

1.5 Estratégia

O uso de interfaces visuais associadas a conhecimento de domínio, formalizado em VCs e ontologias, pode auxiliar os usuários a efetuar a anotação semântica de OIs e a especificação de consultas, com maior agilidade e precisão que a livre enumeração de palavras-chaves. As bases de conhecimento obtidas pela anotação dos objetos usando termos padronizados e semanticamente relacionados em ontologias e VCs constituem redes semânticas, sobre as quais podem ser processadas buscas semânticas, usando técnicas como *Spreading Activation (SA)*(QUILLIAN, 1968) (COLLINS; LOFTUS, 1975). SA permite a expansão semântica controlada das palavras-chaves usadas nas consultas e pode ser executada de forma suficientemente eficiente para viabilizar a sua aplicação prática, desde que sejam efetuados ajustes apropriados de seus parâmetros. A implementação e o teste dos mecanimos de anotação e busca semântica propostos neste trabalho visam validar essas hipóteses em ao menos um domínio de aplicação: anotação e recuperação de objetos de aprendizagem da área de saúde. Testes de usabilidade com usuários serão conduzidos para avaliar as interfaces de anotação baseadas em conhecimento enquanto testes com diversos valores de parêmetros para o SA e de carga irão aferir o desempenho computacional do método de recuperação.

1.6 Justificativas

Um suporte adequado à anotação de OIs evita que tais objetos sejam mal descritos em decorrência de fenômenos que ocorrem nas linguagens naturais (e.g., uso de sinônimos, homônimos, termos genéricos, erros de grafia),

o que poderia dificultar ou até inviabilizar a recuperação dos mesmos (SOUZA et al., 2008).

A fim de prover suporte ágil e eficaz à anotação e recuperação de OIs, este trabalho propõe amparar tais tarefas por meio de interfaces Web que permitem explorar interativamente conhecimento de domínio. Isto é útil quando o usuário não tem domínio da área de conhecimento em questão e precisa descrever um certo OI para fins de catalogação. Também é essencial para usuários que durante a formulação de consultas usam palavras-chave que não são os termos utilizados no momento da catalogação para descrever aquilo que querem encontrar (SILVA, 2007). As interfaces ainda podem ajudar o usuário que não sabe como descrever sua necessidade de informação, já que permitem que um processo exploratório sobre o domínio da área de conhecimento em questão seja executado. Tais interfaces também apóiam a navegação no conteúdo de SRI e estipulação de consultas usando palavras-chaves de um VC.

Buscas semânticas suplantam limitações dos métodos de buscas tradicionais por permitirem que objetos descritos com termos diferentes dos usados na consulta também sejam recuperados. Por exemplo, uma busca semântica via termo "AVC" sobre uma base de conhecimento poderia recuperar objetos descritos por sinônimos (e.g. "Acidente Vascular Cerebral", "Derrame Cerebral", "Apoplexia", "Ictus Cerebral".) ou termos associados a esse termo (e. g. "Isquemia Encefálica", "Transtornos Cerebrovasculares", "Cefaléias Vasculares", "Infarto Encefálico").

Embora diversos trabalhos proponham abordagens para anotação semântica de OIs (ROCHA et al., 2008) (EUZENAT, 2002) e processamento de buscas semânticas com SA para recuperação dos objetos anotados (CRESTANI; LEE, 1999) (NILAS; NILAS; MASAKUL, 2007), testes com usuários e testes de carga devem ser conduzidos para demonstrar a viabilidade de tais abordagens em aplicações práticas reais. Além disso, o uso concomitante de técnicas de visualização e acesso facilitado a conhecimento presente em vocabulários controlados e ontologias ainda não foi suficientemente explorado na anotação de objetos, especificação de consultas, navegação em grandes redes semânticas e gerenciamento do conteúdo de repositórios de informação.

Este trabalho visa contribuir para o preenchimento dessas lacunas, com a aplicação de técnicas para visualização de conhecimento na anotação e formulação de consultas que atuam numa implementação de SA robusto e customizável para processamento de buscas semânticas em diversos domínios, além da avaliação (com usuários e do ponto de vista do consumo

de recursos computacionais) desses artefatos em um estudo de caso envolvendo a anotação e a recuperação de objetos de aprendizagem para a área de saúde, com apoio de especialistas desta área.

1.7 Limitações do trabalho

Nossa abordagem está atrelada ao uso de VCs ou ontologias, os quais podem não estar disponíveis ou serem incompletos quanto à cobertura de um domínio específico. No primeiro caso, pode ser difícil ou inviável colocar o sistema em operação, pois construir um VC é trabalhoso e depende da colaboração de especialista e/ou análise de grandes volumes de informação para extração e formalização do conhecimento. No segundo caso, o usuário pode sentir necessidade de usar termos que não estão contemplados na base de conhecimento disponível, a qual não pode ser arbitrariamente expandida sem a tutela de um especialista de domínio, sob pena de comprometer sua confiabilidade. Além disso, não há como definir, *a priori*, a importância de cada tipo de relação semântica que ocorre na base de conhecimento para a sua utilização na recuperação de informação em uma aplicação específica.

Consultas que exploram outros valores de campos de metadados como autor, data, título não são suportadas por nosso método de recuperação semântica. Ele considera nas buscas semânticas somente os valores do campo **assunto** ou **palavra-chave** *(keyword)*, i.e., cujos domínios de valores possíveis e relações semânticas entre esses valores são descritos em um VC ou ontologia.

Como nossas propostas são aplicadas e validadas em um estudo de caso na área da saúde, a validade dos resultados dos nossos experimentos, a princípio, se restringe a tal estudo de caso, embora possa ser testada e verificada em outras situações. Outro fator que pode impactar no trabalho é a quantidade e a qualidade das anotações realizadas pelos usuários do SRI deste estudo de caso. Sabe-se que a anotação adequada dos objetos durante a catalogação depende do conhecimento, do interesse e da boa vontade dos usuários. Eleger termos apropriados para anotar um dado objeto consome tempo. Porém nem todos os usuários que catalogam objetos se dispõem a isso. Além disso, a quantidade e os temas limitados dos objetos catalogados no SRI pode dificultar ou até mesmo inviabilizar a execução de testes ou comprometer seus resultados.

1.8 Metodologia

Este trabalho envolve o projeto *(design)*, a implementação e a avaliação de soluções de tecnologia da informação para solucionar problemas na área de ciências naturais. Assim, uma metodologia baseada na ciência de projeto é recomendada (MARCH; SMITH, 1995) e seguimos a metodologia DSRM *(Design Science Research Methodology)* (PEFFERS et al., 2007). Tal metodologia fornece um modelo de processo de pesquisa com seis passos:

1. **Identificação do problema e motivação** - foi realizada mediante revisão bibliográfica nas áreas de representação e manipulação de conhecimento, anotação semântica, busca semântica e visualização de informação, além da análise do contexto geral da aplicação. Especificamente:
    - Revisão bibliográfica das técnicas de Recuperação de Informação.
    - Estudo de padrões da Web Semântica.
    - Estudo de modelos em rede para representação de conhecimento.
    - Estudo, captura, tratamento e adequação de um VC.
2. **Definição dos objetivos da solução** - baseou-se em análise detalhada da aplicação, usando um estudo de caso, com validação através de entrevistas e *workshops* com especialistas do domínio, donde se estabeleceu como objetivo principal:
    - Desenvolvimento de artefatos de *software* para amparar a anotação e a recuperação de OIs;
3. **Projeto e desenvolvimento da solução** - utilizou prototipação para apoiar um processo em espíral com refinamentos sucessivos da proposta, donde desenvolveu-se as seguintes atividades:
    - Para a Catalogação:
        - Estudo de técnicas para visualização de conhecimento.
        - Estudo de ferramentas e tecnologias para concepção de interfaces baseadas em conhecimento para anotação de objetos na catalogação.
        - Desenvolvimento de ferramentas que, a partir do VC, geram os dados que alimentam as interfaces baseadas em conhecimento para anotação de objetos na catalogação.

  - Desenvolvimento e testes das interfaces baseadas em conhecimento para anotação de objetos na catalogação.
- Para a Recuperação:
  - Desenvolvimento de ferramentas para geração de uma Rede Semântica a partir do VC capturado.
  - Estudo e implementação de métodos para detecção de similaridade entre termos do VC.
  - Estudo de métodos para recuperação associativa de informações.
  - Adaptação do algoritmo de busca (SA).
  - Desenvolvimento de um módulo de busca semântica que vale-se do algoritmo de SA.
  - Desenvolvimento de um protótipo para testes com buscas semântica que utiliza o módulo de busca semântica desenvolvido.
  - Desenvolvimento de um módulo Web para busca semântica (interface de busca + algoritmo SA + Rede Semântica).

4. **Demonstração da solução** - utilizou módulos de *software*, resultantes do passo anterior, para apoiar a anotação semântica e realizar buscas semânticas de objetos de aprendizagem, em um estudo de caso, utilizando um VC amplamente difundido na área de saúde e adaptado para esta finalidade, donde realizou-se:

   - Acoplamento das interfaces baseadas em conhecimento para anotação de objetos no *workflow* de catalogação de um SRI.
   - Acoplamento do módulo Web para busca semântica a um SRI.

5. **Avaliação da solução** - a avaliação das funcionalidades de anotação semântica usou o mesmo estudo de caso e contou com a participação de usuários (especialistas da área de saúde e outros) para a realização de testes empíricos, enquanto a avaliação das buscas com SA se limitou a testes de desempenho, com alguns dados reais e outros sintéticos, devido à indisponibilidade de grandes volumes de objetos anotados até então e dificuldades para conseguir apoio dos usuários. Especificamente trabalhou-se na:

   - Definição de testes de usabilidade com usuários para avaliar as interfaces baseadas em conhecimento.

- Aplicação de testes de usabilidade com usuários para avaliar as interfaces baseadas em conhecimento.
- Melhoria das interfaces baseadas em conhecimento em decorrência dos testes de usabilidade aplicados.
- Elaboração e execução de testes com o protótipo de buscas semântica segundo variações nos seus parâmetros de configuração.
- Proposta de configuração de valores para os parâmetros do SA.
- Elaboração e execução de testes de desempenho e escalabilidade com o protótipo de busca semântica.
- Proposta de melhorias no módulo de busca semântica em decorrência dos testes de desempenho.

6. **Comunicação dos resultados obtidos** - tem sido feita através de trocas de informações com usuários e os especialistas de domínio que contrataram o projeto de pesquisa e desenvolvimento no qual este trabalho se insere, além da submissão e publicação de artigos e da realização de apresentações orais, com ampla abertura para discussão dos resultados obtidos, possíveis melhoramentos e trabalhos futuros.

1.9 Resultados esperados

Ao término deste trabalho espera-se:

- A proposição de interfaces Web baseadas em conhecimento para amparar a anotação ágil e precisa de OIs assim como a especificação de consultas, economizando tempo do usuário.
- O desenvolvimento um módulo para recuperação de OIs que opere buscas associativas explorando conhecimento de domínio.
- A proposição de configuração de parâmetros para orquestrar a execução das buscas associativas.
- A avaliação do desempenho computacional do módulo de recuperação de OIs frente ao crescimento do número de OIs.

1.10 Estrutura da dissertação

Este trabalho está estruturado da seguinte forma:

- **Capítulo 01 - Introdução**: apresenta uma visão geral sobre SRI, contextualizando e expondo os problemas que os afetam. Ainda aborda os objetivos a serem perseguidos por este trabalho, suas limitações, as justificativas e as motivações que o guiam. Além disso, baseado na hipótese de trabalho, é definida a metodologia utilizada para a concepção deste trabalho.

- **Capítulo 02 - Catalogação de Objetos de Informação**: embasa o trabalho com os arcabouços a cerca do processo de catalogação, onde a descrição de objetos informacionais se dá através da associação de metadados a tais objetos via anotações semânticas.

- **Capítulo 03 - Recuperação de Objetos de Informação**: revisa estruturas para representação de conhecimento e técnicas de processamento para operar recuperação associativa de informação sobre estas estruturas.

- **Capítulo 04 - Visualização de Conhecimento**: apresenta uma resenha das técnicas e ferramentas para a visualização de conhecimento. Ainda faz uma análise comparativo entre algumas ferramentas e aponta as duas mais promissoras para a visualização de conhecimento.

- **Capítulo 05 - Visualizações de Conhecimento na Anotação**: discorre sobre nossa proposta de uso de visualizações do conhecimento na anotação de objetos de informação.

- **Capítulo 06 - Recuperação Semântica**: aborda nossa proposta de Recuperação Semântica de objetos de informação via buscas associativas com a técnica de SA operando sobre estruturas de conhecimento e anotações.

- **Capítulo 07 - Estudo de Caso**: discorre sobre o contexto onde as idéias e implementações apresentadas neste trabalho são aplicadas, destacando a iniciativa de se fomentar o ensino a distância na área da saúde.

- **Capítulo 08 - Testes e Avaliação dos Testes**: descreve os objetivos, os dados, as tarefas e o modo como os testes realizados junto a usuários

do estudo de caso e testes de desempenho procuram testar e validar as propostas apresentadas neste trabalho. Também avalia os resultados obtidos com os testes realizados junto aos usuários do estudo de caso e dos testes de desempenho, aferindo o potencial de nossa proposta.

- **Capítulo 09 - Trabalhos Relacionados**: apresenta um apanhado dos trabalhos afins ao tema de anotação e recuperação semântica de objetos, bem como pesquisas na área de visualização de informação que serviram de inspiração para as idéias aqui abordadas.

- **Capítulo 10 - Conclusão e Trabalhos Futuros**: apresenta as conclusões, discussões, contribuições e direções acerca de trabalhos futuros envolvendo a abordagem apresentada.

- **Anexos**: reúnem recursos suplementares (certificado, questionário, fichas, etc.) que figuram neste trabalho e documentam e detalham alguns aspectos técnicos das implementações.

2 CATALOGAÇÃO DE OBJETOS DE INFORMAÇÃO

Este capítulo discorre sobre a catalogação de OIs, uso e padrões de metadados, VCs, ontologias, anotações semânticas e linguagem para descrição de recursos.

De acordo com (AFONSO, 2010), "catalogar um objeto significa descrevê-lo por meio de seus diferentes aspectos e características. O objetivo da catalogação é fornecer uma representação do objeto, permitindo identificá-lo, localizá-lo e representá-lo". No contexto deste trabalho, considera-se Objeto de Informação:

Definição 1 ***- Objeto de Informação (OI):***

Um objeto de informação oi_i é um texto, imagem, som, vídeo ou conteúdo multimídia em formato digital.

Segundo (MATHES, 2004), a criação de metadados para catalogação de objetos geralmente tem sido abordada de duas formas: criação por profissionais da informação e criação por autores. Embora a catalogação de objetos remeta a um processo manual, muitas vezes executado por especialistas em catalogação, existem outras formas de agregar metadados à recursos para descrevê-los. Uma delas é a a catalogação automática realizada pelo Citeseer [1], por exemplo. Nesta abordagem, algoritmos de extração operam sobre documentos estruturados (e.g. artigos, teses) e elegem termos relevantes que estão presentes em determinados trechos do documento (e.g. título, resumo, autores) ou termos que ocorrem com mais freqüência no texto. É útil salientar que um **termo** é uma "designação de um conceito ou instância definida através de uma palavra-chave simples como "doença" ou composta como "acidente cerebral vascular" (GOMES; MOTTA; CAMPOS, 2009). Então o conjunto desses termos elegidos pelo algoritmo passa a representar o documento. Posteriormente o documento é descrito e armazenado, podendo ser recuperado mediante buscas dirigidas por tais termos. Note que a descrição sumariza em um conjunto de termos o conteúdo de todo o texto (MAXIMINO; MARTINS, 2004).

Outra forma de adicionar metadados a objetos de informação é a técnica de Folksonomia (WAL, 2007a), na qual os usuários descrevem livremente e com seu próprio vocabulário os objetos, como ocorre na catalogação

[1] http://citeseerx.ist.psu.edu

de vídeos no YouTube [2]. Neste contexto surgiu recentemente o fenômeno da Web 2.0 (Web social) (O'REILLY, 2005), onde além do compartilhamento, troca de informações e colaboração, usuários associam livremente termos a objetos para descrevê-los.

Diante do exposto até aqui, é notório que conjuntos de metadados são indispensáveis para a descrição de objetos durante a catalogação.

2.1 Metadados

De acordo com (AFONSO, 2010), metadados são "um conjunto de palavras ou sentenças (elementos) que resumem e descrevem o conteúdo de um recurso digital.Os metadados visam facilitar a gestão e o compartilhamento da informação e representam informações como título, autor, descrição, localização, tipo, formato, entre outras, permitindo um número maior de campos para pesquisas".

É fundamental definir campos de metadados adequados para descrever objetos informacionais. Diversos padrões definem campos de metadados genéricos (e.g., *Dublin Core* (HILLMANN, 2005)) ou para domínios específicos (e.g., *Learning Objects Metadata* (HODGINS, 2002)). Inclusive, em 2009 foi lançado um padrão de metadados para *e-government* totalmente brasileiro, o Padrão de Metadados do Governo Eletrônico (e-PMG) (YAMAOKA, 2009).

Esta pluralidade de tipos de metadados pode comprometer a interoperabilidade entre SRIs, impossibilitando assim a combinação e reuso de OIs que estejam armazenados em SRIs que adotam diferentes padrões de metadados para descrever seus objetos. Em vista disso, um conjunto básico de metadados serve de referência para os processos de compatibilização de metadados entre sistemas: o padrão *Dublin Core (DC)* (HILLMANN, 2005).

2.2 Ontologias

Os termos do universo de discurso de um determinada área referenciam conceitos que podem estar formalizados em ontologias de domínio. De acordo com (GRUBER, 1993), ontologia é "uma especificação formal e explícita de uma conceitualização compartilhada". Ontologias representam conhecimentos e conceitos de um domínio particular do mundo real. Elas capturam o conhecimento de um domínio de forma genérica e provêem um

[2] http://www.youtube.com

entendimento comum deste domínio. Então tal conhecimento pode ser reusado e compartilhado entre aplicações e grupos.

Ontologias contém termos e relacionamentos entre estes termos. Termos são freqüentemente chamados de classes ou conceitos. Segundo (GOMES; MOTTA; CAMPOS, 2009), **conceito** "é uma unidade de conhecimento constituído por características que refletem as propriedades significativas relevantes atribuídas a um objeto ou a uma classe de objetos, sendo expresso comumente por signos lingüísticos. Assumindo que existe uma relação de identidade entre o conjunto de conceitos e um conjunto de termos, pode-se usar estas expressões de modo intercambiável, visto que o termo denota um conceito e é sua forma física visível e manipulável" (adaptado de (GOMES; MOTTA; CAMPOS, 2009)).

O relacionamento entre as classes de uma ontologia pode ser expresso por uma estrutura hierárquica: super-classes representam conceitos de alto nível e sub-classes representam conceitos especializados. Estes conceitos especializados herdam todos os atributos e características que os conceitos de alto nível possuem (YU, 2007). Além do relacionamento entre classes (super-classes e sub-classes), há um outro nível de relacionamento que é expresso através de propriedades. Elas descrevem várias características e atributos dos conceitos e também podem ser usadas para associar diferentes classes (YU, 2007).

Ontologias podem ser expressas pela *Web Ontology Language (OWL)*[3], que é uma extensão do esquema *Resource Description Framework (RDF)* (que será visto em detalhes na seção 2.5.) e é recomendada pela W3C (DEAN; SCHREIBER, 2004). OWL permite a representação de relacionamentos mais complexos entre classes, além de restrições mais precisas na especificação de classes e propriedades.

No espectro da definição de ontologias cabem vários tipos de representações de conhecimento tais como glossários, vocabulários controlados, taxonomias e thesauros, que são chamados de ontologias lingüísticas. Uma espécie de ontologia menos precisa são os vocabulários controlados (BORGO, 2004).

2.3 Vocabulários controlados

Tão importante quanto definir os campos de metadados a utilizar na descrição de OIs em um SRI é definir os possíveis valores que tais campos podem assumir, além de garantir o preenchimento correto destes campos.

[3] http://www.w3.org/TR/owl-ref

Adotar uma estratégia adequada para preenchimento de campos de metadados pode evitar problemas decorrentes das limitações das linguagens naturais (e.g., uso de sinônimos, homônimos, termos genéricos, erros de grafia). De acordo com a semântica de cada campo, diferentes estratégias de preenchimento podem ser adotadas. Por exemplo, para se preencher o campo "autor", somente autores previamente conhecidos (cadastrados no SRI) poderiam ser escolhidos. Já para o campo "título", qualquer seqüência de caracteres seria permitida.

Além desses campos supracitados, de suma importância na descrição de objetos é o preenchimento do campo "assunto" ou "palavra-chave". Em aplicações específicas, o preenchimento de tal campo pode ser dirigido a termos do universo de discurso de uma determinada área (e.g. medicina, computação, arquitetura). Termos que dizem respeito a um domínio específico podem ser agrupados em um vocabulário controlado. Vocabulários Controlados (VC) (SVENONIUS, 1986b) são úteis por organizar termos que referenciam os mesmos elementos (conceitos ou instâncias) ou elementos relacionados de um universo de discurso. Um VC pode ser usado inicialmente para descrever conteúdo e posteriormente para encontrar este conteúdo através de navegação ou pesquisa (WARNER, 2002).

Um exemplo de VC é o Descritores em Ciências da Saúde (DeCS)[4]. Ele define uma terminologia comum para pesquisa, descrição e recuperação de informações entre os componentes do Sistema Latino-Americano e do Caribe de Informação em Ciências da Saúde. Maiores detalhes sobre o DeCS serão vistos na seção 7.6.

Neste trabalho, um VC como o DeCS pode ser representado por um Grafo:

Definição 2 ***- Grafo do Vocabulário Controlado (VC):***

O grafo que representa o Vocabulário Controlado é denotado por $VC(T,A)$*, onde:*

T *é um conjunto de vértices referindo-se a termos.*
A *é um conjunto de arestas, representando as relações semânticas entre os termos de* T*.*

Definição 3 ***- Termo (T):***

[4] http://decs.bvs.br

Um termo $t \in T$ designa um conceito.

Definição 4 - *Associação (A):*

Uma associação $a \in A$ é uma tripla

$a = (origem_a, destino_a, tipo_a)$ *onde:*

$origem_a$, $destino_a \in T$
$tipo_a$ *depende do VC.*

No caso do DeCS $tipo_a \in \{IS\_A, PART\_OF, INSTANCE\_OF\}$

2.4 Anotações semânticas

As descrições de recursos baseadas em termos que referenciam conceitos de um ontologia são chamadas de Anotações Semânticas (AS)(EUZENAT, 2002). No contexto deste trabalho, considera-se Anotação Semântica:

Definição 5 - *Anotação Semântica (AS):*

Seja OIS o conjunto de objetos de informação (veja Definição 1).

Seja T o conjunto de termos de um VC qualquer (veja Definição 2).

Anotar um objeto de informação $oi_i \in OIS$ consiste em associar um ou mais termos $t_i \in T$ ao objeto oi_i, compondo assim $AS(oi_i)$.

$AS(oi_i)$ é o conjunto das anotações semânticas de um certo $oi_i \in OIS$.

Exemplo:

Seja $OIS = \{oi_1, oi_2, oi_3, oi_4\}$
Seja $T = \{t_1, t_2, t_3, t_4, t_5\}$

Durante a catalogação, os OIs oi_1 e oi_2 poderiam ter sido anotados assim:

$AS(oi_1) = \{t_1, t_2, t_5\}$
$AS(oi_3) = \{t_2, t_4\}$

Agora graficamente:

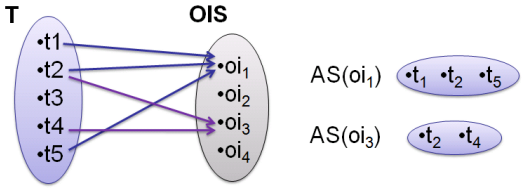


Figura 1: Anotação Semântica de objetos de informação.

AS provêem a OIs um contexto, o qual é ligado ao conhecimento estruturado de um domínio, ou seja, anotações semânticas mapeiam como os objetos estão relacionados com a ontologia (OREN et al., 2006). Estas relações podem ser avaliadas e exploradas em processos de busca, permitindo que objetos que não são explicitamente relacionados com a consulta formulada pelo usuário sejam recuperados. AS podem ser implementadas simplificadamente por meio de tuplas RDF, como será mostrado na seção seguinte.

2.5 Resource Description Framework - RDF

Resource Description Framework (RDF) é uma linguagem do *World Wide Web Consortium (W3C)* que pode ser mapeada para XML e define um modelo de dados formal para descrever informações contidas em recursos da Web. Um recurso Web pode ser uma página, um sitio inteiro ou qualquer item na Web que contenha informação em algum formato. É usada para a construção dos arquivos de metadados que representam os blocos básicos da Web Semântica. RDF está para Web Semântica assim como o HTML está para a Web tradicional (YU, 2007).

RDF é capaz de descrever qualquer recurso, independente de domínio. Provê a base para codificação, troca e reuso de metadados estruturados. A base do RDF é uma tripla do tipo Sujeito-Propriedade-Objeto que representa afirmações (*statements*). Elas assumem a forma Propriedade(Sujeito, Objeto), onde:

- **Sujeito**: é qualquer coisa que está sendo descrito pela expressão RDF. Pode ser um cão, um livro, um ser humano, um sítio Web inteiro, uma

página Web, uma palavra em uma página Web, etc. O sujeito é identificado e nomeado por um *Uniform Resource Identifier (URI)*.

- **Propriedade**: é usada para descrever algum aspecto específico, característica, atributo ou relação de uma dado sujeito.

- **Objeto**: é o valor assumido pela propriedade. Pode ser um literal *(string)* ou um outro sujeito.

Um simples exemplo que usa RDF é mostrado a seguir. A afirmação "Acidente Cerebral Vascular anota o OI_1 e o OI_2" está codificada. "Acidente Cerebral Vascular" é o **Sujeito**, "anota" é a **Propriedade** e "OI_1" e "OI_2" são **Objetos**, ou seja:

anota(Acidente Cerebral Vascular, OI_1) e
anota(Acidente Cerebral Vascular, OI_2).

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
    xmlns:nsRelDeCS="http://decs.org/relationship/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
<rdf:Description
 rdf:about="http://decs/Acidente Cerebral Vascular">
    <nsRelDeCS:anota rdf:resource="http://decs/OI_1"/>
    <nsRelDeCS:anota rdf:resource="http://decs/OI_2"/>
</rdf:Description>
</rdf:RDF>
```

Agora, o modelo RDF do exemplo é mostrado em uma ferramenta gráfica (RDF Gravity)[5]:

Figura 2: Modelo RDF que retrata uma anotação semântica simples.

[5]http://semweb.salzburgresearch.at/apps/rdf-gravity

Neste capítulo apresentamos alguns dos construtos teóricos necessários para a compreensão deste trabalho. Definimos o que são objetos informacionais (OI), vocabulário controlado (VC), ontologias e anotação semântica (AS) segundo os propósitos deste trabalho. Também expusemos o desafio da catalogação de objetos: como prover meios para que um usuário escolha e associe termos de uma ontologia ou de um VC aos OIs de um repositório de forma ágil e precisa?

3 RECUPERAÇÃO DE OBJETOS DE INFORMAÇÃO

Este capítulo aborda a recuperação associativa de OIs baseada em conhecimento. Ela se baseia em representação do conhecimento e técnicas de processamento que operam sobre este conhecimento.

A recuperação de objetos em SRI muitas vezes é realizada por meio de consultas formuladas com palavras-chave. Tal técnica foi popularizada pelos atuais motores de busca da Web e procura fazer a correspondência entre as palavras-chave presentes na consulta formulada pelo usuário e os valores dos metadados que descrevem via anotações os OIs armazenados nos SRI.

Embora tal abordagem seja viável, segundo (CARDOSO, 2000), usuários podem ter dificuldades para descrever suas necessidades de informação através de consultas baseadas em palavras-chave. Esta técnica também é afetada pelas limitações das linguagens naturais (e.g., uso de sinônimos, homônimos, termos genéricos). Além disso, esta abordagem trata cada palavra-chave de forma independente, sem contexto, desconsiderando o significado das mesmas e as relações semânticas que ocorrem em certos domínios de aplicação. Tais relações semânticas podem ser formalizadas em ontologias de domínio e posteriormente exploradas por técnicas de processamento (*reasoning*).

Apesar de todos estes inconvenientes, os motores de buscas de alguns SRI operam deste modo, dificultando a recuperação e reuso de OIs que podem permanecem "perdidos" numa imensidão de conteúdo (GONçALVES, 2007). Segundo (HUSSEIN; NEUHAUS, 2010), o maior problema não é a quantidade de objetos, mas sim encontrar uma forma adequada de se recuperar e filtrar o conteúdo. De nada adianta uma grande coleção de objetos armazenados se não for possível ao usuário recuperar exatamente o que lhe interessa na busca.

No contexto deste trabalho, o problema da recuperação de OIs pode ser definido como:

Definição 6 - *Recuperação de objetos de informação:*

Sejam:

C um conjunto de consultas (queries) passíveis de serem formuladas por um usuário.
$P(c_i)$ um conjunto de palavras-chave p_i que o usuário usa para denotar uma dada consulta $c_i \in C$.
OIS o conjunto que agrupa todos os OIs (veja Definição 1).
$AS(oi_i)$ o conjunto de AS de um certo $oi_i \in OIS$ (veja Definição 5).

Recuperar um Objeto de Informação $oi_i \in OIS$ consiste em se fazer a correspondência entre pelo menos uma palavra-chave $p_i \in P$ (oriunda de uma consulta $c_i \in C$ que visa recuperar oi_i) e alguma anotação $as_i \in AS(oi_i)$.

Por exemplo, um usuário interessado em recuperar OIs que versam sobre Acidente Cerebral Vascular poderia expressar sua necessidade de informação via as seguintes consultas:

- c_1 = O que causa um Acidente Cerebral Vascular?
- c_2 = Que órgãos são acometidos por um Acidente Cerebral Vascular?
- c_3 = Quais os sintomas de um Acidente Cerebral Vascular?
- c_4 = Qual é o tratamento indicado para pacientes acometidos por um Acidente Cerebral Vascular?

De forma genérica:

Seja $C = \{c_1, c_2, c_3, c_4, ...\}$

A consulta c_1 poderia ser traduzida pelo usuário em um conjunto de palavras-chave:

$P(c_1) = \{p_1, p_2, p_3\}$

Seja $OIS = \{oi_1, oi_2, oi_3, oi_4\}$

Durante a catalogação, os OIs oi_1 e oi_3 poderiam ter sido anotados assim:

$AS(oi_1) = \{t_1, t_2, t_5\}$
$AS(oi_3) = \{t_2, t_4\}$

Agora graficamente:

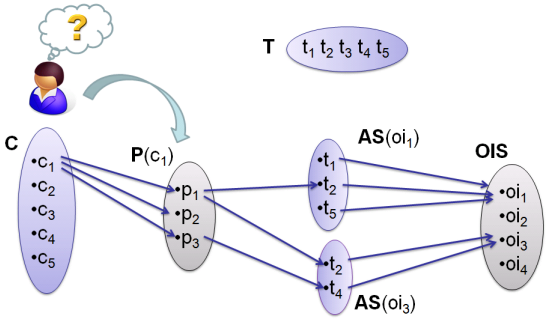


Figura 3: Representação gráfica da recuperação de objetos.

Diante do exposto, o problema da recuperação de objetos consiste em se ampliar as chances de um usuário traduzir sua consulta c_i para um conjunto de palavra-chave $P(c_1)$ e a partir delas um motor de busca retornar todos os objetos cujos termos t_i que os descrevem via anotações estejam relacionados (sintaticamente e/ou semânticamente) a elementos do conjunto $P(c_i)$.

Assim sendo, de acordo com (CRESTANI; RIJSBERGEN, 1993), usuários necessitam de métodos inteligentes e eficientes para acessar esses objetos e a descrição e a organização da informação tornaram-se fundamentais para se alcançar esse propósito. Nesse contexto, a recuperação associativa de objetos apresenta-se como uma alternativa viável. Ela necessita de (i) uma representação do conhecimento e de (ii) técnicas de processamento para operar buscas nesta representação. Respectivamente, tais assuntos serão abordados a seguir.

3.1 Representação do conhecimento

A representação do conhecimento em forma de rede é uma das melhores maneiras de representar o conhecimento de domínio para aplicações de Recuperação de Informação (RI) (do inglês *Information Retrieval - IR*) (BAEZA-YATES; RIBEIRO-NETO, 1999). A principal característica desta abordagem é que ela vê um objeto em termos de suas relações com outros objetos.

A vantagem reside no seu poder de expressão já que em RI o significado de um OI somente pode ser capturado considerando-se suas relações semânticas com outros objetos. Nesta perspectiva, a complexidade dos dados reside nas relações e não nos próprios dados (CRESTANI; RIJSBERGEN, 1993).

Dentre os muitos formalismos que podem ser utilizadas para representar conhecimento de domínio em aplicações de RI, estruturas de representação em rede, como Redes Semânticas são bastante promissoras. Redes Semânticas são a estrutura de representação de conhecimento mais usada para a recuperação associativa de informação. Elas usam a técnica de *Spreading Activation (SA)* como seu paradigma de processamento associativo (CRESTANI, 1997).

3.1.1 Redes semânticas - RS

Desde que foram introduzidas por Quillian em (QUILLIAN, 1968), Redes Semânticas (RS) têm desempenhado um papel significante nas pesquisas sobre representação do conhecimento (CRESTANI, 1997). De acordo com a definição de Quillian, RS expressam conhecimento em termos de conceitos, suas propriedades, e os relacionamentos hierárquicos de sub ou super classes entre os conceitos, formando uma árvore. Cada conceito é representado por um nodo e as relações hierárquicas entre conceitos são representadas por conexões estabelecidas entre nodos através de relações IS-A ou INSTANCE-OF (SCHIEL, 1989).

Os nodos do nível mais baixo da árvore denotam classes ou categorias de indivíduos mais especializados, enquanto que nodos nos níveis mais altos denotam classes ou categorias de indivíduos mais abstratos. Propriedades também são representados por nodos. Uma propriedade que se aplica a um conceito é representada através da ligação do nodo propriedade ao nodo conceito através de uma relação rotulada adequadamente. Normalmente, uma propriedade está ligada ao conceito mais elevado na hierarquia conceitual onde ela se aplica. Se uma propriedade é anexada a um nodo, presume-se que ela se aplica a todos os nodos que são descendentes desse (CRESTANI, 1997). Uma RS adaptada de (CRESTANI, 1997) é mostrada na Figura-4:

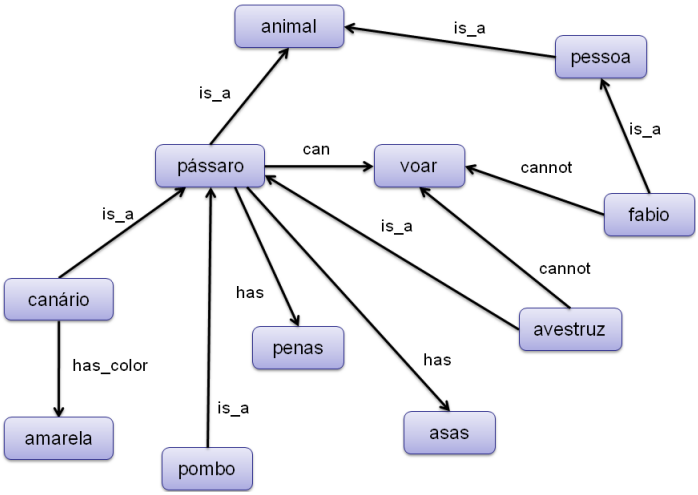


Figura 4: Exemplo de rede semântica.

Dentre as muitas informações que estão representadas nesta RS, é possível se inferir que "o canário, que é animal, especificamente um pássaro, tem asas, tem penas, tem cor amarela e pode voar, enquanto que o avestruz, que por sua vez também é uma animal e mais especificamente um pássaro, apesar de ter asas e de ter penas, não pode voar."

O termo RS tem sido usado em um sentido muito mais geral na literatura de representação do conhecimento do que foi descrito acima, e no que diz respeito a RI, os pesquisadores freqüentemente tem usado o termo RS para se referir a uma Rede Associativa (CRESTANI, 1997).

3.1.2 Redes associativas - RA

Segundo (CRESTANI, 1997), uma Rede Associativa (RA) é uma rede genérica de elementos de informação em que os ítens de informação são representados por nodos e ligações que expressam algumas vezes relações associativas indefinidas e não rotuladas entre elementos de informação. Nodos podem corresponder a termos, documentos, artigos, revistas, assunto, autores, etc. Não há homogeneidade na rede. Um nodo pode representar qualquer coisa. Ligações indicam a associação de um nodo com outro nodo, como, por exemplo, um autor está associado com um documento que ele escreveu,

como mostra a Figura 5, adaptada de (CRESTANI, 1997).

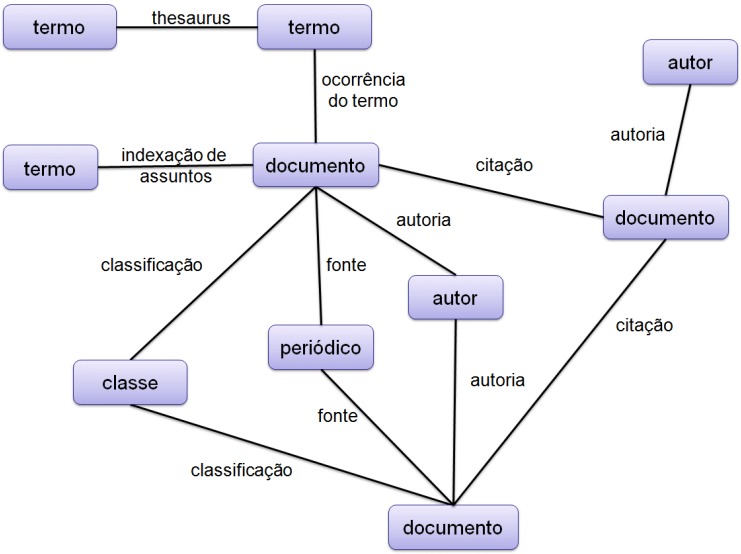


Figura 5: Exemplo de rede associativa.

Na RI moderna, onde as técnicas estatísticas são utilizadas na fase de indexação para associar pesos aos termos do índice, as relações entre os elementos de informação podem ser ponderadas com um peso, adicionando à rede a medida da força das associações entre os elementos. Porém esta abordagem exige muito tempo para a configuração das associações entre os ítens de informação (CRESTANI, 1997)

Tanto Redes Semânticas quanto Redes Associativas são processadas por meio de uma técnica chamada de *Spreading Activation*, a qual é explicada na próxima seção.

3.2 Técnicas de processamento

Inúmeras técnicas de processamento podem operar sobre representações do conhecimento (*Spreading Activation*, Redes Neurais, Redes de Inferência) (CRESTANI, 1997). Enquanto as técnicas de *Spreading Activation* foram inicialmente aplicadas a RS, a recuperação associativa construiu o caminho para aplicações em RI. A idéia fundamental da recuperação associativa

é que informações relacionadas estão conectadas em uma rede. Então informações relevantes podem ser recuperadas, considerando as associações entre conceitos representados nesta rede (BERTHOLD et al., 2009).

A técnica de *Spreading Activation* facilita a extração de subgrafos, nodos, ou arestas relevantes para uma dada consulta disparada sobre uma rede que representa o conhecimento. (BERTHOLD et al., 2009).

3.2.1 Modelo Spreading Activation puro

Spreading Activation (SA) foi proposta por Quillian (QUILLIAN, 1968) e Collins (COLLINS; LOFTUS, 1975) como uma técnica para consulta de redes de informação. Ela tem suas raízes no campo da Psicologia e seu modelo é resultado de estudos sobre os mecanismos da memória humana (ANDERSON, 1983). Foi introduzido pela primeira vez em Ciência da Computação na área de Inteligência Artificial para fornecer um *framework* de processamento para RS (CRESTANI, 1997).

De acordo com (CRESTANI, 1997), o modelo SA puro consiste de (i) uma rede estruturada de dados em que (ii) técnicas de processamento são aplicadas. As próximas seções abordam tais assuntos, respectivamente.

3.2.1.1 Spreading Activation Network - SAN

A rede estruturada de dados do modelo SA é formada de nodos interligados por ligações e de acordo com (TSATSARONIS; VAZIRGIANNIS; ANDROUTSOPOULOS, 2007) é chamada de *Spreading Activation Network* (SAN). Os nodos podem representar objetos ou características dos objetos. Os nodos são geralmente rotulados com os nomes dos objetos que eles representam. As ligações modelam as relações entre os objetos ou as características dos objetos. Ligações podem ser rotulados (com web-link) e/ou ter um peso associado, que segundo (CRESTANI, 1997), adiciona à rede a medida da força da associação entre os objetos. As ligações geralmente são direcionadas, refletindo a relação estabelecida entre os nodos conectados. As relações entre dois nodos diretamente conectados são chamadas de **relações de primeira ordem**. As relações entre dois nodos conectados por meio de um nodo intermediário são chamadas de **relações de segunda ordem** e assim por diante.

Uma SAN é muito semelhante a uma RS, porém é mais geral que a definição da RS apresentada na seção 3.1.1. Uma SAN poderia representar uma RS, mas também uma RA mais genérica (CRESTANI, 1997).

3.2.1.2 Processamento de uma SAN

De acordo com o modelo SA puro, o processamento de uma SAN se baseia unicamente na natureza associativa da representação em rede como estrutura de controle da busca. O processamento de uma SAN se dá por uma seqüência de iterações que ocorrem até que uma condição de término seja atingida. Cada iteração consiste na propagação de ondas de ativação. Tais ondas iniciam em um conjunto de nodos semente e se espalham para todos os outros nodos conectados aos nodos semente. Em cada iteração, a onda de ativação pode sofrer uma "penalização" e enfraquecer à medida que se afasta dos nodos semente. A "penalização" controla a ativação dos nodos da rede, impedindo que todos os nodos sejam ativados.

O efeito é similar a uma pedra que cai num lago, como mostra a Figura 6. São geradas ondas, sendo que as mais próximas do local onde a pedra caiu são maiores. À medida que as ondas se afastam da origem, elas vão diminuindo de tamanho até que cessa a perturbação na água.

Figura 6: Metáfora da propagação das ondas de ativação em uma SAN.

Depois desta visão geral, o foco agora são os detalhes do algoritmo de SA usado para processar uma SAN. A técnica de processamento é definida por uma seqüência de iterações, como descreve esquematicamente a Figura-7, baseada em (CRESTANI, 1997). Cada iteração é seguida por outra iteração. O processo segue até ser interrompido pelo usuário ou até que alguma condição de término seja alcançada. Uma iteração é composta por:

1. um ou mais pulsos;
2. verificação de término.

O que distingue o modelo SA puro de outros modelos mais complexos é a seqüência de ações que compõe o pulso. Um pulso é composto de três

fases:

1. pré-ajuste
2. propagação
3. pós-ajuste

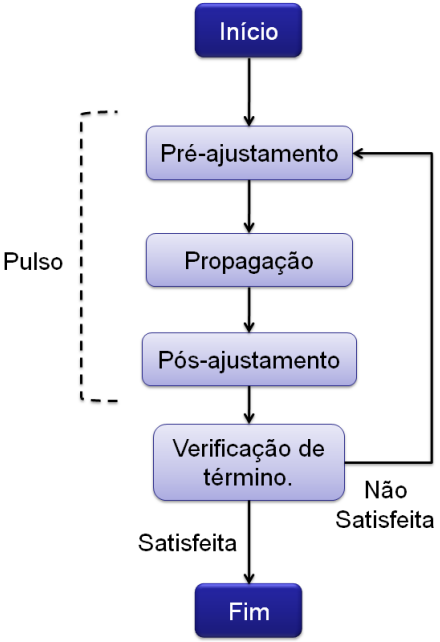


Figura 7: Processamento no modelo SA puro.

Nas fases opcionais de **pré-ajuste** e **pós-ajuste**, alguma forma de decaimento da ativação pode ser aplicado aos nodos ativos. Estas fases são usadas para evitar a retenção de ativação de pulsos anteriores, permitindo controlar tanto a ativação individual de nodos como a ativação global da rede. Estas fases implementam uma forma de "perda de interesse" em nodos que não são continuamente ativados.

A fase de **propagação** consiste da passagem de ondas de ativação de um nodo para todos os outros nodos conectados a ele. Há muitas formas de

disseminar a ativação sobre a rede. Na sua forma mais simples, no nível de um nodo, o algoritmo de SA primeiramente faz o cálculo da ativação do nodo. A ativação de um nodo k no pulso p é dada por:

$$A_k(p) = A_k(p-1) + \sum(O_j(p)).(W_{jk}).(D) \tag{3.1}$$

onde:

$A_k(p)$ é a ativação do nodo k no pulso p
$A_k(p-1)$ é a ativação do nodo k no pulso $(p-1)$
$O_j(p)$ é a saída *(output)* do nodo j conectado ao nodo k no pulso p
W_{jk} é o peso do link que conecta o nodo j ao nodo k
$D \in [0,1]$ é o fator de decaída

Esta definição foi adaptada de (TSATSARONIS; VAZIRGIANNIS; ANDROUTSOPOULOS, 2007) para aqui contemplar a soma coma a ativação alcançada no pulso anterior (p-1) e impor a penalização do fator de decaída D. Esta fórmula não é aplicada aos nodos semente. Geralmente eles recebem o valor de ativação máximo suportado pela aplicação (ex: 1 ou 100%). Quanto aos demais nodos, inicialmente eles recebem o valor mínimo de ativação (ex: 0 ou 0%). O valor de ativação e o peso das arestas normalmente são números reais, no entanto o seu tipo numérico é determinado pelas necessidades específicas da aplicação modelada.

Depois que o valor de ativação de um nodo é calculado, seu valor de saída deve ser determinado. Nos modelos SA, normalmente não há distinção entre "ativação" ou "saída" de um nodo. O nível de ativação de um nodo é o seu valor de saída. Isso geralmente é computado como uma função do valor de ativação:

$$O_j(p) = f(A_j(p)) \tag{3.2}$$

onde:

$O_j(p)$ é a saída *(output)* do nodo j no pulso p.
$A_j(p)$ é a ativação do nodo j no pulso p

Há muitas funções que podem ser utilizados na avaliação da saída. A função mais comumente usada em modelos SA puros é a função de limiar. Ela estabelece um limite que é usado para determinar se um nodo j deve de ser considerado ativo e, portanto, disparar a ativação para outros nodos. Um

exemplo de função de limiar é a seguinte:

$$O_j = \begin{cases} 0 & \text{se } A_j < k_j \\ 1 & \text{se } A_j > k_j \end{cases} \tag{3.3}$$

onde:

O_j é a saída *(output)* do nodo j
A_j é a ativação do nodo j
k_j é valor de limiar para o nodo j

O valor de limiar da função de ativação é dependente da aplicação e pode variar de um nodo para outro. Logo a notação k_j para o limiar do nodo é usada. Depois de computado o valor de saída do nodo j, ele dispara para todos os nodos conectados a ele, normalmente enviando o mesmo valor para cada um dos nodos.

SA é iterativo, composto por uma seqüência de pulsos p e verificação das condições de término. A ativação se inicia na(s) semente(s) e pulso após pulso se espalha pela rede atingindo os nodos que estão longe dos nodos ativados inicialmente. Após um número de pulsos ter sido alcançado ou todos os nodos terem sido disparados, a condição de término é alcançada. Então o processo pára, caso contrário se inicia uma nova série de pulsos. A Figura 8 ilustra tal comportamento, onde "Transtornos Cerebrovasculares" é a semente e 3 pulsos são propagados sobre a rede.

O resultado do processo de SA é um subgrafo composto pelos de nodos que foram ativados até a condição de término. Em nosso exemplo, "Transtornos Cerebrovasculares", "Acidente Cerebral Vascular", "Encefalopatias", "Doenças Vasculares", "Doenças Cardiovasculares" e "Doenças do Sistema Nervoso Central" seriam tais nodos. A interpretação do nível de ativação de cada nodo depende da aplicação e principalmente das características do objeto que está sendo modelado por aquele nodo.

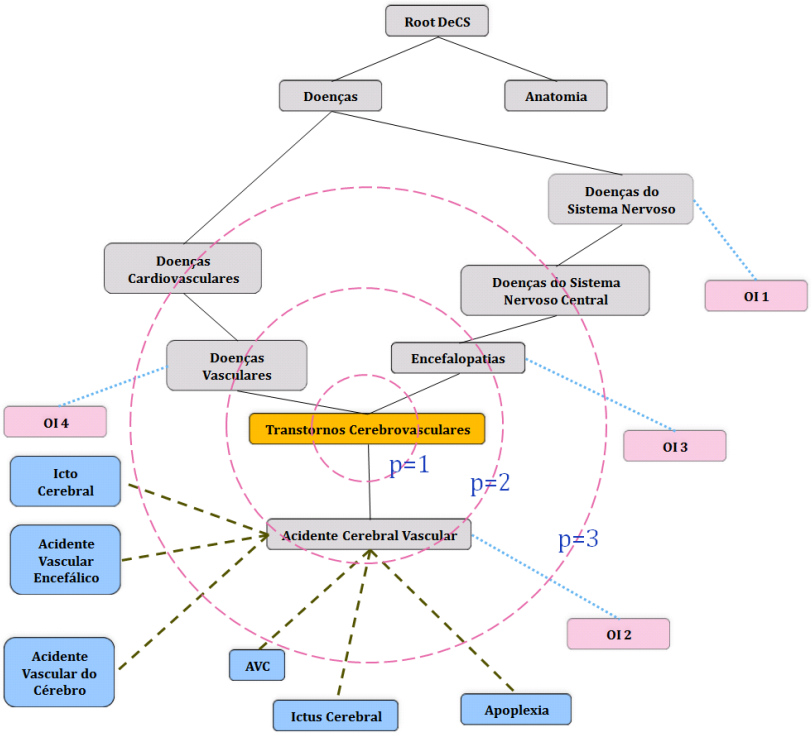


Figura 8: Ondas de ativação do SA.

Na Definição 7 um pseudo código do algoritmo de SA é mostrado [1].

Definição 7 - ***Algoritmo SA:***

Dado um conjunto de Sementes S.

Seja Condição de término:

- *quando todos os nodos são disparados;*
- *quando um número máximo de pulsos é alcançado;*

[1] http://en.wikipedia.org/wiki/Spreading_activation#Algorithm

```
SA sa = new SA();
sa.executarSA(S);

SA{
    NodosParaDisparo = {};
    float LimiteDisparo = 0.5;

    executarSA(S){
        para cada s pertencente a S{
            /* sementes recebem ativação de 100% */
            atribuirAtivaçãoMaxima(s);

            /* apenas altera uma flag do nodo */
            marcarNodosComoDisparado(s);

            /* nodos ligados a "s" por uma
            aresta de tamanho 1 */
            Adjacentes = retornaAdjacentesDe(s);
            marcarNodosParaDisparo(Adjacentes);
        }

        enquanto(not Condição de término) e
            (NodosParaDisparo.size() > 0){
                para cada n pertencente NodosParaDisparo{
                    calcularAtivação(n);
                    dispararNodo(n);
                    NodosParaDisparo.remove(n);

                    /* apenas altera uma flag do nodo */
                    marcarNodosComoDisparado(n);
                }
        }
    }

    dispararNodo(s){
        se (valorAtivação(s) > LimiteDisparo){
            /* nodos ligados a ''s'' por uma
            aresta de tamanho 1 */
            Adjacentes = retornaAdjacentesDe(s);
            marcarNodosParaDisparo(Adjacentes);
        }
    }

    marcarNodosParaDisparo(Adjacentes){
       para cada n pertencente a Adjacentes{
            /* apenas consulta uma flag do nodo */
            se (not disparado(n)){
                /* apenas altera uma flag do nodo
                ele será disparado no próximo pulso */
```

```
                    marcarNodosParaDisparo(n);

                    /* adiciona à lista de nodos
                    marcados para disparo */
                    NodosParaDisparo.add(n);
                }
            }
        }

        calcularAtivação(n){
            /* ver fórmula 3.1 */
        }
}
```

3.2.2 Modelo Spreading Activation aprimorado

Segundo (CRESTANI, 1997), o modelo SA puro apresenta alguns inconvenientes:

- se não for controlada com cuidado, por meio das fases de pré-ajustamento e pós-ajustamento, a ativação acaba se espalhando por toda a rede;
- as informações fornecidas pelos rótulos associados aos *links* não são usadas, ou seja, não há utilização da semântica das associações;
- é difícil de implementar alguma forma de inferência baseado na semântica das associações.

Porém, de acordo com Crestani, estes problemas podem ser atenuados levando-se em conta o significado das relações entre os nodos, ou seja, utilizando-se da informação fornecida pelos rótulos dos *links*. Cada *link* poderia ser tratado e processado de forma especial, de acordo com sua semântica. Desse modo seria possível se implementar alguma forma de heurística ou espalhar a ativação da rede de acordo com regras de inferência. Uma maneira de se propagar a ativação de acordo com regras dá-se por meio de restrições na propagação. Algumas das restrições geralmente usadas em modelos SA são:

- **restrição de distância**: a propagação da ativação deve cessar quando atinge nodos que estão longe daqueles ativados inicialmente (sementes). Isso corresponde a simples regra heurística de que a força da relação entre dois nodos diminui com a distância semântica. É comum

considerar apenas relações de primeira, segunda e, no máximo, terceira ordem. Porém este critério é dependente da aplicação.

- **restrição de espalhamento**: a propagação da ativação deve cessar ao atingir nodos com conectividade muito alta, ou seja, que estão conectados a um número muito grande de outros nodos. Assim evita-se que grandes áreas da rede sejam ativadas.

- **restrição de caminho**: a ativação deve espalhar usando caminhos preferenciais, refletindo as regras de inferência dependentes da aplicação. Isso pode ser modelado via pesos nos *links* ou, se os *links* são rotulados, desviando o fluxo de ativação para um caminho em particular, enquanto deixa de seguir outros caminhos menos significativos.

- **restrição de ativação**: usando a função de limiar a nível de nodos, é possível controlar a propagação da ativação na rede. Isto pode ser conseguido pela alteração do valor de limiar em relação ao nível total de ativação de toda a rede em qualquer pulso. Além disso, é possível atribuir diferentes limiares para cada nodo ou conjunto de nodos em relação ao seu significado no contexto da aplicação. Embora isso possa provocar um aumento no número de cálculos, torna possível implementar várias regras de inferência complexas.

Algumas destas restrições atuam durante a fase de pré-ajustamento (restrição de distância, restrição de espalhamento, restrição de caminho) ou durante a fase pós-ajustamento (principalmente para as restrições de nível de ativação). Estas estratégias de controle atuam na propagação da ativação sobre a rede e visam dar preferência a certas associações, evitando que o espalhamento da ativação ocorra na rede inteira. Além disso, outra vantagem da ativação não se espalhar pela rede inteira é a redução do esforço computacional. Tais restrições podem ser considerados como uma aprimoramento do modelo SA puro.

Neste capítulo o embasamento teórico sobre a recuperação de recursos baseada em conhecimento foi pavimentado. Discorre-se sobre a representação de conhecimento com Redes Semântica (RS) e Redes Associativas (RA), a operacionalização de buscas associativas sobre RS, RA e SAN via a técnica de Spreading Activation (SA). Ainda definiu-se o problema da Recuperação de OIs e o algoritmo de SA.

4 VISUALIZAÇÃO DE CONHECIMENTO

Neste capítulo algumas técnicas e ferramentas de visualização usadas para se explorar conhecimento serão vistas.

VCs e outras formas de denotação de conhecimento, tais como ontologias, apresentam um emaranhado de termos, conceitos e relações semânticas que podem dificultar o seu entendimento (JUDELMAN, 2004). Sendo assim, é essencial prover meios para que usuários explorem e naveguem sobre a estrutura de conhecimento do domínio, tanto durante a anotação quanto na recuperação de OIs. Quais técnicas e ferramentas visuais melhor desempenham este papel? Como os recursos cognitivos visuais humanos envolvidos no entendimento de uma representação do conhecimento podem ser explorados? Estas questões motivaram este trabalho a analisar técnicas e ferramentas de visualização de conhecimento para amparar usuários na escolha dos termos de VCs mais adequados para anotar ou recuperar OIs.

4.1 Princípios básicos

Para se estabelecer o constructo teórico desta seção, é preciso investigar como se dá a percepção humana bem como se estabelecer um entendimento a cerca do que é conhecimento. Isto será feito aqui de forma pragmática e sucinta, dado que várias questões relativas a estas áreas ainda estão em aberto e não há um consenso a respeito de modelos e definições universais tanto sobre conhecimento quanto sobre como se dá a cognição e a percepção humana (JUDELMAN, 2004).

4.1.1 Cognição e percepção humana

Segundo (DIAS; CARVALHO, 2007), o entendimento sobre como funciona a cognição do ser humano mostra-se valoroso quando da elaboração de bens de consumo em geral, estruturas de visualização de informação, *softwares* e suas interfaces. Ainda, o oferecimento de recursos gráficos com a finalidade de apresentar informação "produz a compreensão da mensagem transmitida, pois esta se torna mais natural e exige menos esforço cognitivo". Amparados por este entendimento, tais artefatos podem ser mais bem planejados a fim de se tornarem inteligíveis.

Os mesmos autores também defendem que "a palavra visualizar está

intimamente relacionada a transformar o abstrato em imagens que podem ser modelos mentais ou estruturas gráficas reais. Contudo, o objetivo maior é auxiliar no entendimento de algum assunto que, sem uma visualização, exigirá maior esforço e/ou tempo para ser compreendido". E ainda reconhecem que, "na maioria dos casos, o oferecimento de imagens, figuras, estruturas gráficas e quaisquer outros recursos gráficos, com a finalidade de apresentar uma informação, produz a compreensão da mensagem transmitida, pois esta se torna mais natural e exige menos esforço cognitivo".

Em (JUDELMAN, 2004), Judelman defende que "visualizações aproveitam o poder cognitivo visual e espacial para reduzirem o esforço requerido para o processamento de informações complexas e que, através do mapeamento de parâmetros de dados para localização, cores, ou formas produzem imagens que podem revelar objetos, padrões e relacionamentos indetectáveis quando apresentados na forma de listas ou tabelas". Enfatizando ainda mais o poder das visualizações, o autor afirma que "visualizações podem idealmente ser ferramentas de pensamento e aprendizado, estendendo o processo cognitivo por permitir a ativa exploração de um espaço do conhecimento".

4.1.2 Visualização de informação e de conhecimento

(JUDELMAN, 2004) preconiza que "o grande desafio hoje não é, necessariamente, produzir novos conhecimentos, mas desenvolver modos de melhor trabalhar com ele e dar sentido àquele conhecimento que nós já possuímos". Em vista disso, uma estratégia para se redescobrir conhecimentos já aceitos, bem como facilitar a exploração de novos conhecimentos pode se dar via a apresentação adequada deste conhecimento. Para tanto existem ciências como a Visualização de Informação e a Visualização de Conhecimento que estudam estratégias e técnicas para promover a apresentação adequada das informações e do conhecimento (FREITAS et al., 2001).

Em (BURKHARD; MEIER, 2004), os autores definem Visualização de Conhecimento como o uso de representações visuais para transferir conhecimento entre pelo menos duas pessoas. Intimamente relacionado a este contexto é que surge a Visualização da Informação que, segundo (FREITAS et al., 2001), tem por objetivo o estudo de representações gráficas para a apresentação de informações e visa ajudar a dedução de novos conhecimentos baseados no que está sendo apresentado. Ela é uma ciência que combina aspectos de computação gráfica, interação humano-computador, cartografia e mineração de dados. De acordo com (DIAS; CARVALHO, 2007), "as estruturas de visualização da informação podem contribuir para a tomada de

decisão, descoberta de novos conhecimentos, demonstração de esquemas, representação de idéias e análise das informações, que podem tornar mais ágil a apropriação de conhecimento por parte do usuário, ao observar que tais estruturas oferecem novos conhecimentos que são informados por meio de objetos visuais".

4.2 Técnicas de visualização em GUI

A interface gráfica do usuário (*Graphical User Interface - GUI*) (TUCK, 2001) surgiu, aproximadamente, em meados da década de 80 e denota o meio, a superfície de contato para comunicação humano-computador (DIAS; CARVALHO, 2007). É por meio dela que usuários demonstram suas intenções a um dado sistema. Por exemplo, em sistemas de busca baseados em palavras-chave, é através da entrada de dados via um campo de texto na interface que o sistema de busca passa a receber, por exemplo, a necessidade de um usuário por objetos informacionais que discorrem sobre "Apoplexia".

Tal exemplo simplório retrata uma situação em que o usuário sabe ou faz idéia do que está procurando. Porém, segundo (SILVA, 2007), os usuários podem não saber ou ainda passarem por dificuldades para descrever suas necessidades de informação. As palavras-chave utilizadas pelos usuários na consulta podem diferir daquelas utilizadas na anotação dos OIs durante a catalogação. Além disso, os usuários podem não ter pleno domínio da área de conhecimento em questão, usando palavras-chave que não são os termos utilizados para representar aquilo que pretendem encontrar. Neste caso, interfaces de navegação podem amparar os usuários, pois permitem a execução de processos exploratórios através da navegação em uma visão do conhecimento de domínio, auxiliando-os a descobrir o que desejam.

Interfaces homem-computador podem incluir facilidades gráficas e de navegação que influenciam no entendimento de um domínio, na anotação e na recuperação de OIs. Quais dessas facilidades são as mais adequadas para se desenvolver um sistema para anotação e recuperação de OIs? Essa e outras questões norteiam o trabalho de (SILVA, 2007), que destaca algumas das técnicas em interfaces gráficas que podem ser utilizadas para a visualização e navegação sobre conhecimento. Baseado em tal trabalho é que, doravante, são destacadas as principais técnicas utilizadas para organização e representação de conhecimento em interfaces gráficas, a citar Mapas Conceituais, Mapas Hiperbólicos e os Diagramas Hierárquicos, respectivamente.

4.2.1 Mapas Conceituais - MC

Mapas Conceituais (MC)(NOVAK, 2006) podem ser definidos como um conjunto de conceitos e relacionamento entre conceitos. Eles têm seu referencial teórico baseado na Teoria da aprendizagem ou Teoria de Assimilação de David Paul Ausubel (SILVA, 2007), a qual afirma que o conhecimento apreendido por um indivíduo é armazenado na estrutura cognitiva deste indivíduo. Tal estrutura cognitiva pode ser descrita como um conjunto de conceitos, organizados de forma hierárquica, representando assim o conhecimento e as experiências adquiridas por uma pessoa. O conceito é um termo que representa uma série de objetos, eventos ou situações que possuem atributos comuns.

Baseando-se nessa teoria é que Joseph D. Novak (NOVAK, 2006) desenvolveu a metodologia de MCs, procurando assim representar como o conhecimento é armazenado na estrutura cognitiva de uma pessoa. Segundo Novak, "MCs são interfaces para a organização e representação do conhecimento. Eles colocam os conceitos, geralmente dentro de círculos ou retângulos de algum tamanho, e os relacionamentos entre os conceitos ou as proposições são indicados por uma linha que conecta os dois conceitos. Palavras sobre as linhas especificam a relação entre os dois conceitos". Novak e sua equipe do *Florida Institute for Human and Machine Cognition* (IHMC) criaram um *software* para construção de MCs, o CMap Tools[1], que usamos para elaborar um exemplo de MC, que é mostrado na Figura 9. Tal MC retrata de forma simplificada nosso entendimento acerca do VC DeCS.

[1] http://cmap.ihmc.us

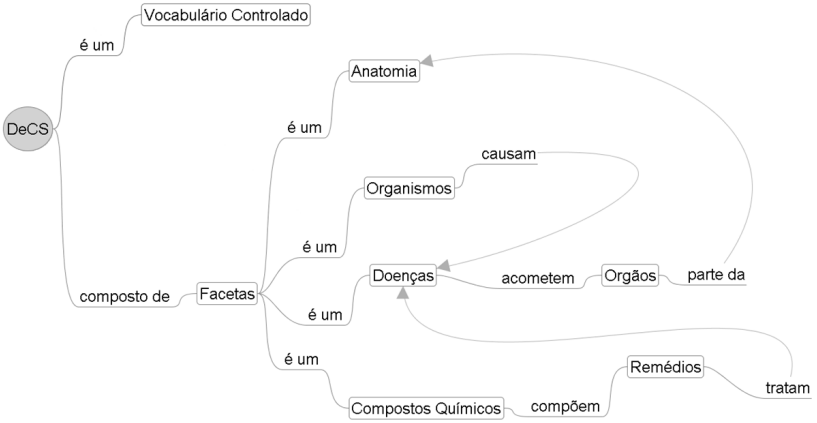


Figura 9: Exemplo de mapa conceitual acerca do DeCS.

Tal MC expressa que o "DeCS" é um "Vocabulário Controlado" composto de "Facetas" (ou categorias) como "Compostos Químicos", os quais compõem os "Remédios" que tratam das "Doenças" que acometem "Orgãos" que são parte da "Anatomia".

4.2.2 Mapas Hiperbólicos - MH

Segundo (RODRIGUES JR. et al., 2008), os neurocientistas afirmam que, embora o cérebro possa perceber vários alvos visuais simultaneamente, ele não pode processá-los em paralelo. A solução então é restringir os objetos apresentados aos olhos. Para isso, o mecanismo da visão se concentra em pequenas regiões do campo visual, considerando objetos únicos, um após o outro, num processo seqüencial regido pelo que se chama de atenção. Esse aspecto é explorado nas visualizações de Mapas Hiperbólicos (MH).

MHs, assim como MCs, apresentam hierarquias de conceitos. Porém, a interface de MHs enfatiza a apresentação do foco (ao centro) em detrimento do seu entorno, cujo nível de detalhamento é decrementado segundo uma função hiperbólica à medida que os objetos apresentados encontram-se afastados do centro em direção à periferia (LAMPING; RAO; PIROLLI, 1995). O resultado é uma interface moderna e atrativa (*fisheye*), que facilita a navegação e a visualização de grandes volumes de conhecimento, pelo estabelecimento de um foco sem perda do contexto. A partir de conceitos genéricos o usuário pode explorar a massa de conhecimento e a partir de certos focos, pode facil-

mente acessar conceitos relacionados.

Este comportamento provido pelos MHs cumpre um papel fundamental na descoberta de novos conhecimentos, principalmente onde a exploração se dá em estruturas que reúnem tal conhecimento e cuja topologia mais representativa é a de um grafo em forma de árvore. Isso fica evidenciado nas palavras de (SILVA, 2007), que prega que a navegação é o método mais adequado para aquele usuário "que não sabe precisamente o que quer ou como conseguir a informação desejada. A partir de conceitos mais genéricos, o usuário pode encontrar conceitos mais específicos que correspondam ao que ele estava procurando." Na Figura 10 é mostrado um exemplo de MH de um trecho do tal VC DeCS.

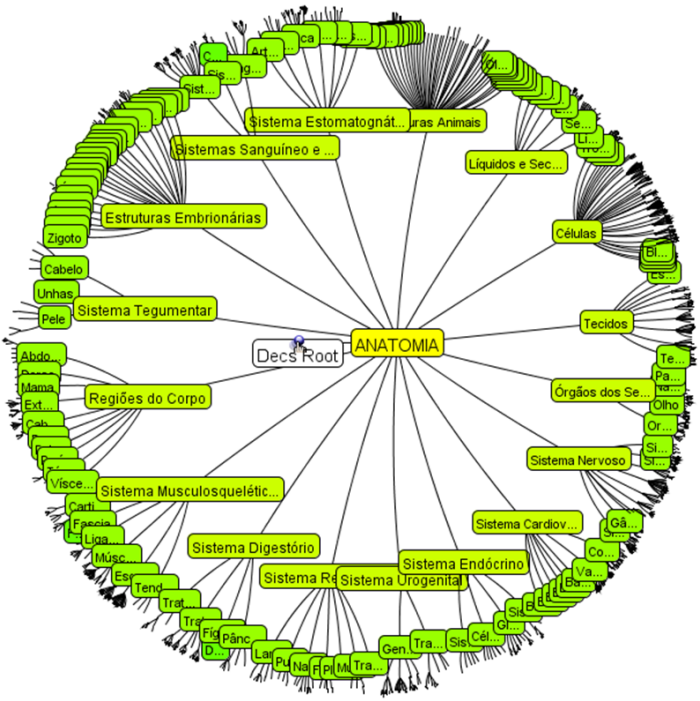


Figura 10: Exemplo de mapa hiperbólico de um trecho do DeCS.

4.2.3 Diagramas Hierárquicos - DH

Técnicas de navegação, como MHs, embora úteis para explorar grandes massas de conhecimento, mostram-se ineficazes na localização de conceitos específicos. Os usuários podem se perder em visões caóticas, preferindo visualizações que oferecem uma navegação mais gradativa e ordenada. Neste contexto destacam-se as visualizações hierárquicas, as quais aproveitam a estrutura semântica para orientar o acesso à hierarquia de conceitos (KATIFORI et al., 2007). Além disso, os Diagramas Hierárquicos (DH) permitem visualizações condensadas, possibilitando que o usuário veja apenas o conhecimento relevante (SILVA, 2007). No DH da Figura 11, novamente é mostrado um trecho do tal VC DeCS.

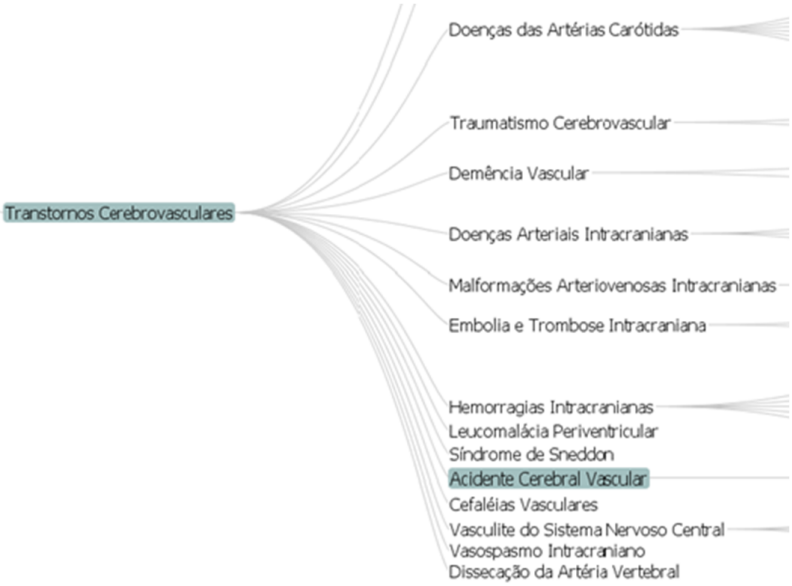


Figura 11: Exemplo de diagrama hierárquico de um trecho do DeCS.

4.3 Ferramentas de visualização: comparativo

Várias ferramentas implementam as técnicas de visualização de conhecimento expostas até aqui. Baseando-se em pesquisas na Web e principal-

mente no sítio da GEMI Bioinformatics[2] elencou-se algumas delas a fim de se fazer uma avaliação comparativa:

1. Prefuse (http://prefuse.org)
2. Treebolic (http://treebolic.sourceforge.net/en/index.html)
3. SpaceTree (http://www.cs.umd.edu/hcil/spacetree)
4. VisNomad (http://visnomad.org/wiki/index.php?title=Main_Page)
5. Last.forward (http://lastforward.sourceforge.net/index.html)
6. yWorks (http://www.yworks.com/en/index.html)
7. ManyEyes (http://www-958.ibm.com/software/data/cognos/manyeyes)
8. Baobab (http://162.38.181.25/ jdutheil/Baobab/Baobab.html)
9. TreeDyn (http://www.treedyn.org)
10. Otter (http://www.caida.org/tools/visualization/otter)
11. TreePlus (http://www.cs.umd.edu/hcil/treeplus)

Das diversas ferramenta disponíveis, procurou-se selecionar primordialmente aquelas de código fonte aberto, com documentação, gratuitas, personalizáveis, com suporte ativo ou atualizações recentes, que são portáveis para a Web e apresentem interatividade e animação. A partir da listagem inicial, este trabalho avaliou 5 dessas ferramentas, para a quais encontrou-se documentação e ao menos código executável para permitir testes:

1. Prefuse
2. Treebolic
3. SpaceTree
4. yWorks
5. TreePlus

[2] http://www.treedyn.org/overview/editors.html

As demais não puderam ser testadas porque não ofereciam sequer código executável ou suporte à personalização exigida pelos exemplos que planejamos inicialmente retratar. Os critérios utilizados na análise comparativa dessas ferramentas foram adaptados de (SILVA, 2007). A análise se deu sob duas óticas: (i) os aspectos funcionais (características conceituais, navegação, comportamento, etc.) e (ii) aspectos não-funcionais (tecnológicos). A análise funcional avalia se as ferramentas conseguem exprimir adequadamente o conhecimento de um domínio de aplicação. Já os aspectos não-funcionais visam aferir se é possível utilizar e estender as ferramentas (código fonte aberto, documentação, são gratuitas, extensíveis, executam na web, etc.).

4.3.1 Aspectos funcionais

Os aspectos funcionais englobam as capacidades de representação conceitual e as facilidades de navegação das ferramentas. A Figura 12 apresenta as capacidades de representação e visualização e a Figura 13 apresenta as facilidades de navegação suportadas pelas ferramentas analisadas.

Capacidade de representar	*Prefuse*	*Treebolic*	*SpaceTree*	*yWorks*	*TreePlus*
Mapas conceituais	X				
Visualização hiperbólica	X	X			
Diagramas hierárquicos	X		X	X	X
Diferentes tipos de conceitos e diferenciá-los		X		X	
Diferentes tipos de relações entre conceitos e diferenciá-los		X		X	
Informações adicionais sobre cada conceito	X	X			
Herança múltipla		X		X	X

Figura 12: Capacidades de representação e visualização.

Facilidade	*Prefuse*	*Treebolic*	*SpaceTree*	*yWorks*	*TreePlus*
Expansão e contração da estrutura	X	X	X		X
Busca textual e destaque nos resultados	X	X	X	X	X
Foco ou ênfase no conceito observado (*fisheye*)	X	X	X		X
Ordenação alfabética dos conceitos subordinados					
Visualização do caminho da cadeia observada até a raiz	X	X	X	X	X

Figura 13: Facilidades de navegação.

4.3.2 Aspectos não-funcionais

Os aspectos não-funcionais das ferramentas estão na Figura 14.

Nome	Status	Customização			Ext.	Código Aberto	Web	Doc.	Ling.	t1	t2
		Nós	Arestas	Dados							
Prefuse	versão Beta	via código fonte	via código fonte	via parâmetro	via código fonte	sim	sim	sim	Java	30	40
Treebolic	concluído em 2008	via parâmetro	via parâmetro	via parâmetro	via código fonte	sim	sim	sim	Java	28	35
jpaceTree	concluído em 2003	?	?	via parâmetro	?	não	sim	sim	Java	12	21
yWorks	ativo	?	?	?	?	não	sim	?	Java.Net Flex	20	18
TreePlus	concluído	via código fonte	via código fonte	via parâmetro	via código fonte	sim	?	sim	C#	40	19

Figura 14: Características não-funcionais das ferramentas.

Esses dados foram coletados de artigos, páginas Web e documentação técnica sobre as ferramentas. O símbolo de interrogação (“?”) em algumas células indica que não foi possível aferir (por falta de informação) a respectiva característica para a ferramenta, mesmo entrando em contato com os responsáveis em alguns casos. A coluna “Status” indica o estado do desenvolvimento da ferramenta e se tem havido continuidade na sua evolução (novas versões). A coluna “Customização” indica as possibilidades de customização da ferramenta no que se refere à representação e apresentação dos nodos, arestas e dados a eles associados. A possibilidade de se aprimorar o comportamento ou incluir novas funcionalidades é contemplada pela coluna “Extensível” (“Ext.”), que apresenta certa correlação com a disponibilidade do código fonte (coluna “Código Aberto”). A coluna “Web” indica se a ferramenta executa no ambiente Web, o que hoje em dia é primordial. A disponibilidade de documentação, aspecto analisado na coluna “Documentação” (“Doc.”) é indispensável para orientar o uso da ferramenta e facilitar alterações no código fonte. A coluna “Linguagem” (“Ling.”) indica a linguagem em que a ferramenta foi desenvolvida, o que tem certa correlação com as plataformas nas quais ela pode ser executada.

Além da análise de artigos e documentação das ferramentas, que resultou na coleta das informações descritas na Figura 12 e Figura 13, foram realizados testes para medir o tempo gasto para construir a visualização apresentada na Figura 20. Tal exemplo serviu como uma espécie de *benchmarck* para dois alunos de graduação da área de tecnologia da informação avaliarem as funcionalidades e a facilidade de uso de cada uma das ferramentas. As colunas “t1” e “t2” apresentam o tempo gasto (em minutos) por cada um dos dois avaliadores para retratar o exemplo da Figura 20 em cada uma das ferramentas. Este teste também ajudou a levantar limitações e potencialidades das ferramentas. A maior dificuldade foi alimentá-las com dados para montar uma estrutura como a da Figura 20. Todas as ferramentas analisadas, exceto a *Treebolic*, exigiram a geração manual de dados em XML.

4.4 Ferramentas de visualização selecionadas

Valendo-se do estudo comparativo realizado, as ferramentas de visualização a serem empregadas na anotação e busca de OIs foram selecionadas. ***Treebolic*** é considerada mais adequada para navegação hiperbólica porque preenche a maioria dos requisitos funcionais e não funcionais. De forma análoga, ***Prefuse*** apresenta os melhores recursos para visualização e navegação em hierarquias e mapas conceituais. A *TreePlus* destacou-se tanto

quanto o *Prefuse* para visualização hierárquica, porém não foi escolhida por exigir ferramentas de desenvolvimento pagas e sua execução na Web demandar recursos adicionais.

Treebolic inclui vários módulos, dentre os quais o *Treebolic Generator*, que facilita a construção de mapas hiperbólicos através de uma interface gráfica que gera um arquivo XML utilizado para alimentar o applet de visualização. Outras fontes de dados podem ser acessadas via provedores de dados (e.g., XML, SQL). O código fonte Java da *Treebolic* está disponível sob licença *OpenSource GPL*, assim como uma vasta documentação, possibilitando a sua customização. O principal ponto negativo da *Treebolic* é que ela não recebe nenhuma atualização desde 10 de Fevereiro de 2009.

Prefuse oferece diversas formas de visualização de mapas conceituais na forma de grafos e estruturas hierárquicas (*GraphView*, *TreeView*, *TreeMap*, etc.). A visualização na forma *TreeView* permite ao usuário navegar deste a raiz de um VC até o termo desejado, com recursos de zoom, deslocamento de componentes da estrutura e destaque de componentes, entre outros. O código fonte Java está disponível nos termos da licença *BSD (Berkeley Standard Distribution)*, tornando a ferramenta livre inclusive para uso comercial. O *Prefuse* conta uma boa documentação técnica, facilitando o seu uso e customização, embora alguns manuais ainda estejam "em construção". Os pontos negativos do *Prefuse* são que a última versão disponível é de 21 de Outubro de 2007, sem receber nenhuma atualização deste então. Este panorama tende a se manter, já que seu mentor não mais faz parte do projeto. Além disso, a customização do formato dos nodos e das arestas demanda codificação, já que não pode ser feita via parametrização. A complexidade para se utilizar a *Application Programming Interface (API)* do *Prefuse* também é outro ponto negativo.

Neste capítulo discorremos sobre técnicas de visualização de conhecimento, a citar, Mapas Conceituais (MC), Mapas Hiperbólicos (MH) e Diagramas Hierárquicos (DH). Também comparamos ferramentas de visualização que implementam tais técnicas, sendo que as ferramentas *Prefuse* e *Treebolic* se destacaram e foram eleitas para a implementação de interfaces gráficas úteis para se explorar conhecimento.

5 VISUALIZAÇÕES DE CONHECIMENTO NA ANOTAÇÃO

Este capítulo descreve em detalhes nossa proposta para amparar usuários com visualizações do conhecimento durante a anotação de OIs.

Este trabalho preconiza o uso de interfaces Web baseadas em conhecimento para amparar a anotação de OIs, facilitar a recuperação e promover o reuso de tais OIs. Este objetivo ambicioso se apóia numa estrutura que reúne uma base de conhecimento, um método que opera buscas sobre esta base e uma série de interfaces Web baseadas em conhecimento. A operacionalização de tais módulos faz parte de nossa proposta. Agora os meandros da mesma serão vistos.

5.1 CIBELE: visão geral

CIBELE é o nosso sistema de recuperação de informação baseado em conhecimento que suporta recuperação associativa de OIs via *Spreading Activation* sobre uma Rede Semântica. Esta abordagem para a operacionalização de sistemas baseados em conhecimento e técnicas de visualização para anotação e recuperação do conteúdo armazenado em SRIs é composta de 4 passos enumerados na Figura 15. Tal figura ilustra o encadeamento lógico e o fluxo de dados que ocorre entre os componentes dispostos na camada de interfaces, de serviços e de persistência, os quais operacionalizam nossa proposta. No Anexo Q um diagrama elaborado com a *Unified Modeling Language* (UML) descreve tal estrutura segundo a dependência dos componentes.

A Figura 15 ilustra que, inicialmente, a partir da adaptação de um VC de um domínio (1), uma base de conhecimento é criada. Tal base é gradativamente enriquecida pela inserção de AS relacionando OIs com termos do VC, os quais são selecionados durante a etapa de anotação dos objetos na catalogação (2). Ferramentas de visualização apresentam visões da base de conhecimento para apoiar a seleção dos termos a serem utilizados como valores de metadados na anotação e recuperação de objetos armazenados no repositório. As visualizações utilizadas (indicadas pelas letras A, B e C na Figura 15.) serão apresentadas na seção 5.1.2 (Figura 17, Figura 18 e Figura 19, respectivamente).

Quanto à recuperação dos objetos (3), a máquina de busca semântica processa as consultas dos usuários, expressas por coleções de palavras-chaves, por meio da busca dos termos correspondentes na base de conhe-

cimento e expansão semântica sobre o grafo representado as relações entre termos e o uso dos mesmos na anotação de objetos armazenados no repositório. Por fim, técnicas de visualização auxiliam o gerente do sistema a analisar o conteúdo do repositório (4), segundo as hierarquias de conceitos presentes na base de conhecimento. Ao gerente também cabe a tarefa de configurar a base de conhecimento, pela definição das coleções, tipos de conceitos e relações semânticas a serem utilizados na anotação e recuperação de OIs.

Os processos de **Adaptação inicial do conhecimento** (1) e **Catalogação** (2) são de interesse deste capítulo, além dos componentes de interface identificados pelas letras A, B e C na Figura 15. Os demais módulos serão vistos no próximo capítulo.

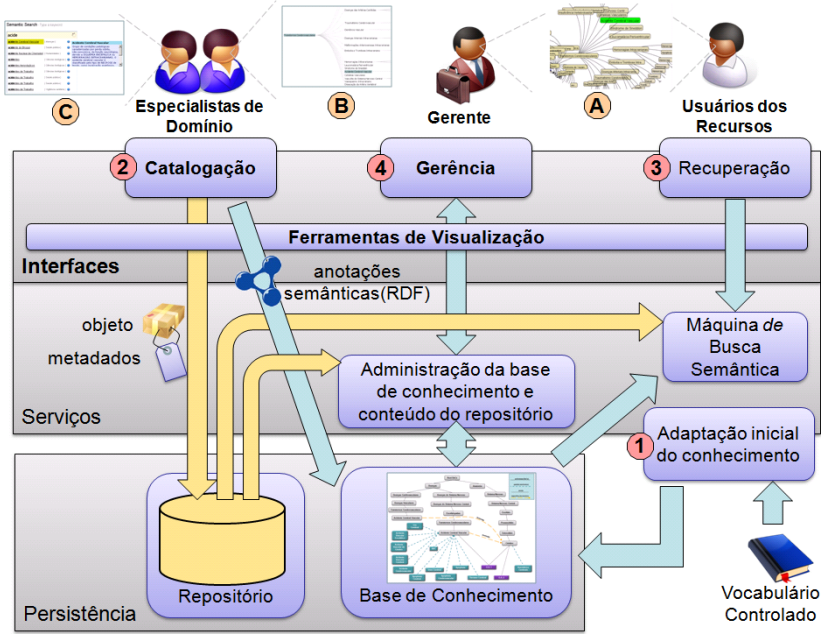


Figura 15: Estrutura do CIBELE.

5.1.1 Adaptação inicial do conhecimento

Levando-se em consideração os inconvenientes relacionados a anotação de OIs de forma livre, os quais foram previamente expostos na seção 1.2 e valendo-se da maturidade de certas áreas quanto à organização e classificação do conhecimento, esta pesquisa optou por conduzir a anotação de OIs via termos definidos em VCs. Isso elimina também problemas que a entrada de texto livre ocasiona, tais como erros de ortografia e interpretação. Sendo assim, a adaptação de conhecimento para uso na anotação e recuperação de OIs parte de VCs disponíveis no domínio. Porém nem todos os termos e relações de um VC disponível para um domínio são úteis a uma aplicação. Recortes temáticos podem ser feitos visando elencar somente porções de conhecimento convenientes para anotação e recuperação de informação. Estas porções devem ser arranjadas como conjuntos de termos parcialmente ordenados. Tal ordenação parcial é definida por relações semânticas, binárias, anti-simétricas (direcionadas) e transitivas entre termos.

Um exemplo de coleção de termos parcialmente ordenados na área de saúde é a hierarquia de classes de **doenças**, ligadas através de relações do tipo IS_A (classe-subclase), onde, por exemplo "Acidente Cerebral Vascular" IS_A "Transtornos Cerebrovasculares", que por sua vez IS_A "Doenças Vasculares". Um trecho da hierarquia de doenças adaptada do VC DeCS é ilustrado na porção superior esquerda da Figura 16.

Outro exemplo é a hierarquia de termos referentes à **anatomia**, ligados através de relações do tipo PART_OF (composição), onde, por exemplo, "Encéfalo" PART_OF "Sistema Nervoso Central", que por sua vez PART_OF "Sistema Nervoso". Um trecho da hierarquia referente à anatomia também adaptado do VC DeCS é ilustrado na porção superior direita da Figura 16.

As relações entre os termos usados para referenciar conceitos (e.g., Acidente Cerebral Vascular) e termos que podem ser utilizados como sinônimos para referenciar tais conceitos (e.g., Icto Cerebral, Acidente Vascular Encefálico, Acidente Vascular do Cérebro, Acidente Vascular Cerebral, Acidente Cerebrovascular, AVC, Apoplexia Cerebral, Ictus Cerebral, Apoplexia Cerebrovascular, Apoplexia, Derrame Cerebral) são representadas por linhas tracejadas na Figura 16.

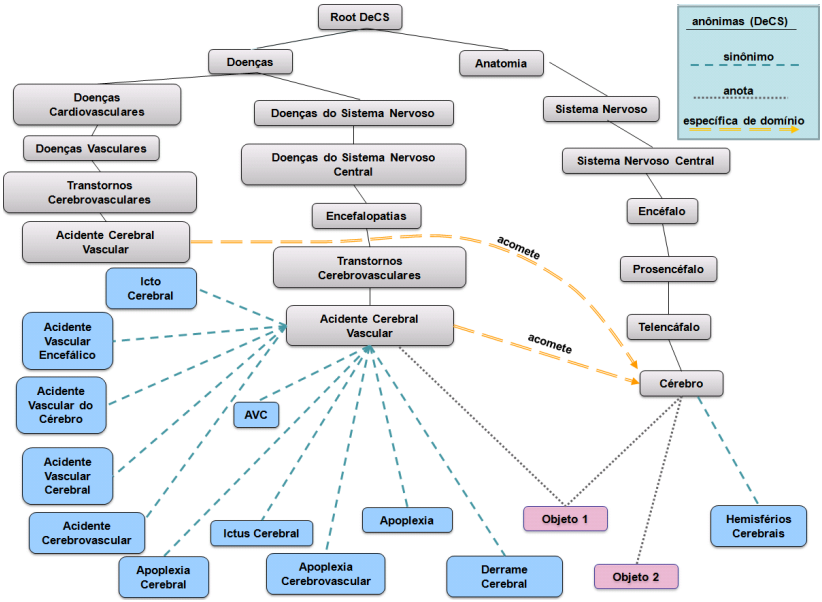


Figura 16: Base de conhecimento: termos ligados por relações semânticas.

5.1.2 *Catalogação: uso do conhecimento na anotação de objetos*

As técnicas de visualização facilitam a navegação em visões da base de conhecimento e a escolha dos termos relevantes do VC para a anotação dos objetos que estão sendo catalogados. A **anotação** de um OI usando um termo referente a um conceito ou instância da base de conhecimento para descrevê-lo implica no estabelecimento de uma ligação entre o objeto e a definição do termo na base de conhecimento. A Figura 16 ilustra, através de linhas pontilhadas, o uso dos termos "Acidente Cerebral Vascular" e "Cérebro" para anotar o "Objeto 1" e "Cérebro" para anotar o "Objeto 2".

A fim de se prover suporte eficiente à anotação de objetos de informação, este trabalho propõe amparar tais tarefas por meio de interfaces Web que permitem explorar interativamente conhecimento de domínio. Este é o primeiro passo para se definir um processo de anotação e recuperação de objetos amparados por técnicas da Web semântica (BERNERS-LEE; LASSILA, 2001).

Assim sendo, no Capítulo 4 este trabalho analisou técnicas e ferramentas de visualização de conhecimento. Tal embasamento foi necessário para alicerçar o desenvolvimento de componentes e interfaces Web para amparar usuários na escolha dos termos de VCs mais adequados para anotar OIs durante a catalogação de tais objetos em SRIs. Doravante estes componentes serão apresentados.

5.1.2.1 Interface Hiperbólica

A Figura 17 ilustra tal interface.

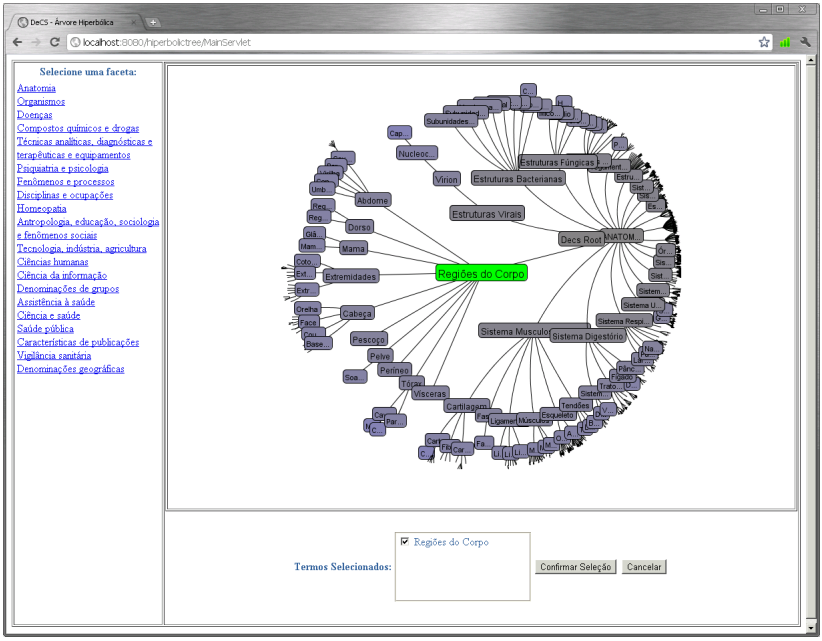


Figura 17: Interface Hiperbólica.

A Interface Hiperbólica é um componente que permite a execução de processos exploratórios através da navegação em uma visão do conhecimento de domínio, auxiliando assim o usuário a descobrir precisamente o que deseja. Ela exibe uma faceta (categoria) do VC por vez, já que devido ao tamanho total do mesmo, o VC não poderia ser exibido completamente sem perda de eficiência na navegação. A Interface Hiperbólica baseia-se principalmente nas funcionalidades providas pela biblioteca *Treebolic*, através de um applet[1].

As facetas podem ser selecionadas no canto superior esquerdo da interface. É possível se explorar a estrutura de conhecimento através do deslocamento do cursor do *mouse* sobre os nodos que representam os termos do VC. Como o foco é mantido no centro da interface, os termos que são exibidos nesta área recebem maior ênfase. Repousando o *mouse* sobre algum dos termos, informações adicionais são exibidas. Então o usuário, por meio de um duplo clique, seleciona os termos desejados, os quais são remetidos a caixa de "Termos Selecionados". Termos já selecionados não podem ser incluídos novamente na caixa de "Termos Selecionados".

Tal interface foi implementada como uma aplicação Web (Servlet[2]) e faz uso da linguagem Java[3] na camada lógica. Para a camada de apresentação utilizou-se JSP[4], HTML[5], CSS[6], além de JavaScript[7]. Maiores detalhes da implementação podem ser vistos no Anexo A.1.

5.1.2.2 Interface Hierárquica

É um componente que também permite a execução de processos exploratórios através da navegação em uma visão do conhecimento de domínio, porém de uma forma mais condensada e mais bem comportada. Faz uso das funcionalidades providas pela biblioteca *Prefuse* e implementa a visualização através de um applet. A Figura 18 ilustra tal interface.

[1] http://java.sun.com/applets
[2] http://www.oracle.com/technetwork/java/javaee/servlet/index.html
[3] http://www.oracle.com/technetwork/java/index.html
[4] http://www.oracle.com/technetwork/java/javaee/jsp/index.html
[5] http://www.w3.org/MarkUp
[6] http://www.w3.org/Style/CSS/
[7] http://www.w3schools.com/js/default.asp

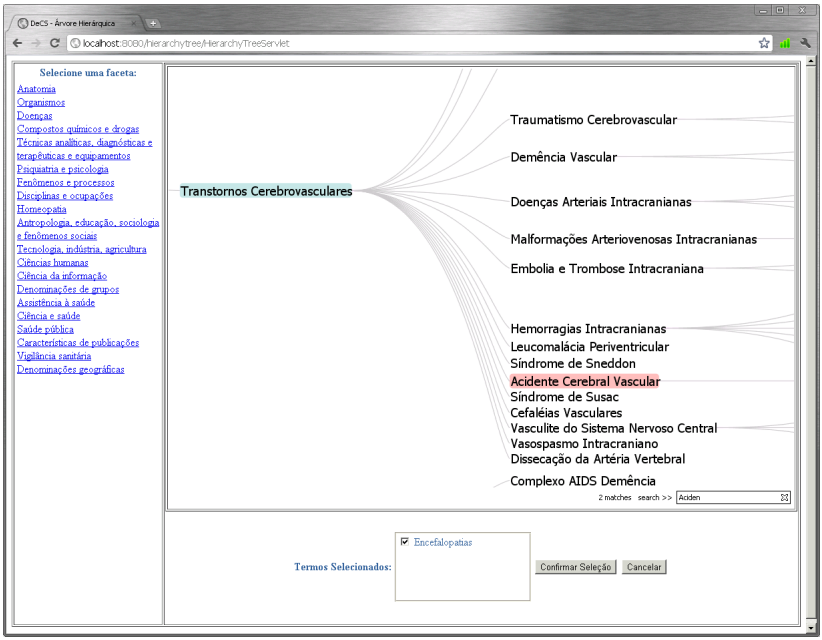


Figura 18: Interface Hierárquica.

As facetas podem ser selecionadas no canto superior esquerdo da interface. É possível se explorar a estrutura de conhecimento através de funções de *zoom* e deslocamento que são disparadas pelo botões do *mouse*. Um clique no botão direito do *mouse* faz o enquadramento da estrutura na tela. O *zoom in* e *zoom out* é conseguido via pressionamento e arrasto do botão direito do *mouse*. Um clique simples com o botão esquerdo do *mouse* sobre um nodo que representa um termo do VC faz com que o ramo que agrupa os termos situados abaixo do termo clicado seja aberto. Mantendo pressionado o botão esquerdo do *mouse* e movimentando o mesmo é possível se deslocar por sobre a estrutura. Então o usuário, por meio de um duplo clique com o botão esquerdo do *mouse*, pode selecionar os termos desejados, os quais são remetidos a caixa de "Termos Selecionados". Termos já selecionados não podem ser incluídos novamente na caixa de "Termos Selecionados".

A fim de facilitar a busca por termos na interface, ainda é disponibilizada uma caixa de busca. À medida que o usuário digita nela, os termos presentes na estrutura de conhecimento que possuem correspondência léxica

com os caracteres digitados na caixa de busca e que estão visíveis (situados em ramos abertos) são destacados (com fundo vermelho) na interface. A quantidade de resultados encontrados também é exibida, porém ela independe dos termos encontrados se situarem em ramos abertos, ou seja, todas as ocorrências são contabilizadas.

Tal componente é uma aplicação Web (Servlet) e faz uso da linguagem Java na camada lógica. Para a camada de apresentação utilizou-se JSP, HTML, CSS, além de JavaScript. Maiores detalhes da implementação podem ser vistos no Anexo A.2.

5.1.2.3 Interface Autocompletar

Além das técnicas de visualização hierárquica, quando o usuário já tem conhecimento sobre o domínio, uma interface capaz de completar e sugerir termos à medida que ele digita uma palavra-chave é útil para apoiar uma anotação ágil e precisa dos objetos durante a catalogação. A Figura 19 ilustra tal interface.

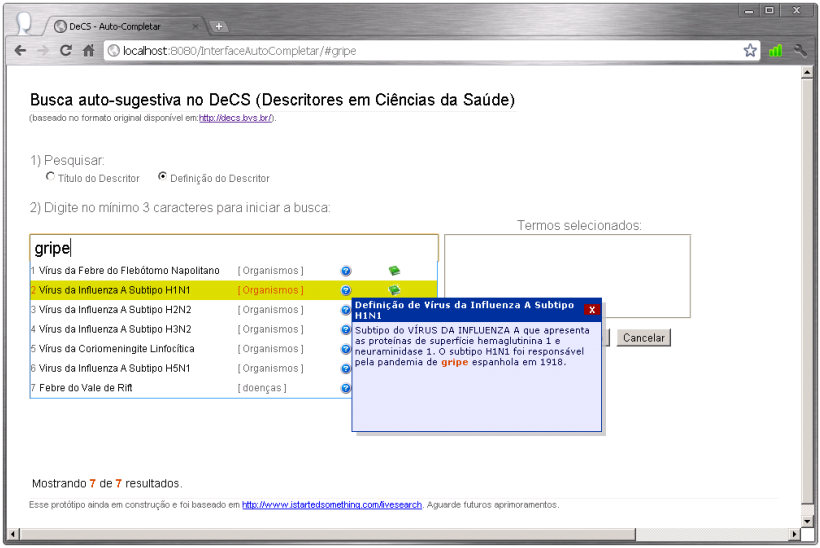


Figura 19: Interface de sugestão e autocomplemento.

Após o campo de entrada de texto conter pelo menos 2 caracteres, se inicia uma pesquisa na base de conhecimento a cada letra que o usuário

insere e, através de correspondência léxica *(matching)* com sinônimos ou palavras presentes na descrição do conceito, são sugeridos termos que denotam conceitos do domínio relacionados ao que está sendo digitado. A categoria do VC que abriga cada termo é mostrada entre colchetes. Já a definição de cada termo, segundo o VC, pode ser vista via posicionamento do *mouse* sobre o ícone . Eventuais sinônimos dos termos podem ser visualizados posicionando-se o *mouse* sobre o ícone . Isso contribui para a desambiguação e esclarece o contexto de cada termo.

Por meio de um clique, os termos selecionados pelo usuário são remetidos a caixa de "Termos Selecionados". Neste contexto, através de um *hiperlink*, o usuário pode acessar o VC (o DeCS, no caso) e então visualizar a definição do termo selecionado. Termos já selecionados não podem ser incluídos novamente na caixa de "Termos Selecionados".

Tal componente também foi implementado como uma aplicação Web (Servlet) e faz uso da linguagem Java na camada lógica. Para a camada de apresentação utilizou-se JSP, HTML, CSS, além de JavaScript e a técnica *Asynchronous Javascript and XML (AJAX)*[8] para prover comportamento dinâmico na apresentação dos resultados. As buscas são suportadas pelo Apache Lucene [9], que durante a inicialização da aplicação Web, a qual pode rodar num container de servlets como o Apache Tomcat[10], indexa arquivos contendo um extrato de cada termo presente no VC. Assim compõe um índice que é alvo das buscas disparadas pela Interface Autocompletar. Maiores detalhes da implementação podem ser vistos no Anexo A.3.

5.1.2.4 Gerência de conhecimento e conteúdo

Apresentamos aqui uma proposta para gerência de OIs baseada em visualizações do conhecimento. Tal módulo é identificado pelo item 4 da Figura 15, a qual ilustra a arquitetura do CIBELE.

Muitas vezes é importante saber a quantidade de OIs disponíveis nos SRI para cada assunto específico do domínio, principalmente à medida que aumenta o número de objetos disponíveis no repositório, que é alimentado dinâmica e colaborativamente. Tal informação pode embasar decisões sobre investimentos na produção de OIs sobre determinados temas. Porém as tradicionais interfaces de gerenciamento dos SRIs são meramente hipertextuais e não exploram e muito menos enfatizam as relações estabelecidas entre

[8] http://www.w3schools.com/ajax/default.asp
[9] http://lucene.apache.org
[10] http://tomcat.apache.org

os valores de metadados que descrevem os objetos, comprometendo assim a identificação dos temas que melhor representam o conteúdo gerido (DUPRIEZ; SCHUBNEL, 2009).

Este trabalho propõe a visualização desse tipo de informação de maneira sintética sobre hierarquias de termos da base de conhecimento, como ilustrado na Figura 20. Os termos mais usados na anotação de objetos catalogados no repositório são destacados dos demais pelo tamanho, pelo tom da cor e pelo seu rótulo, que exibe o número de objetos do repositório anotados com aquele termo. Na Figura 20, o termo "Cérebro" é o mais destacado, por ser usado para anotar 3 objetos, enquanto o termo "AVC" é usado para anotar 2 objetos.

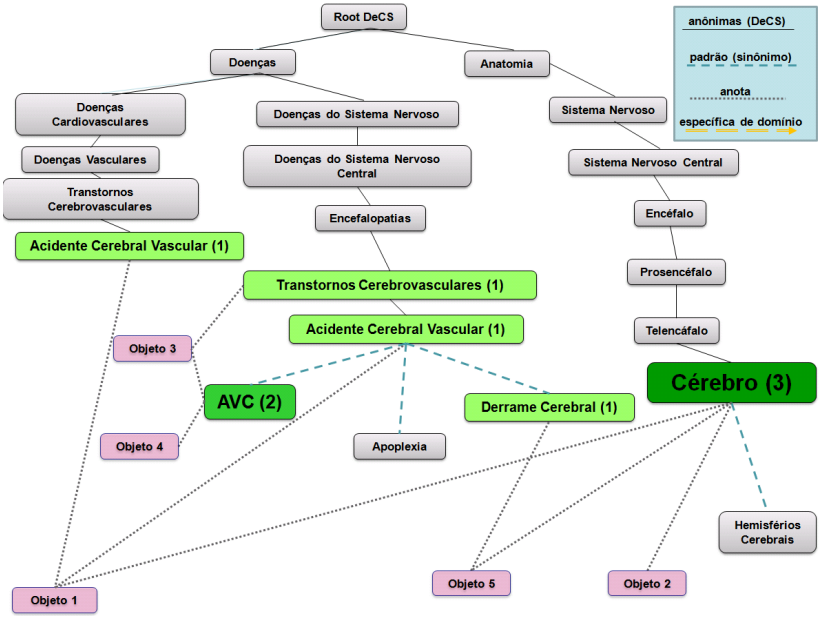


Figura 20: Proposta para gerência de conteúdo baseada em conhecimento.

Porém a implementação de tal modelo com ferramentas de visualização demonstrou que a legibilidade e a navegabilidade no conteúdo do SRI fica comprometida, segundo avaliação empírica de alguns colaboradores deste trabalho. A Figura 21 mostra a implementação do modelo via *Prefuse*.

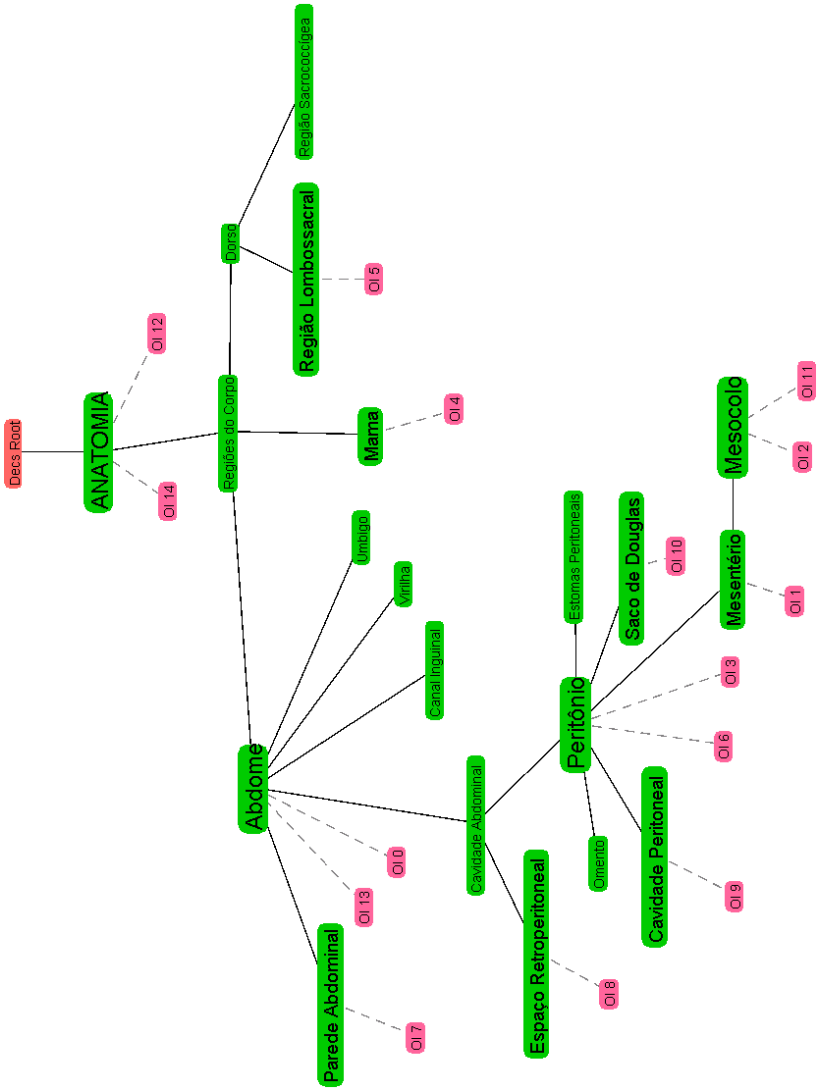


Figura 21: Implementação preterida para gerência de conteúdo baseada em conhecimento.

O problema da legibilidade e da navegabilidade é agravado a medida que aumenta a quantidade de OIs geridos. Sendo assim, a implementação de tal modelo foi alterada. A nova proposta foi implementada com a ferramenta *Prefuse* e assemelha-se a uma nuvem de *tags*. O tamanho dos nodos que representam os termos são destacados conforme sua relevância para o domínio (quantidade de anotações que promovem sobre os OIs), da mesma forma que na proposta anterior. Porém a questão da navegabilidade foi resolvida via árvore hierárquica, onde as relações semânticas agora guiam o gestor de conteúdo, como mostra a Figura 22. Nela, percebe-se que a quantidade de objetos que cada termo anota é exibida junto ao rótulo dos nodos. Por exemplo, o termo "Regiões do Corpo" anota 24 objetos, enquanto que termos mais específicos que "Regiões do Corpo", segundo a hierarquia do VC DeCS, anotam 220 objetos de um SRI hipotético.

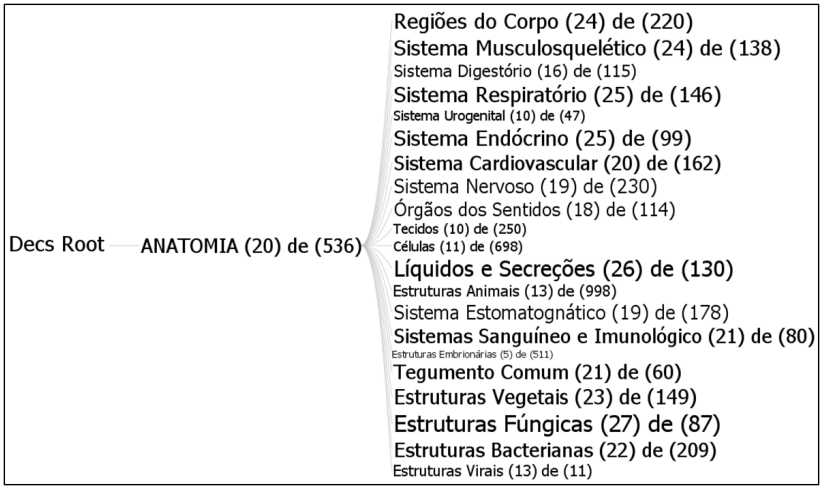


Figura 22: Gerência de conteúdo catalogado baseada em conhecimento.

Salientamos que a interface para Gerência de conhecimento e conteúdo ainda não figura no SRI do estudo de caso, que será abordado no Capítulo 7. Testes de usabilidade com usuários gestores devem ser conduzidos antes de migrarmos nossa solução para um ambiente de produção. Trabalhos futuros tratarão disso.

Neste capítulo apresentamos a estrutura do CIBELE, nosso sistema que opera buscas associativas sobre representações do conhecimento e

anotações. Aqui exploramos o módulo de **Adaptação inicial do conhecimento** e o módulo de **Catalogação**, o qual utiliza conhecimento de domínio na anotação de objetos. Também descrevemos os 3 artefatos de software para amparar a anotação de OIs durante a catalogação de tais objetos em SRIs. Por fim, apresentamos uma proposta para **Gerência de conhecimento e conteúdo**.

6 RECUPERAÇÃO SEMÂNTICA

Este capítulo detalha nossa proposta para a recuperação semântica de OIs via buscas associativas sobre representações do conhecimento e anotações.

Retomando a explicação em torno da arquitetura exposta na Figura 15, este capítulo explora o ítem 3, ou seja, o módulo de **Recuperação**. Ele se apóia em uma base de conhecimento que pode ser vista como um grafo cujos nodos representam termos de algum VC ou objetos armazenados no repositório e cujas arestas representam as relações entre eles, como mostra a estrutura apresentada na Figura 16. No contexto do CIBELE, formalmente tal estrutura é definida como uma Rede Semântica (RS):

Definição 8 ***- Rede Semântica (RS):***

Seja OIS o conjunto que agrupa todos os objetos de informação (veja Definição 1).

O grafo que representa a Rede Semântica é denotado por $RS(Nodos, Arestas)$*, onde:*

cada $ns \in Nodos$ *é um:*

(i) ***termo*** $t \in VC$ *(veja Definição 2) ou*

(ii) ***objeto de informação*** $oi_i \in OIS$

cada $as(t_i, t_j) \in Arestas$ *é uma:*

(i) ***relação semântica*** *de* t_i *para* t_j*, onde* t_i *e* $t_j \in VC$ *ou*

(ii) ***anotação semântica*** *onde* $t_i \in VC$ *e* $t_j = oi_j$*, sendo que* $oi_j \in OIS$*.*

Tal estrutura permite expandir semanticamente as buscas, a partir dos termos usados como **palavras-chave** na especificação das consultas, valendo-se da técnica de *Spreading Activation (SA)*. Consultas que exploram outros valores de campos de metadados como autor, data, título não são suportadas

por esta abordagem.

6.1 Implementação da RS

A RS é construída a partir de um vocabulário controlado (VC). No caso do presente trabalho, ela é baseada no VC DeCS. Os nodos e a estrutura RS se baseiam, respectivamente, nos termos e nas relações hierárquicas que o VC define para os termos, as quais são representadas por linhas contínuas na Figura 16. Informações como a descrição e os sinônimos dos termos também foram levadas em conta para a construção da RS, ou seja, além de espelhar a estrutura hierárquica dos termos, a RS é enriquecida com nodos que representam os sinônimos que o VC define para certos termos. Assim buscas dirigidas pelos sinônimos dos termos presentes na RS também são suportadas. Por exemplo, um OI que foi anotado com o termo "Acidente Vascular Cerebral" poderia ser recuperado a partir de consultas dispararas via sinônimos deste termo: "AVC", "Derrame Cerebral", "Apoplexia", "Ictus Cerebral", entre outros.

A RS é construída em RDF e suporta consultas via RDQL (linguagem de consulta para RDF). No Anexo B, os detalhes do processo da geração da RS são descritos.

6.1.1 Atualização da RS

À medida que a catalogação de objetos é realizada no SRI, anotações semânticas em RDF são geradas. Tais anotações são armazenadas em um banco de dados do módulo de recuperação, ficando disponíveis para a atualização da RS. Tal atualização pode ser *on-line*, ou seja, assim que um novo objeto é catalogado no SRI, sua representação já passa a figurar na RS; ou a qualquer tempo, bastando o gerente invocar tal tarefa ou definir uma política de atualização da RS: uma vez por dia, por exemplo. A segunda abordagem é menos custosa já que a carga da RS e das diversas anotações realizadas seria feita de uma só vez, consumindo menos recursos da máquina que hospeda o módulo de recuperação do CIBELE.

Durante a atualização, a RS e as anotações semânticas persistidas no banco de dados do módulo de recuperação são mescladas via *framework* semântico Jena[1]. Então a RS também passa a considerar os novos objetos catalogados e devidamente anotados com termos do VC através da adição de

[1] http://jena.sourceforge.net

novos nodos. As anotações semânticas realizadas sobre os novos OIs catalogados no SRI promovem a adição de arestas ligando os termos da RS aos novos nodos que representam tais OIs. Estas anotações são representadas por linhas pontilhadas, como mostra a Figura 16.

6.2 Implementação do algoritmo de SA

O algoritmo de SA implementado utiliza a biblioteca da Texai[2]. Ela fornece uma API para a construção programática de uma SAN, além disso disponibiliza métodos de processamento que suportam os diversos parâmetros que regem a execução do SA:

- **Máximo nodos (N)**: o número máximo de nodos que podem ser disparados pelo SA;
- **Máximo pulsos (P)**: o número máximo de pulsos que o algoritmo SA deve executar;
- **Máximo nodos/pulso (NP)**: o número máximo de nodos que podem ser disparados por pulso;
- **Fator de decaída** ($D \in [0,1]$): o fator de decaimento aplicado quando o SA se propaga através de uma aresta;
- **Limiar de disparo** ($L \in [0,1]$): valor de ativação mínimo para o disparo de um nodo (*firing threshold*);
- **Sementes** ($K \subset T$): o conjunto de palavras-chave de T fornecidos na consulta do usuário;
- **Ativação das Sementes** ($A \in [0,1]$): o valor de ativação inicial atribuído aos $k \in K$;
- **Peso das Relações** ($W_{ij} \in [0,1]$): o peso da aresta $RS(n_i, n_j)$, atribuídos de acordo com seu tipo;

Como o algoritmo SA fornecido pela biblioteca Texai não pode executar diretamente sobre a RS em RDF, uma SAN é gerada programaticamente a partir da RS quando o módulo de recuperação é inicializado. Na SAN, todos os nodos, relações e pesos definidas pela RS são mantidos. Porém uma

[2] http://sourceforge.net/projects/texai/files

adequação quanto à direção das arestas é feita: as arestas unidirecionais provenientes da RS tornam-se simétricas na SAN, possibilitando que as ondas de ativação se propaguem tanto generalizando quanto especializando as buscas. Isso permite que mesmo que o SA seja disparado a partir de termos folha da RS (sem filhos), as ondas de ativação se propaguem em direção aos pais de tais termos.

6.2.1 Saída do SA

Assim que alguma condição de término é alcançada (não existem mais nodos para disparo, ou N ou P atingem os valores estabelecidos), o algoritmo de SA cessa o espalhamento da ativação sobre a RS. Então o nível de ativação alcançado pelos nodos é usado para computar a relevância de cada nodo frente à consulta formulada. O nível de ativação de cada nodo $\in [0,1]$. Formalmente, o subrafo RS' resultante do processamento do SA sobre uma RS pode ser assim descrito:

Definição 9 - ***Saída SA:***

Sejam:
Nodos e Arestas os conjuntos da Definição 8.

$Nodos' \subset Nodos$.
$Arestas' \subset Arestas$.
$A(ns')$ *o valor de ativação do nodo ns' calculado pelo SA via 3.1.*
L é o Limiar de disparo definido em 6.2.

O subgrafo que representa a saída do SA é denotado por $RS'(Nodos', Arestas')$, *onde:*

$ns' \in Nodos'$ *e* $A(ns') > L$.
$as' \in Arestas'$ *e* $as'(ns'_i, ns'_j)$, *sendo que* ns'_i *e* $ns'_j \in Nodos'$.

Os nodos ns' ativados podem ser ordenados de forma decrescente de forma que os nodos com maior nível de ativação apareçam no topo da listagem. Daqui para frente o valor de ativação de cada ns' será chamado de **Relevância semântica**. Porém, dependendo da aplicação, nem todos os nodos que aparecem na listagem são de interesse do usuário. Em aplicações de RI, como é o caso deste trabalho, é possível que apenas aqueles que denotam OIs sejam considerados, ou seja, aqueles $ns' \in OIS$.

6.3 Melhorias na abordagem proposta

A topologia que a RS assume e por conseqüência a SAN, está diretamente relacionada às característica do grafo do VC que a originou: ciclos, termos ligados a um grande número de outros termos (alta conectividade), grande profundidade ou largura dos ramos do grafo, etc. Testes empíricos com uma RS originada a partir de porções do VC DeCS mostraram que tais características podem tornar a aplicação do algoritmo de SA puro ineficiente, exigindo alguns refinamentos de modo que a busca espalhe-se de uma maneira mais adequada, ou seja, de encontro às peculiaridades da RS. Esta percepção motivou melhorias na estrutura de conhecimento e no algoritmo de processamento.

6.3.1 Refinamentos no modelo SA

6.3.1.1 Refinamentos na RS

Como a recuperação associativa é baseada nas relações estabelecidas entre os nodos da estrutura de conhecimento, é possível se diferenciar através de pesos atribuídos às relações o impacto de cada tipo de relação durante o processo de busca. Neste trabalho, as relações usadas, as quais são dependentes do VC empregado, e os respectivos pesos são exibidos na Tabela 1:

Tipo de Relação	Peso	Característica
Anônima	0,9	relação hierárquica definida pelo VC que ocorre entre dois nodos $\in T$.
Sinônimo	0,9	relação de sinonímia definida pelo VC e que ocorre entre dois nodos $\in T$.
Anotação	1,0	é uma aresta (t_i, oi_j) onde $t_i \in T$ e $oi_j \in OIS$.
Domínio	1,0	relação entre dois nodos $\in T$ estabelecida por especialistas de domínio.

Tabela 1: Relações semânticas e pesos.

Os valores de tais pesos foram estabelecidos em virtude de diversos testes empíricos realizados previamente. Também considerou-se que as relações oriundas do VC (Anônimas e Sinônimos) não são completamente confiáveis, ou seja, pode existir um "ruído semântico", o que justifica o valor 0,9 ao invés de 1,0. Já as relações de Anotação e de Domínio são explicitamente estabelecidas, respectivamente, por usuários e especialistas, o que

justifica uma acertividade de 100%, ou seja, o peso atribuído é de 1,0.

Tais relações também são podem ser tratadas de maneira específica, conforme a direção das mesmas (CRESTANI, 1997). A partir das sementes é possível saber se a propagação das ondas de ativação esta generalizando ou especializando a consulta. Sendo assim, para contemplar tal regra, um fator adicional foi incorporado às arestas. Ele é mostrado na Tabela 2:

Sentido da Relação (S)	**Peso**
Especialização	1,0
Generalização	0,8

Tabela 2: Direção das relações semânticas e pesos.

A diferenciação nos pesos é amparada por (SAVOY, 1992), que defende que na primeira onda de propagação, o SA deve percorrer as arestas em qualquer direção, enquanto que nas demais ondas, somente arestas na direção de especialização devem ser seguidas. Além disso a intuição diz que é melhor concentrar as buscas por OIs em regiões mais específicas da RS do que em regiões mais generalistas, onde os termos denotam conceitos mais gerais como Anatomia, Doenças e Organismos, por exemplo.

6.3.1.2 Refinamentos na técnica de processamento

Em vista dos aprimoramentos discutidos na seção anterior, o valor de ativação de um nodo é calculado via:

$$A_k(p) = A_k(p-1) + \sum(O_j(p)).(W_{jk}.(S)).(D) \tag{6.1}$$

onde:

$A_k(p)$ é a ativação do nodo k no pulso p
$A_k(p-1)$ é a ativação do nodo k no pulso $(p-1)$
$O_j(p)$ é a saída (*output*) do nodo j conectado ao nodo k no pulso p
W_{jk} é o peso da aresta que conecta o nodo j ao nodo k
$D \in [0,1]$ é o fator de decaída
S é o peso atribuído de acordo com o sentido da aresta que conecta o nodo j ao nodo k

Note que S aqui é novidade em termos da equação já apresentada em 3.1.

6.3.2 Proposta de enriquecimento da RS

Alguns testes empíricos com a RS gerada a partir do VC DeCS mostraram que para determinadas consultas seria praticamente impossível obter-se bons resultados. Na Figura 23 uma RS hipotética ajuda a ilustrar este problema. Suponha que um usuário tenha interesse em recuperar objetos que versem sobre "doenças que acometem o coração". Então ele traduz sua necessidade de informação no seguinte conjunto de palavras-chave: "Doenças" e "Coração". Embora exista um termo mais adequado para expressar esta consulta, "Doenças Cardiovasculares", digamos que o usuário o desconheça. Como o termo "Doenças" pertence à categoria (ou faceta) **Doenças** e o termo "Coração" pertence à categoria **Anatomia** do VC, o espalhamento das ondas de ativação atua localmente na RS (apenas nos ramos oriundos de cada uma das categorias do VC) e não é suficiente para propagar-se pelos termos que melhor exprimem a intenção do usuário, como mostra a Figura 23.

A partir desta constatação, veio à tona a necessidade de se construir "pontes" entre os ramos da RS. Porém estas "pontes" não poderiam ser enxertadas a esmo. Deveria haver algum critério, possivelmente baseado em algum tipo de conhecimento. Logo visualizou-se a oportunidade de se explorar outras características do VC de modo a se obter tais "pontes" e então enriquecer a RS com elas. Tais "pontes" podem ser obtidas a partir do cálculo da similaridade sintática entre as descrições dos termos do VC.

Um exemplo de "ponte" é mostrado na Figura 24, com uma linha largamente tracejada. Ela estabelece uma nova relação semântica entre o termo "Coração" e "Cardiopatias.". Como "Coração" é uma das sementes da busca, o termo "Cardiopatias" seria atingido pelo segundo pulso do algoritmo SA. A abrangência das ondas de ativação agora é maior, possibilitando que o OI 3, que está anotado com um termo que denota um sinistro que ocorre no coração (Aneurisma Cardíaco), seja recuperado.

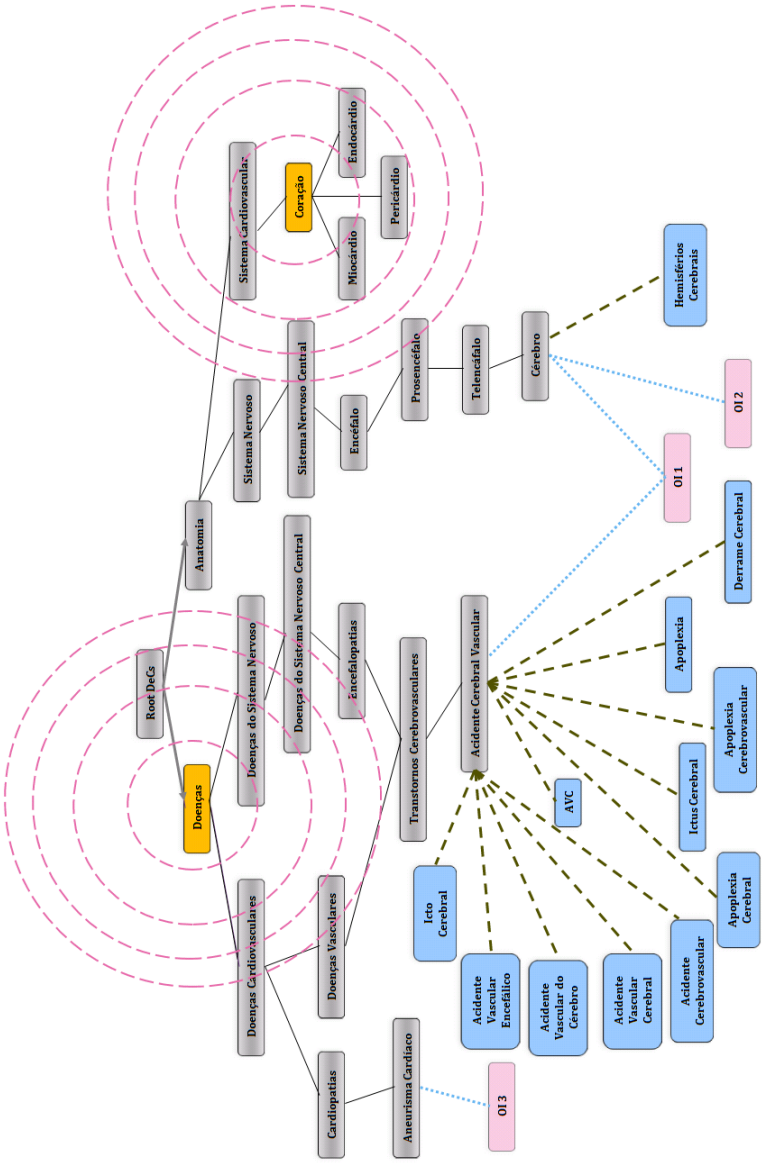


Figura 23: Modelo SA não atende satisfatoriamente determinadas consultas.

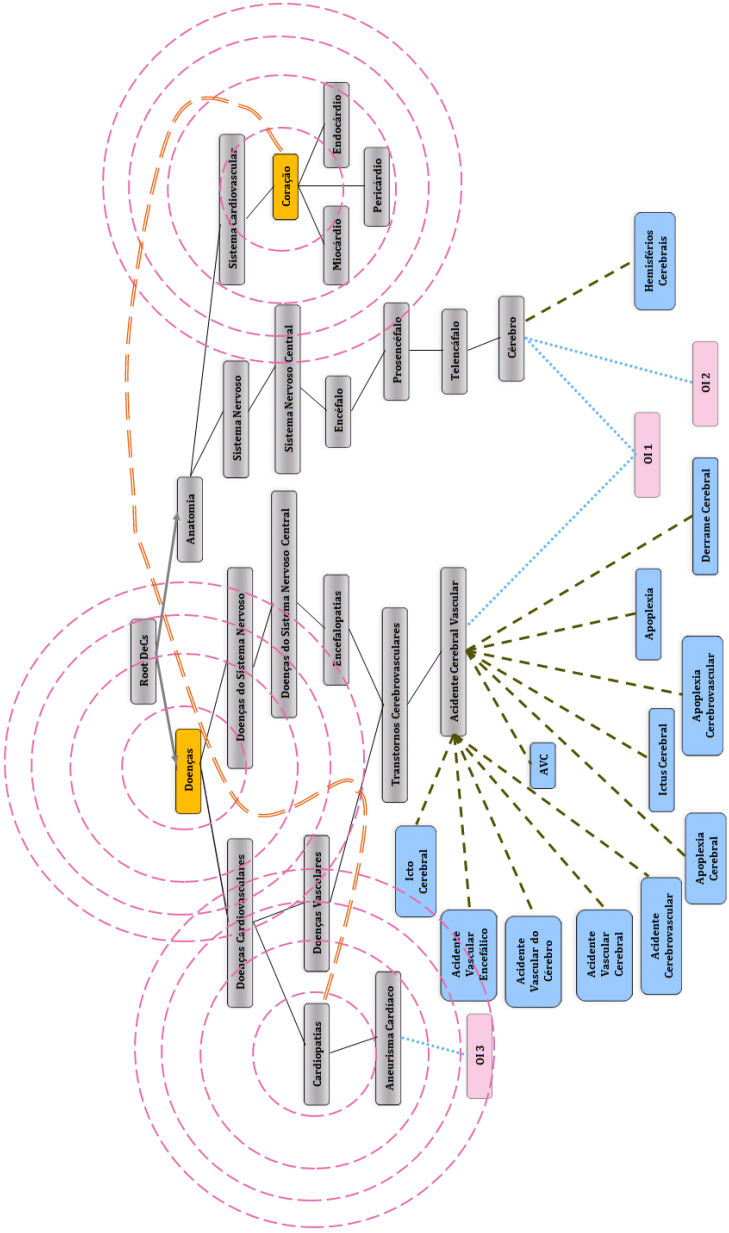


Figura 24: Enriquecimento da RS via nova relação semântica.

A título de exemplo, a partir do cálculo da similaridade sintática entre as descrições dos termos das categorias “Anatomia” e “Doenças” do VC DeCS foi possível se construir a tabela mostrada na Figura 25:

	A	B	C	D
1		Anatomia	Doenças	Similaridade
2		Divertículo Ileal	Divertículo Ileal	1
3		Células Oxífilas	Adenoma Oxífilo	0,67612344
4		Sistema Biliar	Colecistolitíase	0,6123724
5		Colo Descendente	Doenças do Colo Sigmóide	0,5714286
6		Periodonto	Doenças Periodontais	0,5714286
7		Conteúdo Gastrointestinal	Gastroenteropatias	0,54772252
8		Sistema Biliar	Neoplasias do Sistema Biliar	0,53033012
9		Ácido Gástrico	Acloridria	0,51639777
10		Conteúdo Gastrointestinal	Hemorragia Gastrointestinal	0,50709254
11		Sistema Biliar	Neoplasias da Vesícula Biliar	0,5
12		Ureter	Pielocistite	0,49613893
13		Ureter	Pielocistite	0,49613893
14		Braço	Traumatismos do Antebraço	0,46291006
15		Colo Descendente	Doenças do Colo	0,46291006
16		Trato Gastrointestinal Superior	Gastroenteropatias	0,46291006
17		Esfíncter da Ampola Hepatopancreática	Neoplasias do Ducto Colédoco	0,46225014
18		Leucócitos	Agranulocitose	0,45291084
19		Leucócitos	Agranulocitose	0,45291084
20		Sistema Biliar	Adenoma de Ducto Biliar	0,44721359
21		Colo Sigmóide	Proctocolite	0,44721359
22		Ápice Dentário	Fraturas dos Dentes	0,44721359
23		Canal Arterial	Permeabilidade do Canal Arterial	0,44095856
24		Dente Canino	Erosão Dentária	0,42905816
25		Trato Gastrointestinal Superior	Hemorragia Gastrointestinal	0,42857143

Figura 25: Similaridade: Anatomia x Doenças.

Ela relaciona termos da categoria “Anatomia” e termos da categoria “Doenças” através de relações que bem poderiam ser chamadas de “acometido por”. Por exemplo, o “Colo Sigmóide” é acometido por “Proctocolite”. Este esboço mostra o potencial da proposta de se usar conhecimento já estabelecido para a geração de novos conhecimentos, possibilitando a inferência de novas relações semânticas úteis para ligar ramos distintos da RS. Tais relações são agrupadas em um modelo RDF que pode ser usado para enriquecer a RS, como mostra a Figura 26:

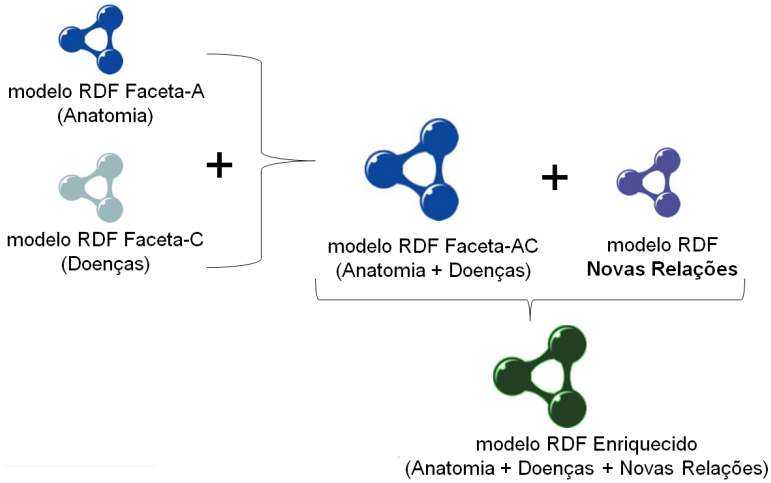


Figura 26: Enriquecimento da RS.

Este tipo de enriquecimento requer supervisão de especialistas de domínio para aferir a qualidade das relações sugeridas. Somente se elas forem julgadas coerentes é que devem ser incorporadas à estrutura de conhecimento. Por limitações de tempo e de recursos, não foi possível contar com o apoio de tais especialistas. Logo tal proposta será explorada em trabalhos futuros. Atenção especial deverá ser dada ao aumento da conectividade do nodos, já que novas relações implicam na adição de novas arestas ao grafo da RS. Como foi visto na seção 3.2.2, nodos com alta conectividade espalham a ativação sobre grandes áreas da RS, podendo comprometer a performance de execução do SA e a qualidade dos resultados.

Detalhes da implementação do cálculo de similaridade entre termos

do VC podem ser vistos no Anexo C.

6.3.3 Composição de relevâncias

Em um SRI, além de consultas baseadas em campos de metadados como assunto ou palavra-chave, é usual que as consultas também sejam dirigidas por campos de metadados como autor, data, título, etc. Assim o usuário poderia recuperar objetos que versem sobre um certo tema denotado pelas palavras-chave mas que são de um determinado autor. Visando suportar tais consultas, o processo de busca pode ser desdobrado em uma busca sintática e uma busca semântica e no final do processo um único *ranking* deve ser montado. Tal *ranking* é composto pelos:

- resultados da busca sintática realizada pelo motor de busca tradicional do SRI, o qual já opera sobre os campos de metadados em questão (autor, data, título,etc.);
- resultados da busca semântica via palavras-chave que é proposto por este trabalho.

Sabe-se que muitos SRI como o DSpace fazem uso do Lucene para operar a recuperação sintática de objetos. O Lucene gera um *ranking* que afere a cada objeto um grau de relevância frente a consulta formulada: um valor 1.0 significa que o objeto recuperado atende exatamente à consulta formulada (*matching* exato), enquanto que valores menores atestam que o objeto recuperado atende parcialmente à consulta ou, no caso de não atender, o valor de relevância atribuído é 0. Daqui para frente este valor de relevância será chamado de **Relevância sintática**.

Em vista disso, o cálculo que define a relevância resultante $R(oi_i)$ de um OI oi_i recuperado por este método híbrido de consulta valendo-se da relevância sintática $R_{Sintática}(oi_i)$ e da relevância semântica $R_{Semântica}(oi_i)$, ambas $\in [0,1]$, pode ser definido como:

Definição 10 ***- Cálculo da composição de relevâncias:***

Seja F o fator de relevância tal que $(0 <= F <= 1)$.

$$R(oi_i) = (F * R_{Semântica}(oi_i)) + ((1-F) * R_{Sintática}(oi_i))$$

Note que o resultado de cada tipo de busca pode ser enfatizado ou atenuado pelo fator de relevância F: um alto valor de F enfatiza os resultados

obtidos pela busca semântica em detrimento aos resultados da busca sintática. Já um baixo valor de F tem o efeito inverso.

Neste capítulo exploramos o módulo de Recuperação do CIBELE. Ele opera recuperação semântica de OIs via buscas associativas sobre representações do conhecimento e anotações. Para suportar tal abordagem, definimos uma RS que modela as associações entre o conhecimento e os OIs anotados na catalogação. Também discorremos sobre a implementação do algoritmo de SA, que atua sobre o conhecimento modelado pela RS. Por fim apresentamos algumas melhorias implementadas: refinamentos no SA, na RS, suporte a buscas híbridas (sintáticas e semânticas) via composição de *rankings* e uma proposta de enriquecimento da RS via análise sintática da descrição dos termos que compõe o VC que a origina.

7 ESTUDO DE CASO

Este capítulo apresenta o ambiente real que abriga muitas das contribuições geradas por este trabalho. Tal ambiente testa nossas contribuições e vale-se delas para fomentar o ensino a distância.

7.1 A UnA-SUS

A análise das propostas apresentadas neste trabalho dá-se no contexto do programa Universidade Aberta do SUS (UnA-SUS). Ele visa criar condições para o funcionamento de uma rede colaborativa de instituições acadêmicas, serviços de saúde e gestão do Sistema Único de Saúde (SUS) destinada a atender as necessidades de formação e educação permanente dos profissionais do SUS através de ações focadas em *e-learning*, intercâmbio de experiências, compartilhamento de material instrucional, cooperação para desenvolvimento e implementação de novas tecnologias educacionais em saúde [1]. Para tanto, foi elaborado um Ambiente Virtual de Ensino e Aprendizagem (AVEA). Ele é um facilitador das tarefas de ensino e aprendizagem e provê acesso a Objetos de Aprendizagem.

7.2 Objetos de aprendizagem - OA

Um Objeto de Aprendizagem (OA) é qualquer material que pode ser usado no processo de ensino e aprendizado (HODGINS, 2002). Quando em formato digital um OA pode ser organizado na forma de material educativo a ser usado na formulação de cursos suportados por tecnologia ou Ensino a Distância (EaD). Cursos baseados em OAs e amparados pela Web facilitam a disseminação do conhecimento e a qualificação de profissionais, segundo um processo de ensino/aprendizagem em que o professor e o aluno estão geograficamente distantes e em que a interação entre eles é preferencialmente estabelecida via meios eletrônicos (GONçALVES, 2007). OAs podem ser descritos usando o padrão de metadados *Learning Objects Metadata (LOM)* (HODGINS, 2002) e depois armazenados em SRI.

[1] http://portal.universidadeabertadosus.org.br/?q=node/1

7.3 O AVEA UnA-SUS - UFSC

O AVEA UnA-SUS UFSC se assenta sobre a plataforma Web e é acessível via qualquer navegador Web. Ele é composto de dois sistemas de domínio público: um SRI e um Sistemas de Gestão da Aprendizagem (SGA) (do inglês *Learning Management System* (LMS)), que podem ser acessados por diversos tipos de usuários, como ilustrado na Figura 27.

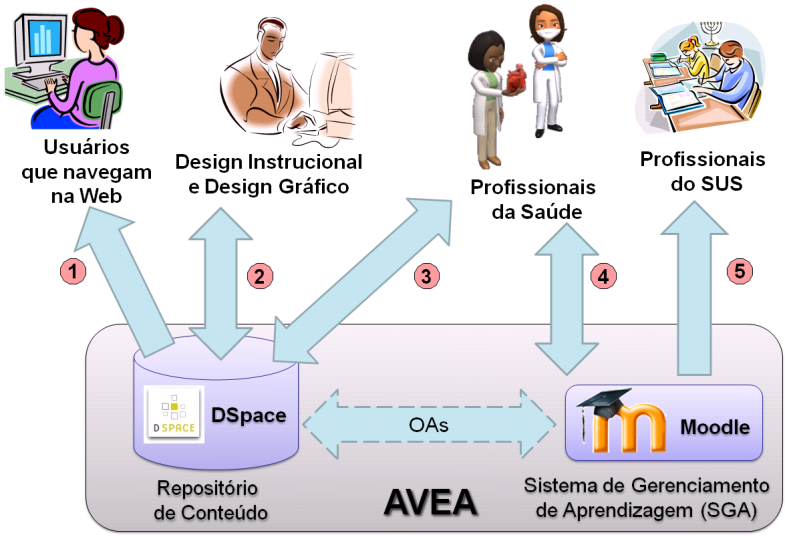


Figura 27: Tipos de usuários do AVEA UnA-SUS - UFSC.

O repositório de conteúdo UnA-SUS UFSC está disponível via Web[2] e pode ser consultado por qualquer pessoa (1). Porém, a inserção de OAs é restrita a especialistas de design (2) e da área de saúde (3), sendo os últimos responsáveis pela montagem e gerenciamento de cursos no SGA (4), os quais estão atualmente disponíveis somente para profissionais do SUS (5).

O repositório de conteúdo mantém os objetos informacionais (OAs, no caso da UnA-SUS) devidamente descritos com metadados para permitir a sua recuperação. O DSpace [3] (TANSLEY et al., 2003) foi a base para a nossa implementação do repositório. Ele foi personalizado e enriquecido com fun-

[2] http://repositorio.unasus.ufsc.br
[3] http://dspace.org

cionalidades da Web 2 (Web Social) e nossas interfaces gráficas baseadas em conhecimento para apoiar a anotação e a recuperação de OAs. O SGA usado é o Moodle[4] (SILVA, 2010), que disponibiliza cursos previamente elaborados a partir de OAs, possibilitando a aprendizagem colaborativa à distância valendo-se dos OAs que podem ser obtidos do repositório e assim reusados em diversos cursos. A geração de tais conteúdos envolve o trabalho de conteudistas e designers, e obedece ao fluxo ilustrado na Figura 28:

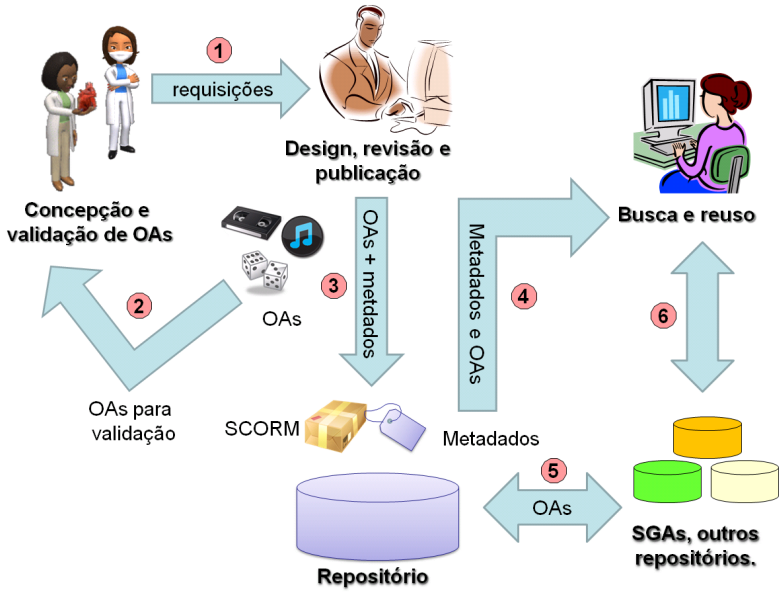


Figura 28: Fluxo de produção, armazenamento e uso de OAs.

Primeiramente os conteudistas, que são profissionais da saúde e tem o papel de conceber OAs, requisitam à equipe de Design Instrucional (DI) e Design Gráfico (DG) a elaboração de OAs sob determinados temas (1). Estes objetos então são produzidos pela equipe de DI donde ferramentas como o Adobe Flash[5] são utilizadas nesse processo de desenvolvimento. Findada a etapa de desenvolvimento dos objetos, eles passam pela avaliação da equipe de conteudistas (2). Se tais objetos forem aprovados, os mesmos são

[4]http://www.moodle.org/
[5]http://www.adobe.com/products/flash

empacotados segundo o padrão *Sharable Content Object Reference Model* (SCORM)[6]. Posteriormente, na catalogação, os OAs são descritos, anotados com termos do VC DeCS e persistidos no repositório (3), ficando à disposição para serem usados em processos de aprendizagem (5), composição de novos objetos e reuso (4) ou intercâmbio (6) com outros repositórios ou SGAs. Em caso de não aprovação, os OAS voltam à equipe de DG, onde sofrem as alterações necessárias (2).

7.4 Repositório de conteúdo DSpace

O DSpace (TANSLEY et al., 2003) é um repositório digital desenvolvido pelo *Massachusetts Institute of Technology (MIT)* e Hewlett-Packard *(HP)* e objetiva armazenar, gerir e disseminar conteúdo digital. É disponibilizado livremente sob a forma de um produto de código aberto desenvolvido em Java. Pode ser livremente adaptado e expandido funcionalmente, nos termos da *BSD Open source license*.

O DSpace possibilita a criação de coleções que aglutinam um conjunto de objetos afins. Permite a definição de *workflows* específicos para cada coleção, onde os diversos campos de metadados que descrevem o objeto catalogado são preenchidos. É muito flexível quanto à personalização do *workflow* de submissão de objetos para cada coleção: novos passos podem ser criados assim como determinados campos de metadados podem ser omitidos. Suporta o esquema de metadados DC e usa o banco de dados PostgreSQL[7] para a persistência. O DSpace é flexível quanto aos tipos de interface visual disponibilizados:

JSPUI: é a interface padrão do DSpace. Não é modular, porém é fácil de se trabalhar. Admite alterações pontuais no código fonte em JSP.

Manakin XMLUI: segue o paradigma de *templates* de interface, componentizando os elementos gráficos via schema *Digital Repository Interface (DRI)*[8] codificado em XML. Porém alterações pontuais exigem grande esforço, sobretudo se elementos não nativos (campos, ancoras com funções, comportamento dinâmico, etc.) precisam ser incorporados.

[6] http://www.adlnet.gov/capabilities/scorm

[7] http://www.postgresql.org/

[8] http://www.dspace.org/1_7_0Documentation/DRI Schema Reference.html

7.5 Melhorias no SRI implementado com DSpace

O SRI do AVEA UnA-SUS - UFSC utiliza o DSpace versão 1.7.1, que é uma aplicação Web desenvolvida em Java 6 e acessa um banco de dados PostgreSQL 9.0.4. O DSpace foi hospedado em um *container* de servlets Apache Tomcat[9], versão 6 e utiliza o Manakin XMLUI como técnica de renderização da interface. Toda esta infra-estrutura executa sobre um servidor FreeBSD [10] versão 8.1.

7.5.1 Anotação amparada por interfaces gráficas baseadas em conhecimento

As extensões incorporadas ao DSpace para visualização hierárquica, visualização hiperbólica e acesso eficiente a conhecimento foram implementadas com o *Prefuse 2007.10.21*, o *Treebolic 2.0.3* e o *Apache Lucene 2.4.1*, respectivamente. As personalizações exigidas para a integração ao DSpace dos componentes de interface baseados em conhecimento propostas por este trabalho primeiramente foram implementadas em JSPUI e depois em XMLUI. Através de funções JavaScript [11] é que conseguimos tal acoplamento. A Figura 29 mostra a interface Autocompletar integrada ao *worflow* de catalogação do DSpace UnA-SUS UFSC. Tal interface promove a anotação semântica ágil de OAs.

[9] http://tomcat.apache.org
[10] http://www.freebsd.org
[11] http://www.w3schools.com/js/default.asp

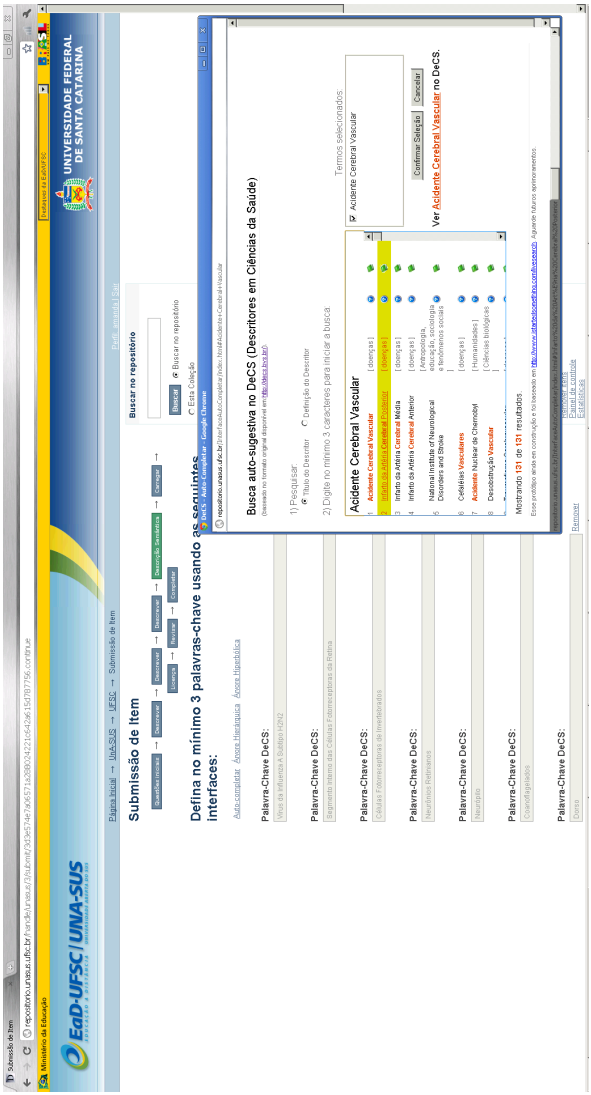


Figura 29: *Workflow* de catalogação: anotação semântica de OAs no DSpace Una-SUS com interface Autocompletar.

As anotações semânticas geradas são gravadas em um banco de dados à parte, usado pelo módulo de recuperação semântica. Tal módulo também foi integrado ao DSpace. Maiores detalhes sobre tal integração e o fluxo de dados podem ser vistos no Anexo N.

7.5.2 *Recuperação semântica de OAs*

Realizamos alterações na interface nativa de busca do DSpace de modo a acomodar um componente de interface que dispara consultas para o módulo de recuperação semântica desenvolvido. Tais alterações podem ser vistas na Figura 30.

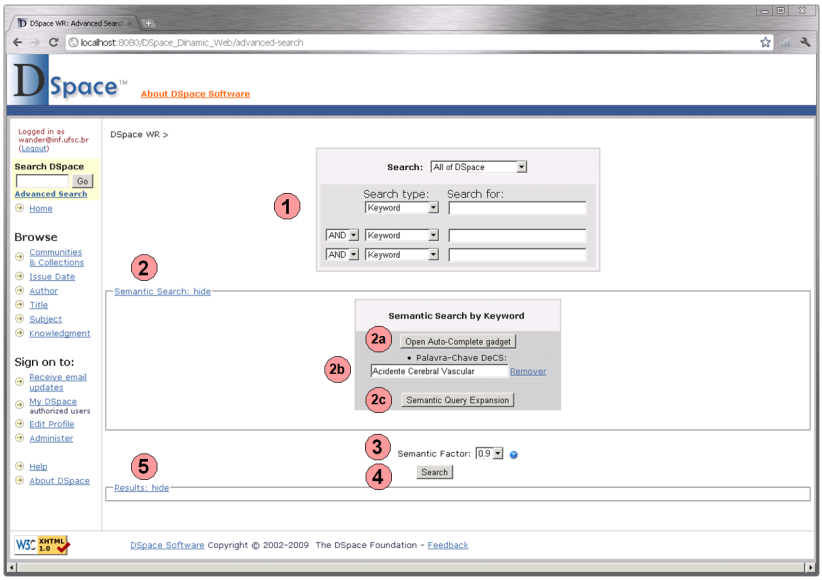


Figura 30: Busca semântica no DSpace.

Mantivemos todos os campos da busca nativa do DSpace (1), os quais realizam apenas buscas sintáticas e exploram valores de metadados tradicionais como autor, data, título, etc. Em (2) nosso componente que dispara buscas semânticas foi adicionado. Através dele é possível se invocar a interface Autocompletar (2a) e então selecionar termos que irão compor a consulta. Neste contexto, a interface Autocompletar também permite digitar livremente palavras-chave. Os termos selecionados ou digitados livremente na interface

Autocompletar são exibidos um abaixo do outro, em (2b), donde também podem ser removidos.

Após selecionar os termos que compõe a consulta, ela pode ser disparada para o módulo de recuperação semântica via ação do botão (4). Tal ação invoca a busca híbrida. Ela leva em conta tanto os valores dos campos nativos do DSpace (se estiverem preenchidos) quanto os providos pela interface Autocompletar. Então, após o processamento da consulta tanto pelo DSpace quanto pelo nosso módulo de recuperação semântica, a composição das listagens obtidas por cada método é feita. Finalmente, o resultado da busca híbrida é mostrado em (5). Maiores detalhes deste processo podem ser vistos no Anexo N.

7.5.2.1 Expansão semântica da consulta

Baseado nos termos selecionados via interface Autocompletar, opcionalmente, a expansão semântica da consulta também pode ser realizada (2c). Ela sugere termos semanticamente relacionados aos fornecidos inicialmente pelo usuário, de modo que o usuário possa refinar sua consulta ou descobrir novos termos para enriquecê-la. Os termos sugeridos são aqueles cujo valor de ativação foi considerado relevante após a execução do SA sobre a RS. A Figura 31 mostra a expansão de uma consulta originada pelo termo "Acidente Vascular Cerebral". Note que os termos são ordenados por significância.

No contexto da consulta expandida (Figura 31), valendo-se dos termos sugeridos que venham a ser selecionados, o usuário tanto pode disparar uma busca semântica (2c.1) ou sintática (2c.2) visando recuperar OIs. Aqui, ao contrário do método exposta na seção anterior, tais buscas são disjuntas e cada uma só retorna uma listagem de resultados. A busca semântica será enviada ao módulo de recuperação semântica. Já a busca sintática será enviada ao mecanismo de busca nativa do DSpace, donde a recuperação será dirigida aos valores do campo de metadado assunto ou palavras-chave *(keyword)*. Além de executar buscas, neste contexto o usuário pode acessar a definição de cada termo no sítio do VC, bastando clicar no nome de cada termo sugerido (2c.3).

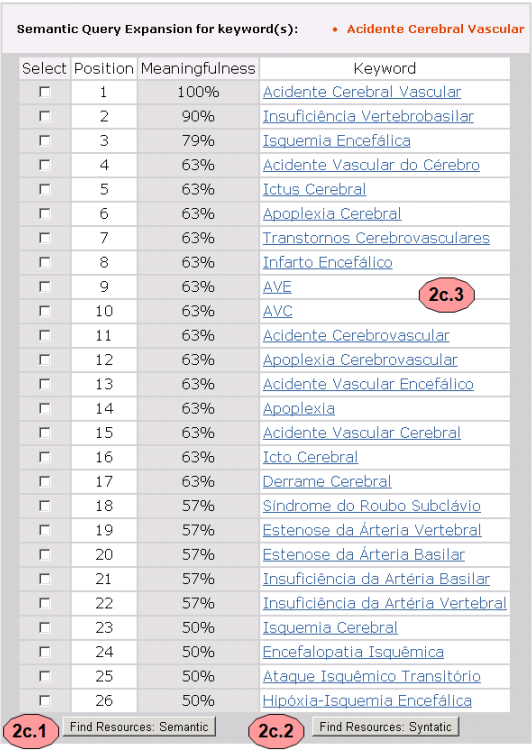


Figura 31: Expansão semântica da consulta no DSpace.

O módulo de recuperação semântica de OIs, apesar de funcional, ainda não figura no ambiente de produção do SRI do AVEA UnA-SUS - UFSC. Ele foi desenvolvido inicialmente com JSPUI e funciona em um ambiente de teste. Um esforço extra seria necessário para portá-lo ao padrão Manakin XMLUI. Implementações futuras tratarão desta questão.

7.6 Vocabulário controlado DeCS

Segundo (PELLIZZON, 2004), o VC Descritores em Ciências da Saúde (DeCS)[12] foi elaborada em 1986 pelo Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME). É originário do *Medical*

[12]http://decs.bvs.br

Subject Headings (MeSH)[13], o qual existe desde 1963 e é de responsabilidade da *United States National Library of Medicine (NLM)*[14].

O DeCS é um vocabulário estruturado trilíngue (português, espanhol e inglês) baseado em coleções de termos, organizados para facilitar o acesso à informação. Foi criado para servir como uma linguagem única na descrição de artigos de revistas científicas, livros, anais de congressos, relatórios técnicos, e outros tipos de materiais, assim como para ser usado na pesquisa e recuperação de assuntos da literatura científica nas fontes de informação disponíveis na Biblioteca Virtual em Saúde (BVS) como LILACS, MEDLINE e outras (BIREME, 2010). Mais detalhes sobre sua estrutura e aquisição podem ser vistos no Anexo O.

Neste capítulo abordamos o projeto UnA-SUS e a implementação de seu viés tecnológico no AVEA UnA-SUS - UFSC valendo-se do repositório DSpace. O DSpace armazena Objetos de Aprendizagem (OA), os quais foram anotados com termos do VC DeCS valendo-se das interfaces Web baseadas em conhecimento desenvolvidas. Ainda mostramos como nosso módulo de busca semântica se integra ao DSpace, possibilitando consultas híbridas que exploram aspectos sintáticos e semânticos dos metadados que descrevem os OAs catalogados.

[13] http://www.nlm.nih.gov/mesh
[14] http://www.nlm.nih.gov

8 TESTES E AVALIAÇÃO DOS TESTES

Os testes que avaliam as contribuições deste trabalho são apresentados neste capítulo. Testes de usabilidade com catalogação de OIs realizados por usuários serão vistos a seguir. Em seguida, testes de desempenho envolvendo o módulo de recuperação semântica de OIs terão vez.

Testes de usabilidade com usuários foram conduzidos visando aferir a aceitação e o potencial das interfaces baseadas em conhecimento para anotação de OIs. Testes para aferir o desempenho do sistema de recuperação semântica também foram realizados. Estes últimos testes não contaram com a colaboração de usuários e visaram principalmente descobrir correlações entre os parâmetros de configuração do módulo de recuperação e aferir o consumo de recursos da máquina à medida que: (i) os valores destes parâmetros são alterados e (ii) a quantidade de OIs catalogados aumenta.

8.1 Testes de usabilidade com catalogação

Visando aferir a aceitação e o potencial das interfaces gráficas baseadas em conhecimento propostas neste trabalho, testes de usabilidade explorando a catalogação, e em especial, o processo de anotação de OAs ambientados no SRI descrito no Capítulo 7 foram conduzidos. Tais testes também serviram para colher sugestões de melhorias nas implementações.

Ressaltamos que tais testes obtiveram o aval de todos os usuários participantes e foram aprovados pelo Comitê de Ética em Pesquisa com Seres Humanos (CEPSH) em processo registrado junto a tal orgão sob o título "Tecnologias de Informação aplicadas na Formação e Aperfeiçoamento de Profissionais da UnA-SUS - UFSC", cujo número do processo é 1827. O certificado que respalda os testes pode ser visualizado no Anexo D.

Para preservar o anonimato, conforme estipulado no Termo de Consentimento Livre e Esclarecido (TCLE), o qual figura no Anexo D e que os usuários tiveram acesso antes dos testes, a identidade dos usuários participantes foi omitida. Deste modo, daqui para diante eles serão chamados de avaliadores (avaliador 1, avaliador 2, e assim por diante).

8.1.1 Objetivos

Os objetivos dos testes de usabilidade com catalogação de OIs são:

1. Avaliar a **aceitação** das soluções propostas e a **satisfação** dos usuários ao fazerem uso de tais soluções.
2. Avaliar o **potencial** das soluções propostas.
3. Avaliar a **eficácia** das soluções propostas.
4. Avaliar a **eficiência** das soluções propostas.
5. Quantificar possíveis **melhorias** alcançadas pela aplicação da abordagem proposta de suporte ao uso de conhecimento de domínio em comparação a uma abordagem desprovida de tal suporte.
6. Avaliar a **eficiência** para se acessar, navegar e explorar as funcionalidades do AVEA UnA-SUS UFSC.
7. Detectar possíveis **falhas** decorrentes da falta de escalabilidade do sistema.
8. Colher **observações**, críticas e sugestões junto aos avaliadores.

Pela aplicação do método *GQM (Goal, Question, Metric)* (BASILI; ROMBACH, 1988), questões e métricas de avaliação foram esmiuçadas a partir do refinamento dos objetivos. A seguir, sucintamente, são esboçados os subprodutos resultantes do método aplicado: as questões e as métricas.

8.1.1.1 Questões

As questões desdobram os conceitos abstratos relatados nos objetivos em subfatores, facilitando o entendimento de tais conceitos. Para o objetivo 1, as seguintes questões foram elaboradas:

Questão	Descrição	Métrica
Q01	Os usuários gostam de usar as interfaces baseadas em conhecimento?	M1, M2, M3, M4
Q02	Os usuários julgam ser fácil usar as interfaces baseadas em conhecimento?	M1, M2 ,M3, M4
Q03	Os usuários preferem qual interface baseada em conhecimento?	M5, M6, M7
Q04	Os usuários julgam que as interfaces baseadas em conhecimento são claras e fáceis de entender?	M1, M2, M3, M4

Tabela 3: Questões do objetivo 1.

No Anexo F estão descritas todas as questões.

8.1.1.2 Métricas

Para responder as questões, métricas foram operacionalizadas. As seguintes métricas dizem respeito às 4 questões do objetivo 1:

Métrica	Coletada via	Descrição
M01	Questionário	Número de avaliadores que escolheram "Discordo fortemente".
M02	Questionário	Número de avaliadores que escolheram "Discordo".
M03	Questionário	Número de avaliadores que escolheram "Concordo".
M04	Questionário	Número de avaliadores que escolheram "Concordo fortemente".
M05	Questionário	Número de avaliadores que escolheram "Autocompletar".
M06	Questionário	Número de avaliadores que escolheram "Hierárquica".
M07	Questionário	Número de avaliadores que escolheram "Hiperbólica".

Tabela 4: Métricas das questões do objetivo 1.

As demais métricas estão descritas no Anexo G. Elas foram coletadas através de um questionário aplicado aos avaliadores, fichas catalográficas

preenchidas pelos avaliadores e ferramentas específicas para a realização de testes de usabilidade, tais como o *software* Morae [1], além da instrumentação do código (*logging*) do SRI.

8.1.2 Ambiente

Os testes de usabilidade com catalogação de OIs foram realizados nos Laboratórios de Informática (LIICT) do Centro Tecnológico (CTC) da Universidade Federal de Santa Catarina (UFSC). Cada avaliador recebeu um *login* e uma senha para acessar via navegador Web o SRI do estudo de caso descrito no Capítulo 7. As máquinas disponibilizadas aos avaliadores contam com Windows XP, processador Intel Core 2 Duo E7500 de 2.93 GHz e 2 GB de memória RAM e navegadores Web Firefox e Internet Explorer.

8.1.3 Avaliadores

Os testes contaram com a participação de 25 avaliadores, os quais foram recrutados conforme o ramo de atuação profissional:

- 18 usuários da área de tecnologia da informação (TI).
- 04 profissionais da área de saúde.
- 03 profissionais da área de design gráfico (DG) e desing instrucional (DI).

Os usuários da área de TI representam os usuários que navegam na Web e contribuem com a visão técnica e possivelmente avaliam mais aspectos de ordem tecnológica, enquanto os profissionais da área da saúde e de DI são os reais usuários do AVEA UnA-SUS UFSC. Esses dois últimos podem contribuir com suas percepções enquanto se ambientam com o sistema. Tais avaliadores foram reunidos em grupos conforme a disponibilidade de horário. Os profissionais da saúde figuram em menor número devido a indisponibilidade de agenda. Já os 3 profissionais de design envolvidos nos testes são a totalidade de membros de DI e DG alocados para o projeto UnA-SUS, no contexto da UFSC.

[1] http://www.techsmith.com/morae.asp

8.1.4 Procedimento dos testes

Visando avaliar a utilização das funcionalidades para anotação de OIs do SRI e dar aos usuários a oportunidade de explorar tais funcionalidades, os avaliadores realizaram uma seqüência de tarefas de catalogação de OAs. Antes da execução destas tarefas, os avaliadores receberam instruções verbais amparadas por uma apresentação, receberam um guia como o mostrado no Anexo I, assistiram a um vídeo sobre o repositório e se familiarizaram com as interfaces baseadas em conhecimento, com o DeCS e com dois OAs, com os seguintes títulos:

- **OA1**: Principais Agravos à Saúde da Criança.
- **OA2**: Infecções Respiratórias Agudas mais Comuns nas Crianças.

As tarefas realizadas foram as seguintes:

- **T1**: Catalogar o OA1 manualmente via o preenchimento de uma ficha catalográfica em papel (vide Anexo J).
- **T2**: Valendo-se da ficha catalográfica do OA2 previamente preenchida por especialistas de domínio (vide Anexo K), catalogar o OA2 no repositório.
- **T3**: Valendo-se da ficha catalográfica preenchida manualmente pelo avaliador em T1, catalogar o OA1 no repositório.

Porém, devido à disponibilidade de tempo, para um grupo de 15 alunos da disciplina de Engenharia de Usabilidade do curso de Sistemas de Informação da UFSC, T2 foi alterada, tornando-se T2' e uma quarta tarefa (T4) foi criada:

- **T2'**: Valendo-se da ficha catalográfica do OA2 previamente preenchida por especialistas de domínio, catalogar o OA2 no repositório utilizando somente a interface gráfica nativa do DeCS [2] para buscar os termos que descrevem o OA2.
- **T4**: Valendo-se da ficha catalográfica do OA2 previamente preenchida por especialistas de domínio, catalogar o OA2 no repositório utilizando as interfaces baseadas em conhecimento para buscar os termos que descrevem o OA2.

[2] http://dspace.org

Ressaltamos que devido à limitação de tempo das equipes, as tarefas T2' e T4 não puderam ser aplicadas aos profissionais das áreas de saúde, design gráfico e desing instrucional.

8.1.5 Questionário

Após a execução das tarefas descritas na seção anterior, aos avaliadores foi aplicado um questionário para coleta dos dados exigidos pelas métricas definidas para avaliar o sistema que implementa a nossa proposta. A maioria das alternativas de resposta são baseadas na escala de Likert de 4 pontos (LIKERT, 1932). O respondente indica o seu grau de concordância ou discordância, selecionando para cada item do questionário um valor de resposta na escala dada. As opções de resposta são:

- Discordo fortemente
- Discordo
- Concordo
- Concordo fortemente

O Anexo H apresenta todas as questões do questionário.

8.2 Avaliação dos testes de usabilidade com catalogação

Os resultados dos testes de usabilidade com catalogação foram analisados quantitativamente (em termos do tempo gasto pelo avaliadores) e qualitativamente (em termos da opinião dos avaliadores ao responder o questionário).

8.2.1 Análise quantitativa

A análise quantitativa utiliza dados colhidos pela instrumentação do código do AVEA UnA-SUS UFSC e pelo Morae. Ela compara o tempo gasto pelos 11 avaliadores para realizarem as anotações propostas nas tarefas T2' e T4, as quais foram descritas na seção 8.1.4. A Figura 32 destaca o quanto a anotação usando as interfaces baseadas em conhecimento (T4) foi mais rápida que a anotação baseada no DeCS (T2'), para cada um dos 11 avaliadores para os quais os dados coletados puderam ser tratados (dos 15 avaliadores que

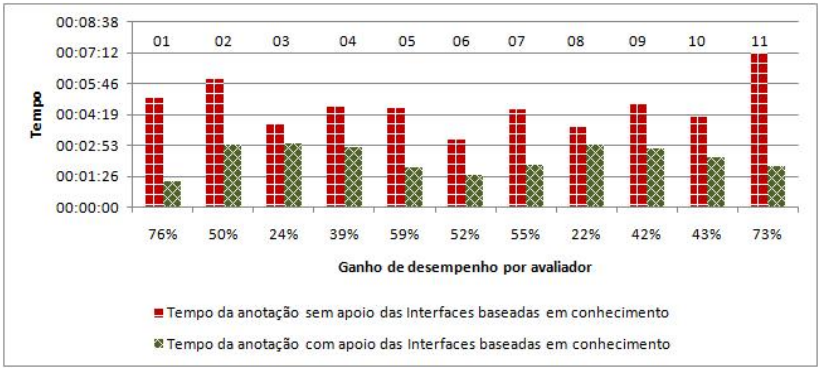


Figura 32: Comparação do tempo de anotação de T2' e T4.

iniciaram o teste de usabilidade, 4 o abandonaram). Na maioria dos casos o ganho ultrapassou os 50% e em dois deles os 70%.

A Figura 33 mostra que, com o apoio das interfaces baseadas em conhecimento, houve melhora de 51% na média do tempo de execução das anotações propostas pela tarefa T4 (00:02:19) em relação à tarefa T2' (00:04:45), que não contou com tal apoio.

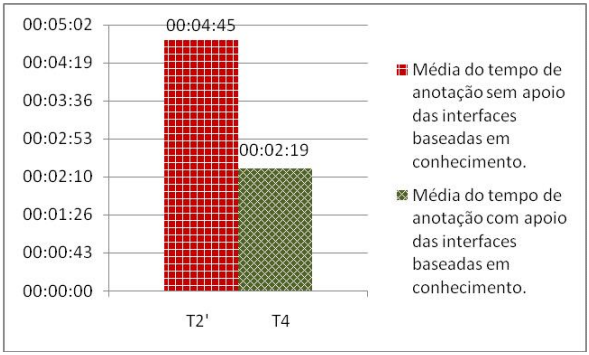


Figura 33: Média do tempo de anotação de T2' e T4.

8.2.2 Análise qualitativa

A análise qualitativa avalia a percepção dos 25 avaliadores ao usar o sistema proposto via dados coletados pelo questionário descrito na seção 8.1.1.1. A Figura 34 sintetiza a satisfação dos avaliadores ao usar as interfaces baseadas em conhecimento. Note que a aprovação foi sempre acima de 50% nos seguintes quesitos usados para aferir o grau de satisfação:

- Q1 - Gosta de usar as interfaces baseadas em conhecimento.
- Q2 - Julga fácil operar essas interfaces.
- Q3 - Essas interfaces são claras e fáceis de entender.

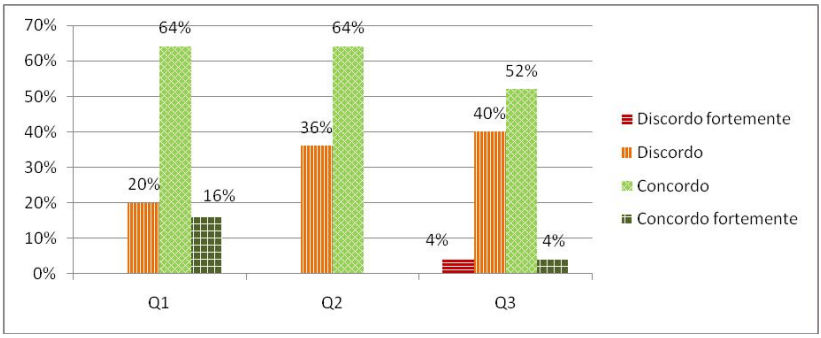


Figura 34: Satisfação do usuário.

A interface preferida por 100% dos avaliadores para a inserção de valores de metadados foi a Autocompletar, em detrimento das interfaces Hiperbólica e Hierárquica. A Figura 35 mostra que 60% dos avaliadores julgaram a interface Hiperbólica como a mais trabalhosa de usar, seguida pela Hierárquica com 32% e a Autocompletar com 8%.

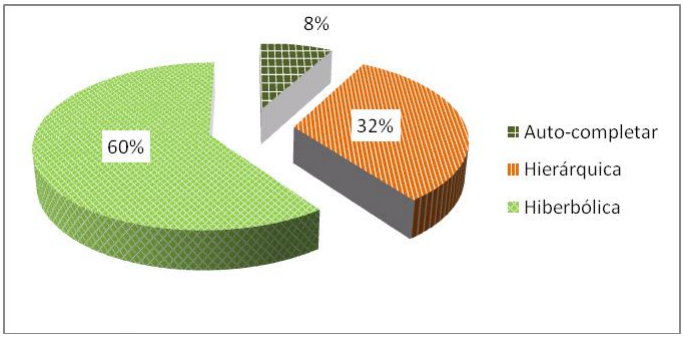


Figura 35: Interface mais trabalhosa.

A Figura 36 retrata a avaliação do potencial de uso das interfaces, de acordo com os seguintes quesitos:

- Q4 - Útil para o preenchimento de outros campos de metadados diferentes do campo palavra-chave.
- Q5 - Útil em outros domínios de aplicação.

Em ambos os quesitos, os valores ultrapassam os 60%.

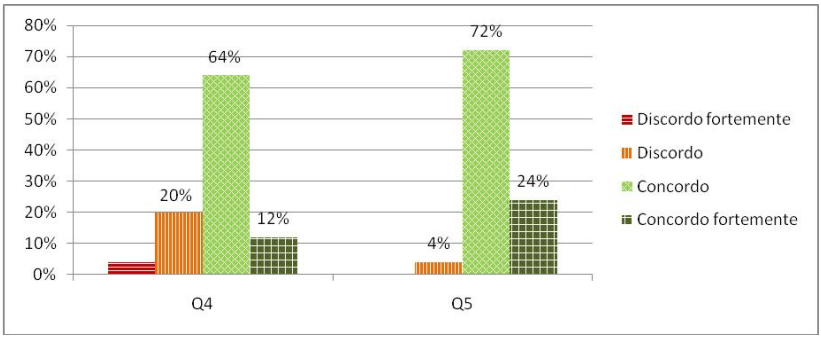


Figura 36: Potencial de uso.

A Figura 37 mostra que 76% dos avaliadores julgaram que as interfaces baseadas em conhecimento facilitaram a execução das tarefas. Isto em

comparação à abordagem de descrição manual ou valendo-se somente da interface de consulta do DeCS em seu formato original[3].

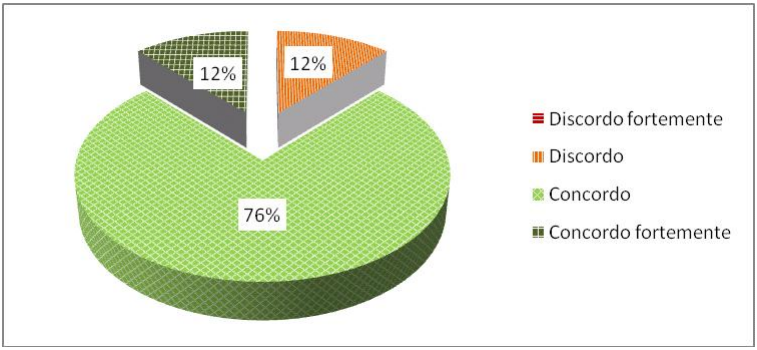


Figura 37: Facilitação na execução das tarefas.

A Figura 38 mostra que 72% dos avaliadores julgaram que os termos oriundos do DeCS e providos pelas interfaces foram suficientes para descrever os OAs.

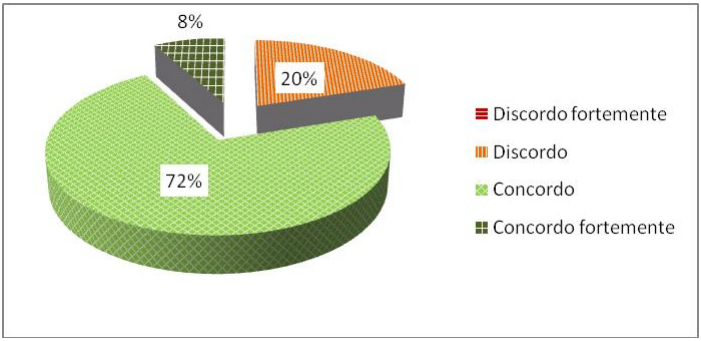


Figura 38: Completude do vocabulário.

A Figura 39 relata que 48% dos avaliadores realizaram as tarefas com facilidade, porém, para 32% dos avaliadores, as tarefas propostas impuseram certa dificuldade.

[3]http://decs.bvs.br

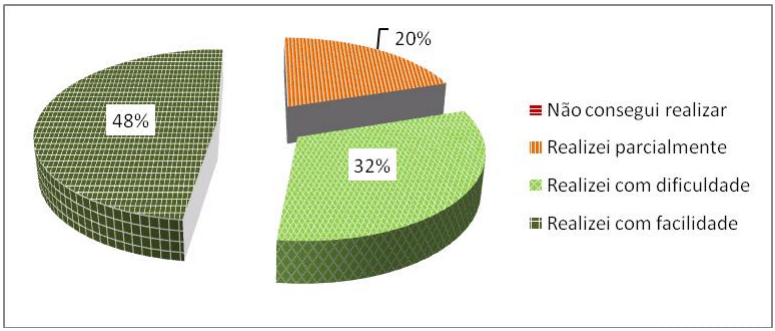


Figura 39: Realização das tarefas.

8.2.3 Discussão

Muitos avaliadores tiveram dificuldade em pesquisar termos no sítio do DeCS[4] e descrever OAs a partir do entendimento do seu conteúdo. Na maioria das vezes, os avaliadores realizam consultas infrutíferas no DeCS por utilizarem termos compostos como "agravos à saúde da criança", o que não é suportado pela interface de busca do DeCS.

O uso das interfaces baseadas em conhecimento aliadas ao DeCS permitiu que erros de grafia inseridos propositalmente nas fichas catalográficas previamente preenchidas fossem detectados pelos usuários e corrigidos durante as descrições dos objetos. O suporte semântico também ajudou a "orientar" a formulação das consultas. O *feedback* da interface Autocompletar é instantâneo, permitindo a correção imediata dos termos usados na consulta.

Nossos testes apontam que com o apoio das interfaces baseadas em conhecimento alguns avaliadores conseguiram ganhos superiores a 70% no tempo de anotação. Tal ganho foi capitaneado principalmente pelo uso da interface Autocompletar. Ela teve a aprovação unânime de todos os avaliadores e foi considerada mais ágil e fácil de operar, justamente por ser similar a ferramentas de busca difundidas atualmente.

O foco limitado da visualização hiperbólica dificulta a visualização e seu comportamento dinâmico atrapalha a navegação e a seleção de informações. Já a visualização hierárquica de grandes trechos de um VC estruturado como o DeCS foi considerada confusa devido ao emaranhado de termos e relações apresentadas, o que também dificulta a navegação. Em

[4]http://decs.bvs.br

decorrência de tais observações coletadas nos testes de usabilidade, melhorias serão implementadas visando facilitar a operação e o entendimento do sistema proposto.

8.2.4 Ameaças à validade dos testes

Na última etapa de aplicação dos testes de usabilidade, dos 15 avaliadores que iniciaram os testes, 4 fizeram uso de seu direito de abandoná-lo após um certo tempo. Tais avaliações foram desconsideradas, porém é temerário que tal comportamento possa ter desestimulado os demais avaliadores. Além disso, a aplicação dos testes para profissionais leigos na área da saúde pode ter causado um desconforto e certas dificuldades adicionais para se manusear o VC, sobretudo porque descrever um OA valendo-se de um conjunto limitado de termos providos pelo DeCS é uma tarefa árdua até para profissionais da saúde. Eles nos relataram que termos jurídicos seriam necessários para se descrever certos OAs ligados, por exemplo, aos Direitos da Criança. Porém, em nossos testes de usabilidade, OAs bem simples foram usados. Além disso, os termos que descrevem o OA envolvido nas tarefas T2, T2' e T4 foram previamente fornecidos.

8.3 Testes com o módulo de busca semântica

Inicialmente testes de recuperação de OIs com o módulo de busca semântica foram planejados, porém o repositório UnA-SUS não conta com um número considerável de objetos armazenados. Hoje são em torno de 40, o que impossibilita testes confiáveis que avaliem a cobertura e precisão das consultas. Além disso, o motivo determinante é que os OIs catalogados no repositório UnA-SUS foram mal descritos e/ou anotados com poucos metadados. Isso de deve a falta de visão e comprometimento de algumas equipes com um dos objetivos mais nobres do projeto UnA-SUS: o compartilhamento de conhecimento através do reuso de OAs devidamente descritos. Sendo assim, direcionamos nossos esforços para testes de desempenho, os quais foram conduzidos com o auxílio de um protótipo (vide Anexo P), que foi desenvolvido para facilitar a execução de tais testes.

8.3.1 Calibração dos parâmetros do SA

Como a execução do SA é regida por diversos parâmetros, é premente saber quais são os valores mais adequados para tais parâmetros, sob a óptica

do desempenho e qualidade de resultados. As pesquisas realizadas por (PEREIRA, 2011) em seu Trabalho de Conclusão de Curso (TCC) valendo-se do protótipo desenvolvido sugerem valores ótimos para a configuração dos parâmetros que regem a execução do SA sobre uma RS originada de uma porção do DeCS. Tais pesquisas foram co-orientadas pelo autor deste trabalho e seguiram as seguintes diretrizes.

8.3.1.1 Objetivos

1. A partir de variações nos valores dos parâmetros de configuração do SA, avaliar qual o impacto de tais parâmetros no desempenho do SA e na qualidade dos resultados das buscas.

2. A partir de variações nos valores dos parâmetros de configuração do SA, descobrir possíveis correlações que ocorram entre tais parâmetros.

3. Sugerir valores de parâmetros de configuração do SA mais adequados ao domínio do estudo de caso.

8.3.1.2 Métricas

As seguintes métricas são objetos de estudo:

- **tempo de execução**: quantos milissegundos o SA levou para executar uma consulta. Envolve a expansão semântica e calculo do *ranking*;

- **quantidade de OIs**: quantidade de OIs obtidos em cada consulta (cobertura).

8.3.1.3 Definição dos testes

Base de conhecimento e coleção de OIs anotados

Os termos $t_i \in T$ que compõe a RS foram tomados das categorias Anatomia, Doenças e Organismos do DeCS e dos sinônimos de tais t_i , totalizando 9.596 nodos. Cada termo t_i foi denotado por um número natural no intervalo [1, 9.596]. Os $oi_i \in OI$ de RS foram gerados pela ferramenta DBGen[5], que produziu 10.000 oi_i sintéticos. Cada oi_i foi denotado por um número natural no intervalo [1, 10.000] e foi anotado com 1 a 7 termos, segundo a distribuição Gaussiana (Normal). Cada anotação associou aleatoriamente um

[5] http://www.gbdi.icmc.usp.br/download?q=node/350

termo $t_i \in [1,9.596]$ a um oi_i. O número de termos usados para anotar cada oi_i assim como a distribuição são mostrados na Tabela 5:

Termos	1	2	3	4	5	6	7
OIs	660	990	1650	3400	1650	990	660

Tabela 5: Número de termos que anotam os OIs.

Consultas (C)

Numa primeira etapa, consultas de tamanho variado foram compostas aleatoriamente a partir dos $t_i \in [1,9.596]$. Em seguida, a tais consultas foram acrescidos 5 $t_i \in [1,9.596]$ mais gerais, de acordo com o DeCS. Estas consultas foram aplicadas como entrada do algoritmo de SA em cada bateria de teste.

Parâmetros observados (P)

Os parâmetros alvos de nossos testes são:

- limiar de disparo *(firing threshold)*
- fator de decaimento
- número máximo de nodos por pulso

A escolha de tais parâmetros se deve ao fato de que em testes empíricos anteriores, estes parâmetros apresentaram certa correlação. Já o número máximo de pulsos e o número máximo de nodos são mantidos em aberto (infinito) com o intuito de não intervir nos testes.

8.3.1.4 Ambiente

Os testes foram realizados em uma máquina Intel Core 2 Duo E6550 @ 2.33 GHz, 4GB de memória principal (RAM), sistema operacional Microsoft Windows 7 Professional, de 32 bits. O código do protótipo foi instrumentado e executado na IDE Eclipse 3.5, sob uma JRE 6.

8.3.1.5 Execução dos testes

Procedeu-se da seguinte maneira:

i) determinou-se 16 conjuntos de valores iniciais para cada parâmetro $p \in P$, conforme Tabela 6.

ii) então selecionou-se um parâmetro $p \in P$ e atribui-se o valor inicial mínimo

iii) fixou-se o valor dos demais parâmetros;

iv) variou-se 20 vezes o valor do parâmetro p com passos constantes;

v) a cada novo valor de p foram executadas 2 consultas no SA.

Testes empíricos mostraram que o SA apresenta o melhor tempo de execução quando o limiar de disparo *(firing threshold)* e o fator de decaimento estão no intervalo [0.3, 0.6]. Desta forma usamos tal intervalo para atribuir os valores iniciais aos parâmetros analisados, variando os valores com passos de 0.1, conforme a Tabela 6. Ao número máximo de nodos por pulso foi atribuído o valor 50, embora os testes empíricos mostraram que tal parâmetro não é relevante para o tempo de execução, quando o SA executa sobre uma RS originada do DeCS.

Teste	**Máx. Nodos/Pulso**	**Threshold**	**Fator de decaída**
01	50	0,3	0,3
02	50	0,4	0,3
03	50	0,5	0,3
04	50	0,6	0,3
05	50	0,3	0,4
06	50	0,4	0,4
07	50	0,5	0,4
08	50	0,6	0,4
09	50	0,3	0,5
10	50	0,4	0,5
11	50	0,5	0,5
12	50	0,6	0,5
13	50	0,3	0,6
14	50	0,4	0,6
15	50	0,5	0,6
16	50	0,6	0,6

Tabela 6: Valores iniciais dos parâmetros.

Os testes partiram dos 16 conjuntos de valores iniciais que foram compostos. Cada teste compreende 20 execuções do SA, na qual cada um dos parâmetros não fixado foi alterado gradativamente com passos constantes: 0,05 para limiar de disparo e fator de decaída e 1, 5, 10 para número máximo de nodos por pulso. Então se coletou o tempo de CPU e a quantidade de OAs obtidos para cada execução do SA. As médias logarítmicas das métricas avaliadas nos 16 testes são mostradas no próximo capítulo.

8.3.2 Desempenho com recuperação

Antes de ser efetivamente usado em um SRI, nosso módulo de busca semântica precisa passar por testes que avaliem o desempenho computacional e a escalabilidade de tal solução.

8.3.2.1 Objetivos

Os testes com o módulo de busca semântica desenvolvido objetivam aferir o desempenho computacional do SA atuando sobre grandes bases de conhecimento.

8.3.2.2 Métricas

Para cada uma das 20 bases de conhecimento elaboradas para os testes, as seguintes métricas são objetos de estudo:

- **tempo de carga**: quantos milissegundos o SA levou para a montagem e carga de cada base de conhecimento em memória.

- **tempo de execução SA**: quantos milissegundos o SA levou para executar 20 consultas sobre cada uma das bases de conhecimento. Envolve a expansão semântica e calculo do *ranking*;

- **consumo de RAM na carga**: quantidade de kilobytes de memória RAM necessária para a carga de cada base de conhecimento em memória;

- **quantidade de OIs**: quantidade de OIs obtidos nas 20 consultas (cobertura) sobre cada base de conhecimento.

8.3.2.3 Definição dos testes

Base de conhecimento e coleções de OIs anotados

Os termos $t_i \in T$ que compõe a espinha dorsal da base de conhecimento *BC* foram tomados das 20 categorias do DeCS. Ela é composta de 65.134 nodos, sendo que 36.330 nodos são sinônimos dos tais t_i. Ainda criamos um nodo raiz artificial, o Root_DeCS, que conecta as 20 categorias do VC via o termo raiz de cada uma delas (Anatomia, Doenças, Organismos, etc.). Cada termo t_i foi denotado por um número natural no intervalo [1, 65.134].

Os $oi_i \in OIS$ foram gerados sinteticamente utilizando-se a ferramenta DBGen[6], que produziu 20 coleções Col_i de OIs sintéticos. O tamanho das coleções Col_i varia de 10.000 OIs ($|Col_{01}|$) a 200.000 ($|Col_{20}|$), em passos de 10.000. Cada $oi_i \in Col_i$ foi denotado por um número natural no intervalo [1, $|Col_i|$]. Tais coleções foram salvas em arquivos RDF.

Anotações

As anotações semânticas foram geradas aleatoriamente via DBGen. Cada OI recebeu de 1 a 7 anotações, segundo a distribuição Gaussiana (Normal). Os OIs de cada Col_i foram particionados em 7 categorias Cat_j de acordo com a quantidade de anotações que receberam. A lei de formação da distribuição das anotações leva em conta o tamanho de cada coleção para então definir a quantidade de OIs em cada categoria Cat_j, segundo os percentuais mostrados na Tabela 8:

Anotações	1	2	3	4	5	6	7
Col_i	6,6%	9,9%	16,5%	34%	16,5%	9,9%	6,6%

Tabela 7: Lei de formação da distribuição das anotações sobre os OIs de cada coleção.

Para a Col_{01}, com 10.000 OIs e Col_{20}, com 200.000 OIs, obteve-se as seguintes distribuições:

A partir das anotações de T sobre cada uma das 20 coleções Col_i de OIs, obtivemos coleções de anotações $AS_i(T, Col_i)$. Tais coleções de anotações também foram salvas em arquivos RDF.

[6] http://www.gbdi.icmc.usp.br/download?q=node/350

Anotações	1	2	3	4	5	6	7
Col_{01}	660	990	1650	3400	1650	990	660
Col_{20}	13330	19995	33325	66700	33325	19995	13330

Tabela 8: Exemplo de distribuição das anotações sobre os OIs das coleções Col_{01} e Col_{20}.

Consultas (C)

Foram geradas 20 consultas. Elas estão agrupadas em famílias, de acordo com as características explicitadas na Tabela 9. A profundidade (Pf) de um nodo é a distância deste nodo até a raiz Root_DeCS da RS.

Família	**Característica**	**Profundidade (Pf)**
F_1: Generalizada	Termos mais gerais na hierarquia do VC.	$3 \geq Pf > 0$
F_2: Especializada	Termos mais específicos na hierarquia do VC.	$Pf \geq 8$
F_3: Sinônimos	Termos que são sinônimos segundo o VC.	Pf qualquer
F_4: Normal	Termos que não se encaixam nas famílias anteriores.	$8 > Pf \geq 3$

Tabela 9: Famílias das consultas.

Cada consulta c_i é composta de 1 até 5 palavras-chave p_i, pois segundo (JANSEN, 2000), 93% das consultas na Web têm menos de cinco termos. Cada palavra-chave é denotada por um número natural no intervalo [1, 65.134], correspondendo a um termo $t_i \in T$, como mostra a Tabela 10.

Família	Consultas	Termo(s)
F_1	c_{01}	{39742}
	c_{02}	{10981, 52307}
	c_{03}	{59835, 64503, 56152}
	c_{04}	{47008, 22547, 1, 9333}
	c_{05}	{58001, 48064, 28, 12525, 19422}
F_2	c_{06}	{35588}
	c_{07}	{503, 908}
	c_{08}	{1479, 32528, 2496}
	c_{09}	{36301, 63964, 138, 63003}
	c_{10}	{1710, 2666, 29959, 33639, 50869}
F_3	c_{11}	{34}
	c_{12}	{314, 13439}
	c_{13}	{18904, 25569, 32536}
	c_{14}	{39953, 43514, 49465, 65124}
	c_{15}	{64398, 64039, 63090, 61486, 57437}
F_4	c_{16}	{831}
	c_{17}	{65020, 3849}
	c_{18}	{56827, 101, 64482}
	c_{19}	{3727, 8770, 54465, 38952}
	c_{20}	{13, 2869, 65133, 9913, 26597}

Tabela 10: Consultas.

Os termos do DeCS que compõe as consultas da família F_1 são mostrados na Tabela 11. Eles estão separados por ponto e vírgula.

Consulta	Palavras-chave
c_{01}	misturas complexas
c_{02}	doenças; terapêutica homeopática
c_{03}	políticas, planejamento e administração em saúde; localizações geográficas; educação
c_{04}	psiquiatria e psicologia; compostos químicos e drogas; anatomia; vírus
c_{05}	assistência à saúde; transtornos mentais; histeria dissociativa; neoplasias; doenças cardiovasculares

Tabela 11: Exemplos de consultas da família F_1.

8.3.2.4 Ambiente

Os testes foram realizados em uma máquina Intel Core 2 Duo T8100 @ 2.10 GHz, 4GB de memória principal (RAM), sistema operacional Microsoft Windows Vista, de 32 bits. O código do protótipo foi instrumentado e executado na IDE Eclipse 3.6, sob uma JRE 6.

8.3.2.5 Configuração dos parâmetros do SA

Como exposto previamente na seção 8.3.1, o trabalho realizado por (PEREIRA, 2011) sugere valores ótimos para a configuração dos parâmetros que regem a execução do SA sobre uma RS originada de uma porção do DeCS. Valemo-nos de tais insumos para então configurar, via protótipo, nosso módulo de busca semântica. Os parâmetros considerados e seus respectivos valores estão descritos na Tabela 12:

Parâmetro	Valor
Limiar de disparo *(firing threshold)*	0.45
Fator de decaimento	0.65
Número máximo de nodos por pulso	50
Peso da relação de sinonímia	0.9
Peso da relação de anotação	1.0
Sentido: quando especializa	1.0
Sentido: quando generaliza	0.8

Tabela 12: Valores dos parâmetros do SA.

8.3.2.6 Execução dos testes

Procedeu-se da seguinte maneira:

1. A partir das 20 consultas c_i formuladas;
2. A partir das 20 coleções Col_i de OIs;
3. A partir das 20 coleções de anotações $AS_i(T, Col_i)$;
4. Via *framework* semântico Jena, cada $AS_i(T, Col_i)$ foi adicionada a *BC*, gerando $RS_i(T, Col_i)$.

5. As 20 consultas c_i foram submetidas ao SA executando sobre cada $RS_i(T, Col_i)$, então as métricas foram colhidas via ferramentas de análise, como a VisualVM [7] e instrumentação do código (*logging*) do protótipo.

8.4 Avaliação dos testes com o módulo de busca semântica

Doravante discutimos os resultados dos testes realizados com o módulo de busca semântica. Primeiramente os testes que visaram observar o comportamento do SA segundo variações nos valores de seus parâmetros de configuração. Em seguida mostramos os resultados dos testes de escalabilidade.

8.4.1 Calibração dos parâmetros do SA

A Figura 40 mostra a média do tempo de CPU gasto para executar o SA com variação nos parâmetros limiar de disparo *(firing threshold)*, fator de decaimento e número máximo de nodos por pulso.

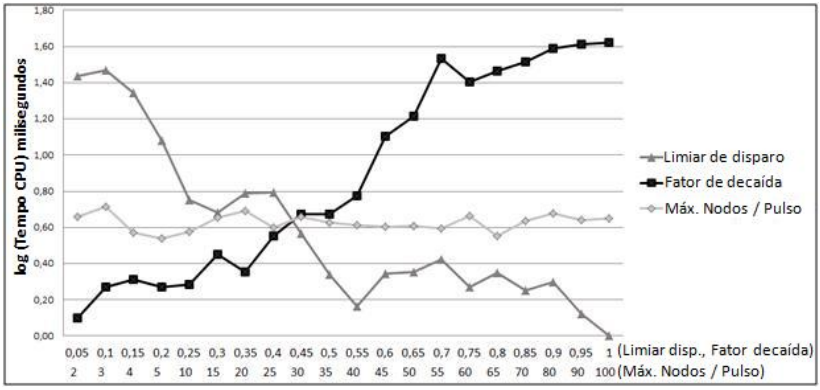


Figura 40: Média do tempo de CPU gasto para executar o SA com variação nos parâmetros.

Tal métrica está em escala logarítmica para facilitar a comparação das curvas quando os valores são baixos e estão muito próximos. Observa-se na Figura 40 o efeito inverso provocado pelo incremento dos valores dos

[7] http://visualvm.java.net/

parâmetros limiar de disparo, fator de decaimento no tempo de CPU gasto para executar o SA. O tempo médio de CPU geralmente aumenta com o crescimento do fator de decaimento e a diminuição do limiar de disparo. Isso ocorre porque mais nodos são atingidos (e conseqüentemente devem ser processados) pelas ondas de ativação do SA.

O número máximo de nodos por pulso não influenciou nos resultados. O tempo médio de CPU gasto para executar o SA ficou estatisticamente estável para os diferentes valores escolhidos para este parâmetro. Isso aconteceu porque os valores escolhidos para o número máximo de nodos por pulso nesses experimentos não foram alcançados durante as expansões semânticas sobre a RS originada do DeCS, onde os nodos são normalmente conectados em uma hierarquia, o grau dos nodos são baixos, e o número de arestas é uma função linear do número de nodos.

A Figura 41 mostra o número médio de OIs recuperados durante a execução do SA variando-se os parâmetros limiar de disparo, fator de decaimento e número máximo de nodos por pulsos. O comportamento de tal métrica foi análogo à média do tempo de CPU gasto para executar o SA. O número de objetos recuperados aumenta com o crescimento do fator de decaimento e diminuição do limiar de disparo. Novamente, o número máximo de nodos por pulso não influenciou nos resultados e o número médio de OIs recuperados manteve-se estatisticamente estável para os diferentes valores escolhidos para este parâmetro.

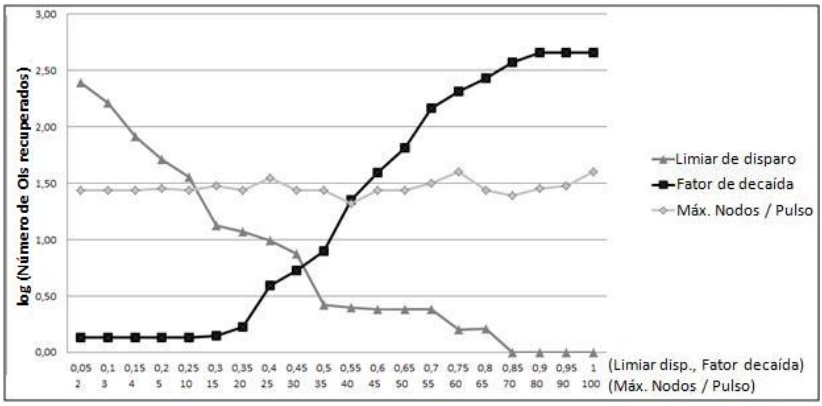


Figura 41: Número médio de objetos recuperados nas execuções do SA com variações nos parâmetros.

Estes resultados mostram que o tempo de execução e o número de objetos recuperados começam a crescer exponencialmente para valores do fator de decaimento além de 0,55 e 0,50, respectivamente. Por outro lado, estas métricas também começam a crescer exponencialmente para valores de limiar de disparo inferiores a 0,25 e 0,35, respectivamente. O melhor compromisso entre tempo de execução e o número de objetos recuperados foi empiricamente obtido com valores de fator de decaimento e limiar de disparo situados no intervalo [0.45, 0.65].

8.4.2 Desempenho com recuperação

Doravante os testes de desempenho com o módulo de recuperação são relatados. Primeiramente os testes de **carga** das redes e depois os de **execução** das consultas sobre elas.

Carga

Cada $RS(T, Col_i)$ é carregada a partir de um arquivo RDF. Inicialmente tentamos persisti-las em banco de dados, porém os modelos RDF persistidos pelo Jena em banco de dados não são lidos adequadamente pela própria ferramenta. Logo optamos pela abordagem de arquivos. A Figura 42 mostra a RAM exigida na carga de cada $RS(T, Col_i)$.

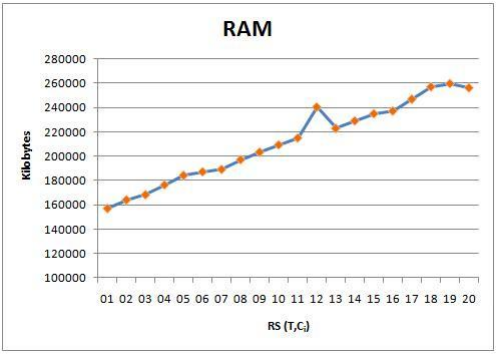


Figura 42: RAM consumida na carga de cada RS.

O crescimento do consumo de RAM é linear com o crescimento do tamanho dos arquivos RDF que armazenam cada RS, que é mostrado na Figura 43.

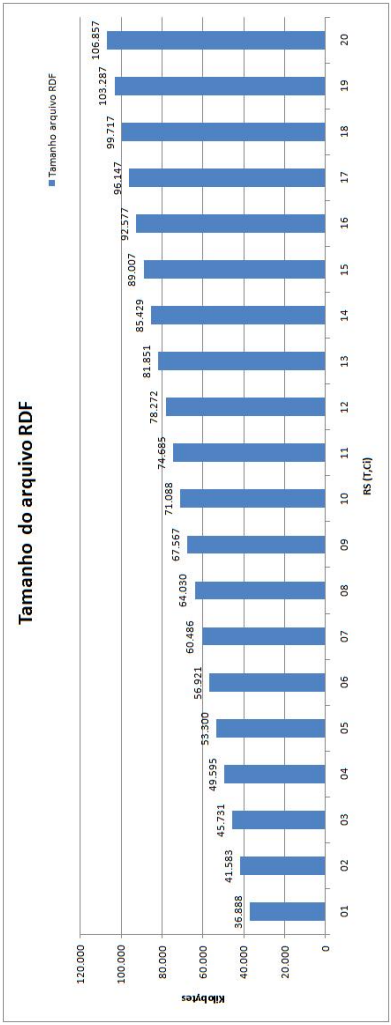


Figura 43: Tamanho do arquivo RDF de cada RS.

A Figura 44 mostra o tempo de CPU gasto na carga de cada $RS(T, Col_i)$.

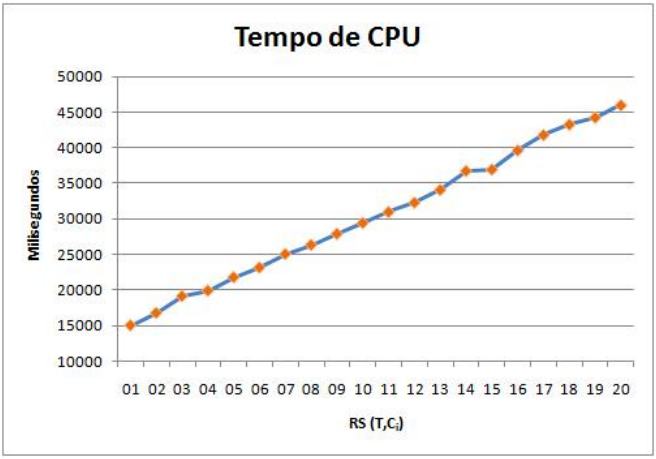


Figura 44: Tempo de CPU gasto na carga de cada RS.

A Figura 45 mostra o tempo gasto na carga de cada $RS_i(T, Col_i)$.

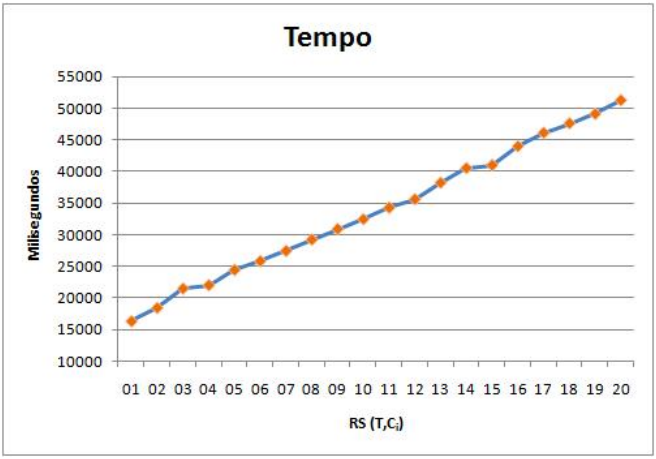


Figura 45: Tempo gasto na carga de cada RS.

Como esperado, mais recursos computacionais são exigidos à medida que RS maiores são carregadas. Percebe-se esta tendência de crescimento tanto no tempo gasto quanto no consumo de RAM.

Execução

A Figura 46 mostra o tempo gasto para a execução do SA de acordo com cada $RS_i(T, Col_i)$. As consultas da família F_1 exigiram mais tempo, principalmente as compostas por 3 a 5 palavras-chave, as quais iniciam a propagação do SA sobre termos genéricos como "Anatomia", "Doenças" e "Organismos". Tais termos encabeçam categorias enormes do VC DeCS, as quais estão altamente associadas a outras. Logo a ativação se espalha por uma grande região da RS, consumindo mais tempo. Já o tamanho das $RS(T, Col_i)$ parece não influenciar no tempo de execução. O tempo gasto segue a mesma tendência para todas as $RS_i(T, Col_i)$.

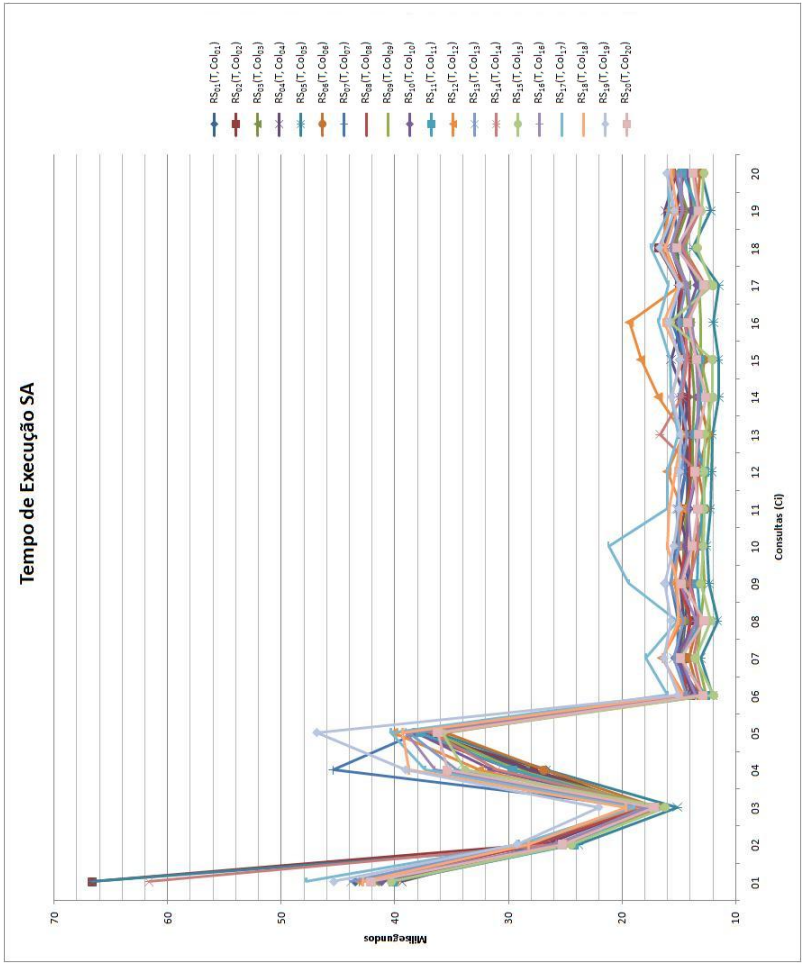


Figura 46: Tempo gasto na execução do SA sobre cada RS.

A Figura 47 compara a quantidade de OIs recuperados na execução do SA em cada $RS_i(T, Col_i)$. Para fins de visualização, omitimos as colunas das $RS_i(T, Col_i)$ onde a recuperação de OIS foi infrutífera. $RS_{20}(T, Col_{20})$ destacou-se na quantidade de OIs recuperados: foram 68 via consulta c_4, a mais pujante.

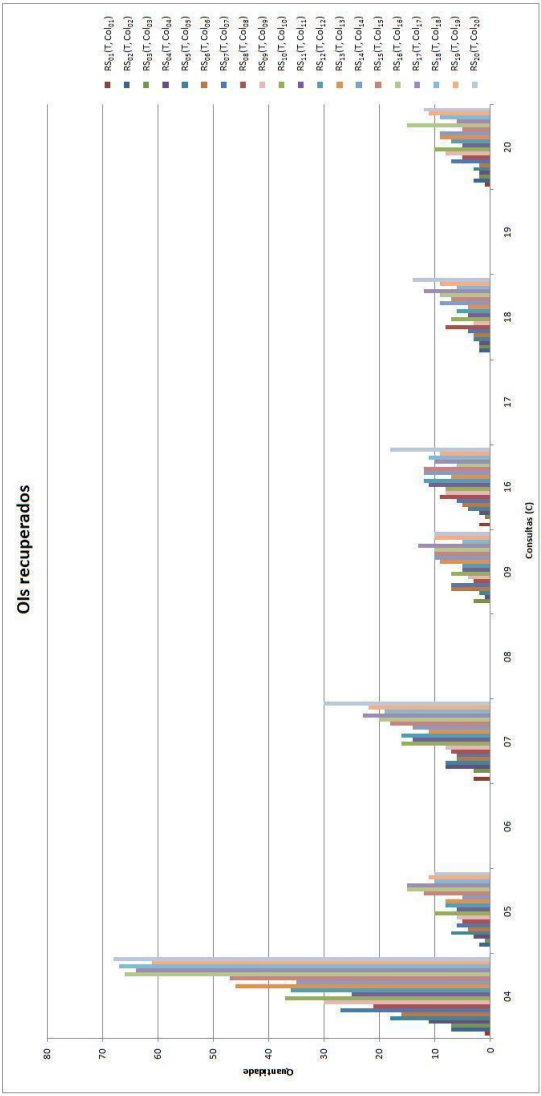


Figura 47: Quantidade de OIs recuperados na execução do SA sobre cada RS.

A Figura 48 mostra a quantidade de OIs recuperados pela consulta c_4 para as diversas RS. A medida que cresce o tamanho das $RS_i(T, Col_i)$, mais OIs são recuperados pela mesma consulta. Isso é esperado já que mais OIs podem estar ligados ao nodos ativados pelas ondas de propagação do SA.

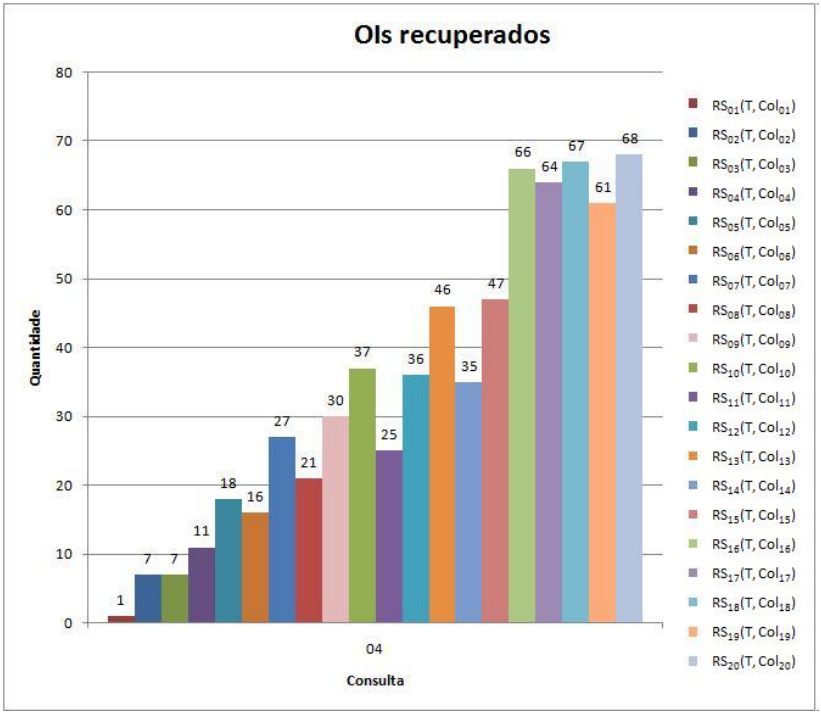


Figura 48: Quantidade de OIs recuperados pela consulta c_4.

A Figura 49 mostra a quantidade de nodos ativados após a execução de cada consulta no SA. Para todas as $RS_i(T, Col_i)$ a quantidade é a mesma. O que influencia aqui são as palavras-chaves (sementes) que disparam as consultas. Note que c_5 se destaca, donde 2214 nodos foram ativados.

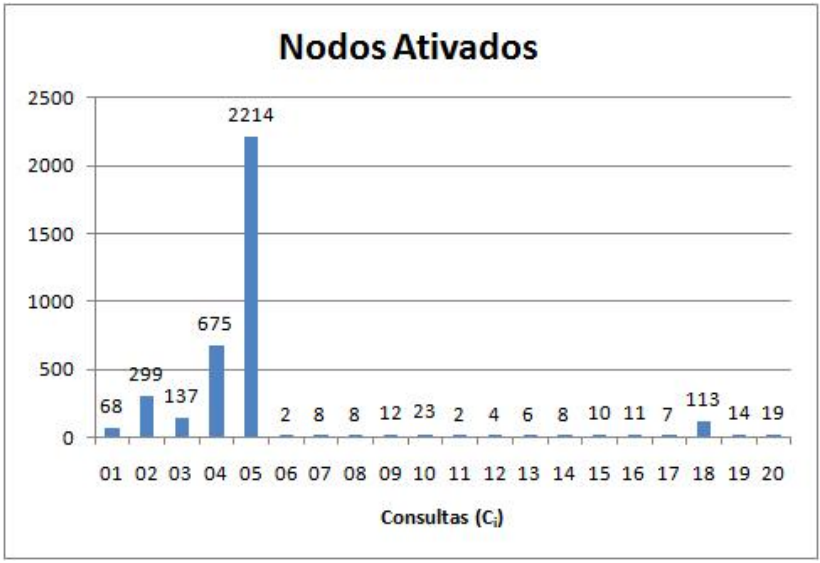


Figura 49: Quantidade de nodos ativados por C_i.

Neste capítulo abordamos os testes de usabilidade realizados com usuários, testes com diversas configurações de parâmetros e os testes de desempenho que avaliam o módulo de recuperação semântica implementado a partir de nossa proposta. Também analisamos os resultados obtidos em tais testes.

9 TRABALHOS RELACIONADOS

Apresenta-se aqui um conjunto de trabalhos que nos serviram de inspiração. Também confrontamos nosso trabalho com outras pesquisas que versam sobre anotação e recuperação semântica de informações.

9.1 Anotação de objetos

Segundo (MATHES, 2004), a criação de metadados para catalogação de recursos geralmente tem sido abordada de duas formas: criação por profissionais da informação e criação por autores. Porém há uma terceira abordagem, na qual os usuários criam os seus próprios metadados para os seus recursos de informação (MATHES, 2004). Neste contexto surgiu recentemente o fenômeno da Web 2.0 (Web social), onde os usuários associam livremente palavras-chave a recursos para indexá-los. Thomas Vander Wal (WAL, 2007b) cunhou o termo folksonomia, para definir a técnica de escolha de palavras-chave de forma livre e de acordo com a linguagem, conhecimento, interesse, opinião ou interpretação do usuário em relação ao recurso em marcação. Entre os pontos negativos de folksonomias estão sua natureza não controlada e fundamentalmente caótica, acarretando problemas de imprecisão e ambigüidade, que VCs bem desenvolvidos conseguem tratar (MATHES, 2004).

Em (ECHARTE et al., 2007) também são citados inconvenientes de folksonomias para anotação de recursos, tais como problemas ligados a polissemia, sinonímia, granularidade e tempos verbais. Tais problemas se devem ao fato das palavras-chave serem definidas livremente e à falta de estruturação das folksonomias, que não incluem hierarquias nem relações entre os metadados. Por outro lado, folksonomias dão liberdade para os usuários construírem o seu próprio vocabulário e descentralizam o controle de um sítio Web (PETERS; BECKER, 2009), livrando a administração do sítio da tarefa de classificar o grande volume de recursos publicados.

O trabalho de (DAHL; VOSSEN, 2008) adota tal abordagem para promover a anotação e recuperação de recursos educacionais (OAs). Segundo os autores, o conceito de *tagging* social na criação de metadados ainda não foi examinado no contexto de repositórios para *e-learning*. Eles pregam que tal abordagem oferece uma grande potencial que pode também ser interessante para a descoberta de OAs. Porém não realizaram experimentos com usuários comprovando isso. Em vista de todos os inconvenientes levantados, em nosso

trabalho, optamos conduzir a anotação de OAs de forma comportada, dirigida a termos de VCs.

Em (KIRYAKOV et al., 2004) os autores apresentam sua visão sobre um sistema holístico que permite anotação, indexação e recuperação de documentos em relação a entidades do mundo real. O KIM implementa parcialmente este conceito e é usado para avaliação e demonstração. Ele disponibiliza interfaces para anotação semântica de documentos via *pluggin* acoplado ao Internet Explorer. É poderoso e robusto ao implementar uma arquitetura que permite navegação desde as anotações até uma base de conhecimento, além de reconhecer e extrair de documentos entidades com correspondência em sua ontologia. Porém testes de usabilidade com usuários não foram conduzidos. À medida que a ontologia cresce, a navegação em árvore hierárquica pode não ser suficiente para a busca e visualização de seus conceitos. Além disso, a dependência do KIM ao navegador limita o uso de tal sistema.

9.1.1 Anotação no DSpace

Em (DUPRIEZ; SCHUBNEL, 2009) são relatados esforços para dotar o DSpace de recursos similares aos que propomos neste trabalho, para anotação e busca de objetos musicais, especificamente partituras e gravações. O preenchimento de certos campos de metadados foi implementado em (DUPRIEZ; SCHUBNEL, 2009) com recursos de autocompletar. Os valores de preenchimento são oriundos de um tesauro. Há também recursos para navegação em árvore hierárquica dotada de mecanismo de filtro via seleção de conceitos. A hierarquia proporciona ainda uma visão geral do conteúdo do repositório. Porém nem todas essas características estão contempladas no repositório disponibilizado em http://www.windmusic.org/dspace, que ainda é baseado na versão 1.4.2 do DSpace, de Maio de 2007. Além disso, os autores não conduziram testes com usuários. Nosso sistema ainda não tem suporte a vários idiomas, como o de (DUPRIEZ; SCHUBNEL, 2009), porém pode ser expandido para acomodar tal característica, inclusive usando versões do DeCS disponíveis em inglês e espanhol. Outros trabalhos, como (SHRIVASTAVA; SHUKLA; VIJAIANAND, 2009) abordam somente a expansão de funcionalidades do DSpace como navegação em coleções e buscas, sem usar conhecimento específico de domínio.

9.2 Visualizações de conhecimento

Silva (SILVA, 2007) analisa três interfaces baseadas em hipertexto para organização e representação de informação: (1) diagramas hierárquicos; (2) mapas conceituais e (3) mapas hiperbólicos. O autor também aborda como aspectos dessas interfaces influenciam na navegação e na recuperação da informação e levanta parâmetros para análise de interfaces, os quais ajudaram a definir os quadros comparativos apresentados na seção 4.3.1 deste trabalho.

(JUDELMAN, 2004) explora a natureza e a estrutura da informação e do conhecimento, os pontos fortes e limitações do sistema cognitivo e do sistema perceptivo humano, o trabalho do conhecimento no contexto social e discurso visual, além da definição e avaliação de metáforas de interface. Ele apresenta o estado da arte em estratégias de visualização e preconiza que "o grande desafio hoje não é, necessariamente, produzir novos conhecimentos, mas desenvolver modos de melhor trabalhar com ele e dar sentido àquele conhecimento que nós já possuímos". Em vista disso, uma estratégia para se redescobrir conhecimentos já aceitos, bem como facilitar a exploração de novos conhecimentos pode se dar via a apresentação adequada deste conhecimento. Isso colaborou para o estabelecimento do propósito deste trabalho e dos critérios para comparação de métodos e técnicas de visualização.

O trabalho de (KATIFORI et al., 2007) estuda métodos e técnicas de visualização de ontologias e categoriza suas características a fim de apoiar a seleção da técnica mais adequada, além de promover pesquisas. Em (FREITAS et al., 2001) os autores apresentam uma introdução à visualização de informações, abordando aspectos considerados fundamentais e técnicas que ilustram esses aspectos. Eles ainda comentam importantes características interdisciplinares dessa área. Em (SOUZA, 2007), a autora também faz um comparativo sucinto entre ferramentas e aponta a *Prefuse* para a implementação do mapa de um site.

9.3 Buscas associativas e SA

No início desta pesquisa, cogitou-se o uso de Redes de Bayesianas (JENSEN, 1996). Segundo (DECAMPOS; FERNáNDEZ-LUNA; HUETE, 2004), elas são modelos probabilísticos no qual o conhecimento representado é expresso em termos de relações de dependência e independência de variáveis. Uma rede Bayesiana consiste de uma parte qualitativa, um grafo acíclico dirigido, e uma quantitativa, uma coleção de parâmetros numéricos, normalmente as tabelas de probabilidade condicional. Se aplicam a elementos da RI onde o fator

incerteza existe. Por exemplo, as consultas são apenas representações aproximadas das necessidades de informação dos usuários. Redes de Bayesianas apresentam inconvenientes: (i) demasiado tempo para avaliar as distribuições (o número de probabilidades condicionais por nodo é exponencial com o número das arestas que chegam nos nodos), (ii) espaço exigido para armazenar a tabela de probabilidade condicionais e (iii) ineficiência para executar as inferências porque, em geral, inferência em redes Bayesianas é um problema NP-Árduo *(NP-hard)*(COOPER, 1990). Além disso, tal abordagem é desaconselhada para ambientes interativos e que comportam grandes coleções. Tais considerações desestimularam o emprego de Redes de Bayesianas.

Como o SA já estava disseminado entre membros do nosso grupo de pequisa, focamos as pesquisas em trabalhos afins. Em (CRESTANI, 1997) é apresentado um *survey* do SA, o modelo original, origem histórica e exemplos de sistemas implementados, como o GRANT, um sistema que utilizava o SA para obter correlacionamentos entre agências e tópicos de pesquisa, sendo este o primeiro sistema a apresentar o problema da configuração dos parâmetros do SA. Com tal *survey* e o trabalho de (TSATSARONIS; VAZIRGIANNIS; ANDROUTSOPOULOS, 2007), que implementa uma SAN a partir de um thesauro foi possível entender como funciona o SA e como definir uma SAN para este algoritmo. Nosso trabalho utiliza as definições de RS os cálculos do SA e adapta estas definições para o contexto de um estudo de caso na área da saúde, com um VC desta área.

A proposição de um mecanismo de buscas híbrido é apresentada por (ROCHA; SCHWABE; ARAGãO, 2004), de maneira a inspirar a busca composta. Outros trabalhos relacionados ao SA, porém com focos distintos do nosso são vistos a seguir. O trabalho de (D'AGOSTINI; FILETO, 2009), que emprega *Relevance Feedback* para capturar o contexto semântico do usuário, ou seja, seu contexto relativo a uma ontologia. O contexto semântico é representado como uma visão da ontologia subjacente. Ele é usado para semanticamente expandir e desambiguar consultas com SA, de acordo com o conhecimento sobre as preferências do usuário coletados em interações anteriores com o sistema de RI. No entanto, a análise do desempenho desta abordagem se concentra em sua precisão e revocação, quando aplicado para a recuperação de artigos da Wikipédia anotados com assuntos da DBpedia. O de (NILAS; NILAS; MASAKUL, 2007), que usa o SA para implementar recomendações em um sistema de *e-commerce*. O de (ASWATH et al., 2005), que usa o SA para identificar termos que afetam positivamente e termos que afetam negativamente um determinado documento, de maneira a treinar um algoritmo de inteligência artificial.

Neste capítulo fizemos um apanhado de pesquisas afins ao tema deste trabalho. Em suma, os demais trabalhos carecem de testes com usuários e testes em larga escala, como mostra a Tabela 13:

Trabalho	**Usa conhecimento de domínio**	**Testado com usuários**	**Testado em larga escala**	**Suporta Anotação**	**Suporta Recuperação**
(DUPRIEZ; SCHUBNEL, 2009)	X			X	X
(KIRYAKOV et al., 2004)	X			X	X
(DAHL; VOSSEN, 2008)				X	X
(D'AGOSTINI; FILETO, 2009)	X	X			X
CIBELE	X	X	X	X	X

Tabela 13: Trabalhos relacionados.

10 CONCLUSÃO

Este capítulo conclui o trabalho, enumerando nossas contribuições e apontando temas para futuras pesquisas.

Este trabalho propõe o uso de visualizações de conhecimento de domínio para apoiar a anotação e a especificação de consultas via seleção de palavras-chaves de um VC. Nosso módulo de recuperação é baseado em conhecimento e vale-se das palavras-chaves fornecidas para, via SA, explorar as relações semânticas estabelecidas na ontologia e aquelas usadas para anotar os OIs.

Testes de usabilidade realizados com usuários em um estudo de caso avaliaram as interfaces para anotação desenvolvidas. Mostramos ganhos no tempo de anotação pelo uso do suporte semântico proposto. Esses testes também indicaram facilidade de uso e viabilidade da proposta no domínio considerado, além da predileção dos usuários por uma interface baseada em autocompletar léxico, ao invés de navegação em árvores.

Testes com diferentes configurações de parâmetros e testes de carga com o módulo de recuperação avaliaram o desempenho e a escalabilidade do SA. Eles mostraram que o SA pode ser empregado em grandes coleções sem perda de *performance*. As configurações dos parâmetros que regem sua execução é que determinam um espalhamento comportado das ondas de ativação. Assim, o crescimento das coleções pode ser administrado com parâmetros de configuração adequados, como aqueles sugeridos para o domínio da saúde no estudo de caso. Tais valores podem ser descobertos com o amparo do protótipo desenvolvido.

Os objetivos traçados inicialmente foram totalmente alcançados de modo que as principais contribuições deste trabalho são:

i) uma análise comparativa de técnicas e ferramentas para a visualização de conhecimento;

ii) o desenvolvimento de um arcabouço de componentes Web para visualização de conhecimento que engloba árvores hierárquicas e hiperbólicas, além de um componente com autocompletar dinâmico para operar buscas por valores de metadados em VCs;

iii) avaliação de um sistema que emprega tais recursos de visualização de conhecimento na anotação de OIs em um estudo de caso com relevante

aplicação na área da saúde;

iv) uma proposta para a gerência de conteúdo de SRI baseada em visualizações do conhecimento;

v) uma proposta de arquitetura para SRI baseados em conhecimento e visualização para a anotação, gerência, e recuperação de OIs;

vi) um modelo de RS que permite a associação de diferentes pesos para cada tipo de relação semântica, a fim de enfatizar a expansão semântica em função da importância relativa de cada tipo de relacionamento;

vii) desenvolvimento de um módulo de *software* para a execução do SA em qualquer RS construída de acordo com o modelo proposto, permitindo diferentes configurações de valores para os parâmetros do SA;

viii) identificação de correlações entre os parâmetros de configuração e sugestão de intervalos de valores para os parâmetros de sintonia do SA, a fim de se reduzir o tempo de execução e gerar listas de resultados ordenados com tamanho razoável;

ix) resultados experimentais que ilustram como o tamanho das coleções de OIs processadas pelo SA influenciam no consumo de recursos computacionais.

As implementações foram validadas junto a usuários do domínio do estudo de caso, onde aferiu-se a potencialidade da nossa proposta. Os resultados foram ilustrados em um estudo de caso na área de saúde, mas podem ser reproduzidos em outros domínios para os quais haja conhecimento formalizado disponível que possa ser adaptado para utilização no sistema proposto.

Nossas idéias foram bem aceitas de modo que elas resultaram nas seguintes publicações:

a) "Performance Evaluation and Tuning of Spreading Activation for Associative Information Retrieval" (FILETO et al., 2011) que foi aceito no *International Association for Development of the Information Society on WWW/Internet 2011* (ICWI 2011);

b) "Anotação de Conteúdo Multimídia em Repositórios com Interfaces Web baseadas em Conhecimento de Domínio" (RIGO et al., 2011), que foi publicado no Simpósio Brasileiro de Sistemas Multimídia e Web (WEBMEDIA) e discorre sobre um sistema para apoiar a anotação e a

navegação semântica sobre o conteúdo de repositórios na Web, donde testes de usabilidade com usuários indicaram ganhos consideráveis no tempo de anotação.

c) "Interfaces Web baseadas em Conhecimento para Anotação de Recursos de Informação e Gerenciamento de Repositórios"(RIGO et al., 2010), que foi publicado no XXI Simpósio Brasileiro de Informática na Educação (SBIE) e propõe o uso de técnicas de visualização de conhecimento para apoiar a seleção de metadados na descrição de recursos de informação e para analisar o conteúdo de repositórios;

Até a data de edição deste trabalho, ainda aguardávamos o resultado da seguinte submissão:

- "Scalability Evaluation of an Information Retrieval System using Spreading Activation" que foi submetido em Setembro de 2011 ao *27th Symposium On Applied Computing (ACM SAC 2012)*[1] e avalia o desempenho computacional do SA atuando sobre grandes coleções de OIs.

10.1 Trabalhos Futuros

Esta pesquisa também desvelou outras idéias promissoras e que podem ser exploradas em trabalhos futuros:

- Componentizar as interfaces baseadas em conhecimento de modo que estas possam ser reutilizadas por outras aplicações. Hoje tais interfaces são módulos Web que se integram ao SRI via Javascript mas podem ser implementadas como *Web Service* para minimizar o acoplamento com a aplicação.

- A partir do processamento sintático do conteúdo do OI que está sendo descrito na catalogação, pesquisar no VC e sugerir os termos mais afins que poderiam anotar tal OI.

- Realizar testes com a técnica de recuperação proposta em diferentes domínios visando-se maximizar a precisão e a cobertura dos resultados retornados.

[1] http://www.acm.org/conferences/sac/sac2012

- Testar a hipótese de enriquecimento colaborativo da base de conhecimento baseado na co-ocorrência de palavras-chave na descrição de objetos ou na especificação de consultas (CHEN; QIN, 2008), visando melhorar o desempenho das buscas.

- Usar informação do perfil semântico do usuário, i.e., do seu contexto relativo ao conhecimento de domínio, colhida através de *relevance feedback*, para aperfeiçoar a recuperação e o ordenamento dos resultados das buscas (D'AGOSTINI; FILETO, 2009).

- Disponibilizar o módulo de buscas semânticas via *Web Service*. Então outras aplicações da área da saúde poderiam fazer uso de nossa técnica de processamento semântico. Detalhes no Anexo L;

- Aprimorar um *front-end* Web que vale-se das interfaces baseadas em conhecimento para especificação de consultas, as quais são processadas pelo módulo de recuperação semântica desenvolvido. Detalhes no Anexo M.

ANEXO A – DETALHES DE IMPLEMENTAÇÃO DAS INTERFACES

A.1 Interface Hiperbólica

É uma aplicação Web que foi concebida segundo a estrutura MVC *(Model-View-Controller)*[1]. Deve rodar no mesmo contexto onde a aplicação que a utiliza está rodando.

Visão geral

A Figura 50 ilustra os bastidores do funcionamento da interface Hiperbólica.

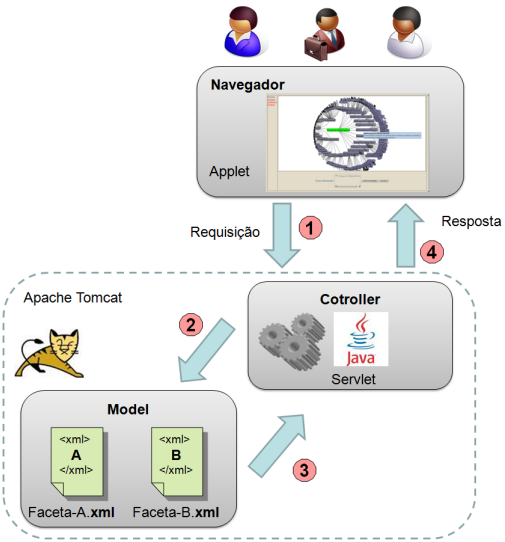


Figura 50: Arquitetura da interface Hiperbólica.

Por meio de um navegador usuários enviam requisições (1) que chegam até o *Controller*, módulo que abriga o Java Servlet[2] da aplicação. Nessas requisições o usuário pode definir que faceta (ou categoria) do VC ele quer visualizar. O *Controller* recebe as requisições e as trata, encaminhando

[1] http://java.sun.com/blueprints/patterns/MVC-detailed.html
[2] http://www.oracle.com/technetwork/java/javaee/servlet/index.html

as solicitações de facetas ao *Model* (2), módulo que abriga os provedores de dados. Então o *Model*, sabendo qual foi a faceta escolhida, seleciona o provedor de dados adequado e o encaminha ao *Controller* (3). De posse do provedor de dados, o *Controller* retorna a resposta ao navegador do usuário (4), o qual abriga a interface Hiperbólica, que é implementada via um applet[3] que suporta as funcionalidades da biblioteca *Treebolic*.

Geração do XML para o applet da interface Hiperbólica

O applet previamente mencionado necessita de um provedor de dados. Nossa implementação faz uso de arquivos XML[4]. Nestes arquivos estão os dados das facetas que podem ser exibidos na interface Hiperbólica. A Figura 51 explicita como estes arquivos XML são gerados.

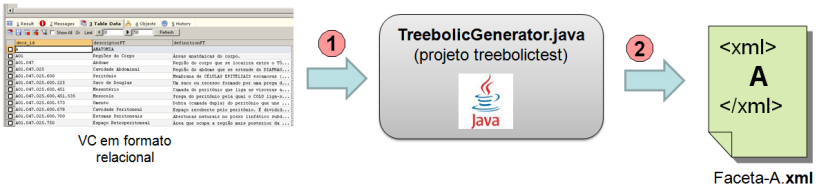


Figura 51: Geração do XML de entrada do applet da interface Hiperbólica.

Os dados das facetas são extraídos do VC que foi salvo em um banco de dados MySQL[5] em formato relacional (1). Eles são tratados por um utilitário Java[6] especialmente desenvolvido (2). Baseado em informações presentes no vocabulário controlado, recursivamente o utilitário recria em memória a estrutura de árvore que representa a faceta do VC. Num segundo passo, ele gera um arquivo XML que reflete a estrutura hierárquica da faceta do VC. Tal XML obedece ao *schema* XML que o applet implementado pela biblioteca *Treebolic* espera. Então este arquivo XML pode alimentar o applet, o qual gera a visualização dos dados da faceta do VC no formato hiberbólico.

[3] http://java.sun.com/applets
[4] http://www.w3.org/XML
[5] http://www.mysql.com
[6] http://www.oracle.com/technetwork/java/index.html

A.2 Interface Hierárquica

É uma aplicação Web que também foi concebida segundo a estrutura MVC *(Model-View-Controller)* e faz uso da biblioteca *Prefuse*.

Visão geral

A Figura 52 esboça a arquitetura da interface Hierárquica.

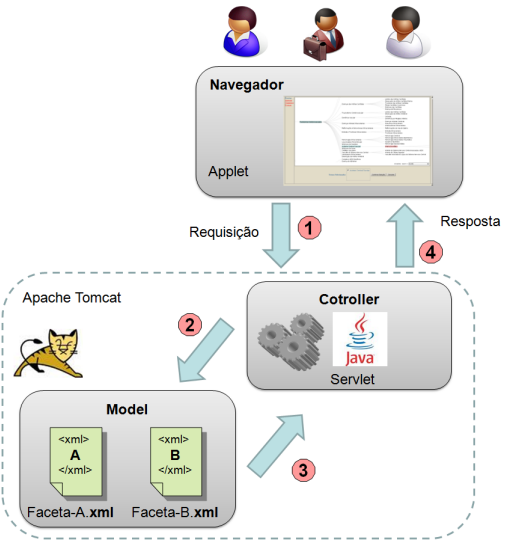


Figura 52: Arquitetura da interface Hierárquica.

Por meio de um navegador usuários enviam requisições (1) que chegam até o *Controller*, módulo que abriga o Java Servlet da aplicação. Nessas requisições o usuário pode definir que faceta ele quer visualizar. O *Controller* recebe as requisições e as trata, encaminhando as solicitações de facetas ao *Model* (2), módulo que abriga os provedores de dados. Então o *Model*, sabendo qual foi a faceta escolhida, seleciona o provedor de dados adequado e o encaminha ao *Controller* (3). De posse do provedor de dados, o *Controller* retorna a resposta ao navegador do usuário (4), o qual abriga a interface Hierárquica, que é implementada via um applet que suporta as funcionalidades da biblioteca *Prefuse*.

Geração do XML para o applet da interface Hierárquica

O applet previamente mencionado necessita de um provedor de dados. Nossa implementação faz uso de arquivos XML. Nestes arquivos estão os dados das facetas que podem ser exibidos na interface Hierárquica. A Figura 53 explicita como estes arquivos XML são gerados.

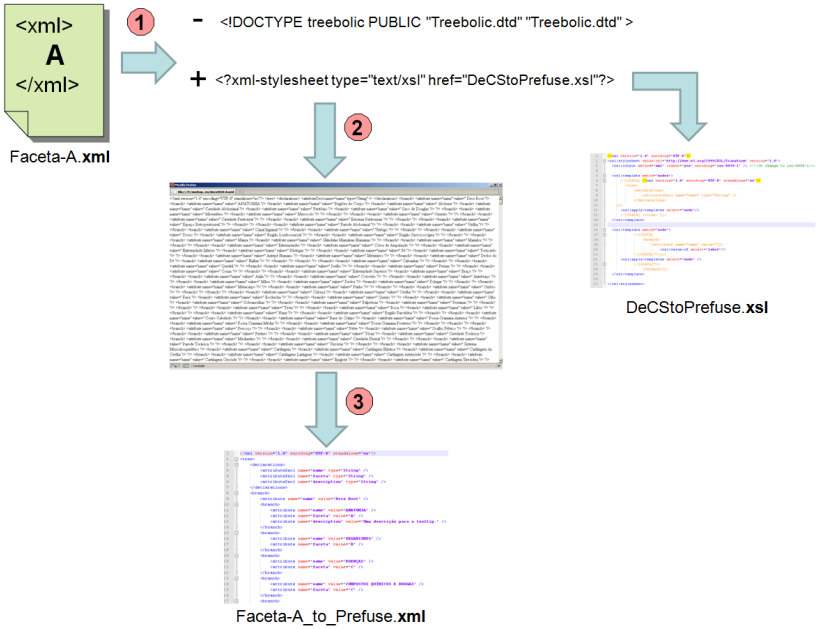


Figura 53: Geração XML de entrada do applet da interface Hierárquica.

A partir do XML de cada faceta (1) obtido via processo descrito na Figura 51, uma transformação via folha de estilo XSL (XML Style Sheet)[7] é aplicada (2), donde obtemos um outro arquivo em XML já de acordo com o schema que o *Prefuse* espera (3).

[7] http://www.w3.org/Style/XSL

A.3 Interface Autocompletar

É uma aplicação Web que, a exemplo das outras, também foi concebida segundo a estrutura MVC *(Model-View-Controller)*.

Visão geral

A Figura 54 ilustra a o fluxo de dados que ocorre por entre as camadas do MVC. As requisições disparadas pelos usuários através de algum navegador (1) chegam até o *Controller*, módulo que abriga o Java Servlet da aplicação. Ele recebe a requisição e a trata, encaminhando a solicitação ao *Model* (2), módulo que abriga a ferramenta Lucene. O Lucene é o responsável por apurar os resultados da busca, os quais são encaminhados ao *Controller* (3). De posse dos resultados, o *Controller* envia os dados para o navegador do usuário, onde eles são formatados e exibidos adequadamente (4).

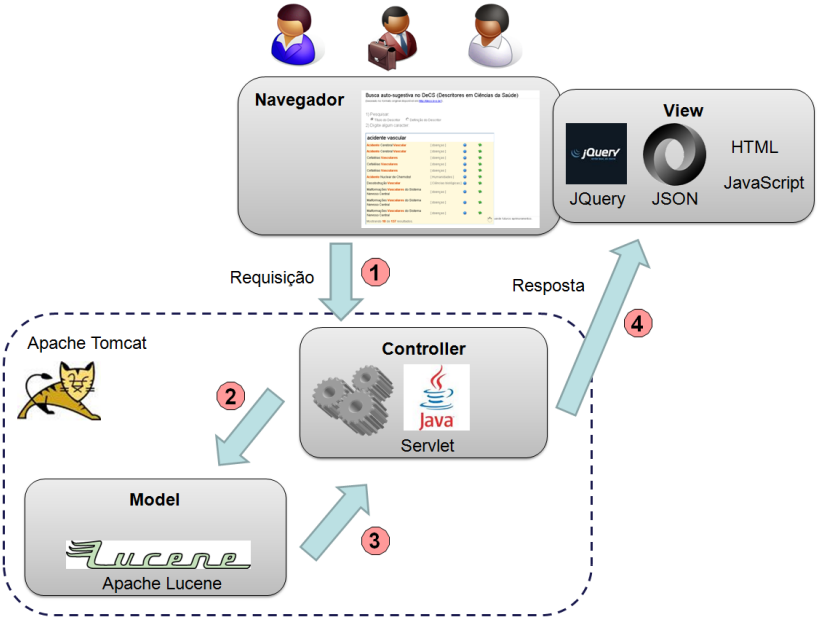


Figura 54: Arquitetura da interface de sugestão e autocomplemento.

Detalhes da busca e indexação

Na inicialização do Java Servlet, o Lucene indexa 30.298 arquivos que contém um extrato de cada conceito presente no VC. Chamaremos cada um desses arquivos de "sumário de um conceito". Cada sumário de um conceito apresenta a estrutura que pode ser vista na Figura 55:

caregoryId=C14.907.253.855

descriptorName=Acidente Cerebral Vascular

definition=Grupo de condições patológicas caracterizadas por perda súbita, não-convulsiva, da função neurológica, devido a ISQUEMIA ENCEFÁLICA ou HEMORRAGIAS INTRACRANIANAS. O acidente cerebral vascular é classificado pelo tipo de NECROSE de tecido, como localização anatômica, vasculatura envolvida, etiologia, idade dos indivíduos afetados e natureza hemorrágica versus não-hemorrágica (Tradução livre do original: Adams et al., Principles of Neurology, 6th ed, pp777-810).

synonym=Acidente Vascular Cerebral; Acidente Vascular do Cérebro; Acidente Vascular Encefálico; Icto Cerebral; Apoplexia Cerebrovascular; Apoplexia Cerebral; Acidente Cerebrovascular; Apoplexia; AVC; Ictus Cerebral; Derrame Cerebral

Figura 55: Exemplo de um sumário de um conceito.

Na indexação realizada pelo Lucene é gerado um arquivo de índices que pode ser mantido na memória ou em disco. O índice faz o mapeamento entre cada termo relevante encontrado em certos trechos (descriptionName, definition, synonym) de cada sumário de conceito e os sumários dos conceitos que contém este dado termo. Este arquivo de índices é consultado pelo Lucene a cada requisição disparada via AJAX pela Interface Autocompletar. Quando ocorre correspondência sintática *(matching)* entre os caracteres digitados pelo usuário e algum termo do índice, os sumários dos conceitos apontados pelo índice são retornados. Este é o resultado da busca.

Geração dos sumários dos conceitos

A geração dos sumários dos conceitos demanda uma série de passos, os quais estão esboçados na Figura 56.

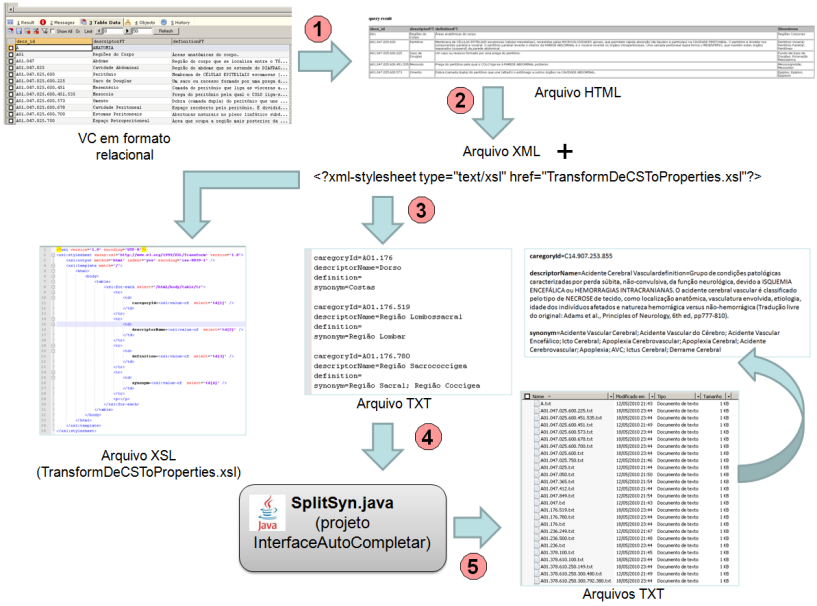


Figura 56: Processo de geração dos sumários dos conceitos.

Os dados são extraídos do vocabulário controlado que foi salvo em um banco de dados MySQL [8] em formato relacional. Eles são exportados em HTML[9] (1) pela ferramenta SQLYog [10]. O arquivo exportado é renomeado para XML e recebe em seu cabeçalho uma instrução para renderizar seu conteúdo usando uma folha de estilo XSL (XML Style Sheet)[11] (2). Esta folha de estilo irá reformatar os dados que estavam em formato tabular no HTML. Assim que este arquivo for aberto em um navegador (3), será possível ver que o novo formato impingido pelo XSL agrupa cada conceito do VC em pequenos fragmentos compostos pelas seguintes informações:

caregoryId: um identificador que o conceito recebe no VC.
descriptorName: o nome do conceito no VC.
definition: a definição do conceito, segundo o VC.

[8] http://www.mysql.com
[9] http://www.w3.org/MarkUp
[10] http://www.webyog.com
[11] http://www.w3.org/Style/XSL

synonym: os sinônimos do conceito, segundo o VC.

A partir do navegador, este arquivo é salvo no formato txt (4). Em seguida ele é processado por um utilitário Java especialmente desenvolvido (5). Ele quebra cada fragmento dos conceitos em arquivos que recebem o nome do **caregoryId** de cada conceito. Estes arquivos são alvo da indexação promovida pelo Lucene.

Componentes da aplicação

Afim de operacionalizar o funcionamento da interface Autocompletar, diversas tecnologias e componentes foram utilizados:

JavaScript[12]: é uma linguagem de script interpretada que adiciona comportamento dinâmico e interativo à paginas da Web.

AJAX(*Asynchronous Javascript And XML*)[13]: não é uma linguagem de programação, mas uma nova forma de usar o JavaScript. Compreende a técnica de intercâmbio de dados entre o cliente (Navegador) e um server (aplicação Web), atualizando somente as partes necessárias na página no cliente. Isso evita que todo o conteúdo da página tenha que ser renderizado a cada requisição.

JQuery[14]: é uma biblioteca JavaScript rápida e concisa que simplifica a travessia de documentos HTML, tratamento de eventos, animações e interações via AJAX.

JSON(*JavaScript Object Notation*)[15]: é utilizado principalmente para tráfego de informações em ambientes heterogêneos via HTTP. Além de diminuir o tráfego sobre o HTTP devido a sua sintaxe concisa, evita que dados em formato XML recebidos pelo navegador tenham que ser parseados para DOM (Document Object Model)[16] antes de serem renderizados na página.

Java Serlvet[17]: É um componente do lado servidor que gera dados

[12] http://www.w3schools.com/js/js_intro.asp
[13] http://www.w3schools.com/ajax/default.asp
[14] http://jquery.com
[15] http://www.json.org
[16] http://www.w3.org/DOM
[17] http://www.oracle.com/technetwork/java/javaee/servlet/index.html

para a camada de apresentação de um aplicativo Web. É basicamente uma classe Java que dinamicamente processa requisições e gera respostas segundo a lógica implementada.

Apache Lucene[18]: o Lucene é uma das mais famosas bibliotecas de código aberto usadas para indexação e consulta de textos. É escrita em Java e pode processar todo o tipo de informação que possa ser convertido em texto. Atua em duas etapas: indexação e pesquisa. A indexação processa os dados textuais gerando um índice, que é uma estrutura de dados inter-relacionada eficiente para a pesquisa baseada em palavras-chave. A pesquisa, por sua vez, consulta o índice pelas palavras digitadas em uma busca e organiza os resultados pela similaridade do texto com a busca.

Apache Tomcat[19]: é um servidor Web para aplicações Java, mais especificamente, um *container* de servlets. Por um longo tempo foi a implementação de referência para as tecnologias Java Servlet[20] e JavaServer Pages (JSP)[21].

[18] http://lucene.apache.org/java
[19] http://tomcat.apache.org
[20] http://www.oracle.com/technetwork/java/javaee/servlet/index.html
[21] http://www.oracle.com/technetwork/java/javaee/jsp/index.html

ANEXO B – GERAÇÃO DA RS

A geração da RS é ilustrada pela Figura 57. A partir do VC armazenado em um banco de dados relacional (1), um utilitário Java especialmente desenvolvido lê as informações sobre os termos e os sinônimos dos termos e então valendo-se do *framework* semântico Jena [1] gera um modelo em RDF (2). Tal modelo pode ser salvo em um arquivo .rdf (3) ou em um banco de dados relacional, como o MySQL (4). Este modelo RDF salvo é a RS.

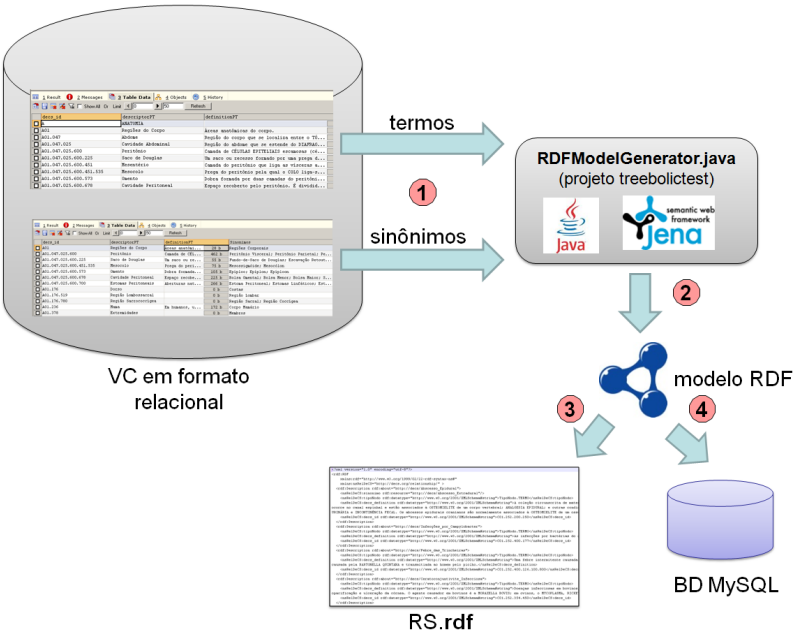


Figura 57: Geração da RS em RDF.

[1] http://jena.sourceforge.net

ANEXO C – SIMILARIDADE ENTRE TERMOS DO VC

A geração de novas relações semânticas em uma representação do conhecimento requer o trabalho e a supervisão de especialista de domínio. Este trabalho apresenta uma proposta para facilitar esta tarefa, apontando possíveis novas relações descobertas através do cálculo da similaridade sintática entre a descrição dos termos de um VC.

Visão geral

O processo de cálculo da similaridade sintática entre termos de facetas do VC é esboçado na Figura 58:

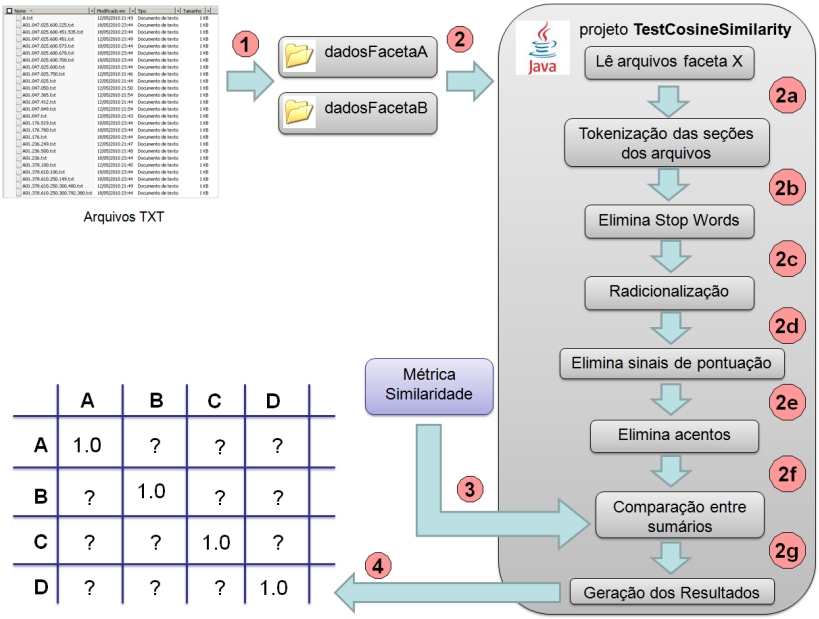


Figura 58: Processo de cálculo da similaridade sintática entre termos do VC.

Primeiramente, duas facetas (categorias) do VC são escolhidas (1). Então para cada um dos termos de tais facetas é montado um sumário (2). Tais sumários originam-se dos arquivos gerados pelo passo 5 da Figura 56 e são compostos das seguintes seções:

a) Rótulo: o termo que é usado para identificar um dado conceito no VC.

b) Definição: uma descrição textual provida pelo VC e que discorre sobre o conceito que tal termo denota.

c) Sinônimo(s): os eventuais sinônimos definidos pelo VC para um dado termo.

A Figura 59 mostra um exemplo de sumário. O termo em questão é "Acidente Cerebral Vascular":

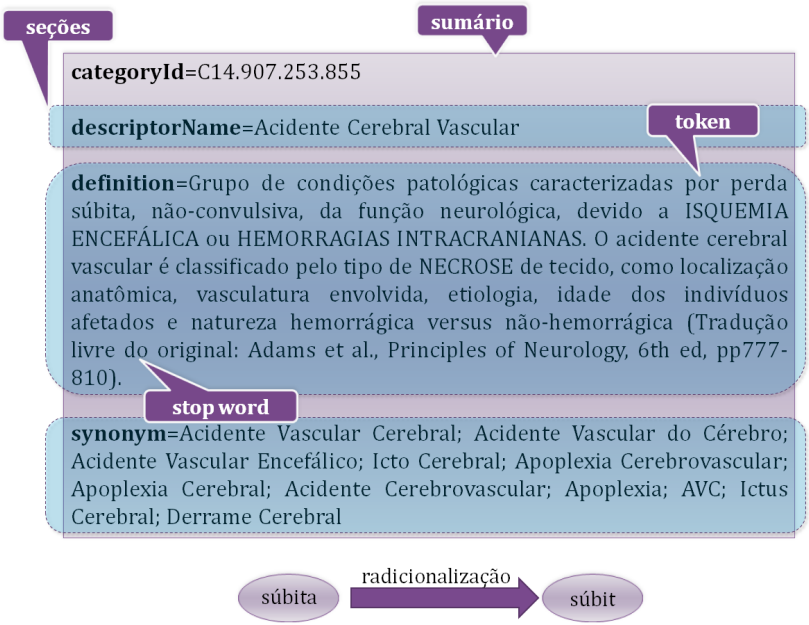


Figura 59: Exemplo de sumário.

À medida que os arquivos que armazenam o sumário de cada termo são lidos (2a), cada seção do sumário é quebrada em *tokens* (2b). Então

as *stop words* são eliminadas das seções (2c). *Stop words* são palavras que aparecem com freqüência em texto e não detém significado, logo podem ser consideradas irrelevantes. Exemplos: as, e, os, de, para, com, sem, foi. Em seguida os *tokens* restantes são radicionalizados por um *steammer* (2d), ou seja, apenas o radical de cada palavra é mantido. Em seguida, os sinais de pontuação (e.g. vírgulas, pontos, parênteses, conchetes) são retirados das sentenças (2e), assim como a acentuação (2f). Posteriormente, os sumários dos termos de cada uma das duas facetas são confrontados, sendo que as seções afins de cada sumário são emparelhadas (2g). Então uma métrica de similaridade (3), no caso, Similaridade Cosseno, é usada para calcular a similaridade sintática entre os radicais pertencentes às seções afins. Em seguida é calculada a Similaridade Resultante. Ela leva em conta a similaridade calculada para cada seção, ponderada por um peso:

Definição 11 - *Similaridade resultante:*

Seja A e B termos pertencentes a facetas distintas de um dado VC.
Seja x o valor da similaridade sintática entre as seções "Rótulo" dos termos A e B.
Seja y o valor da similaridade sintática entre as seções "Definição" dos termos A e B.
Seja z o valor da similaridade sintática entre as seções "Sinônimo(s)" dos termos A e B.
Seja X coeficiente importância da seção "Rótulo", sendo que $(0 <= X <= 1)$.
Seja Y coeficiente importância da seção "Definição", sendo que $(0 <= Y <= 1)$.
Seja Z coeficiente importância da seção "Sinônimo(s)", sendo que $(0 <= Z <= 1)$.

$$SimilaridadeResultante(A,B) = (x*X + y*Y + z*Z)/3$$

Por fim é gerada uma matriz que relaciona cada um dos termos das duas facetas e afere a Similaridade Resultante para cada inter-relacionamento entre os termos (4). Definido um valor de limiar para a Similaridade Resultante, pode-se dizer que os valores maiores que este limiar aferem que existe algum tipo de relação entre os termos de facetas distintas.

ANEXO D – CERTIFICADO EMITIDO PELO COMITÊ DE ÉTICA

UNIVERSIDADE FEDERAL DE SANTA CATARINA
Pró-Reitoria de Pesquisa e Extensão
Comitê de Ética em Pesquisa com Seres Humanos

CERTIFICADO **Nº 1827**

O Comitê de Ética em Pesquisa com Seres Humanos (CEPSH) da Pró-Reitoria de Pesquisa e Extensão da Universidade Federal de Santa Catarina, instituído pela PORTARIA Nº 0584/GR/99 de 04 de novembro de 1999, com base nas normas para a constituição e funcionamento do CEPSH, considerando o contido no Regimento Interno do CEPSH, **CERTIFICA** que os procedimentos que envolvem seres humanos no projeto de pesquisa abaixo especificado estão de acordo com os princípios éticos estabelecidos pela Comissão Nacional de Ética em Pesquisa – CONEP.

APROVADO

PROCESSO: 1827 **FR:** 399658

TÍTULO: Tecnologias de Informação aplicadas na Formação e Aperfeiçoamento de Profissionais da UnA-SUS - UFSC (Universidade Aberta do Sistema Único de Saúde - Universidade Federal de Santa Catarina)

AUTOR: Aldo von Wangenheim, Paulo Eduardo Battistella, Wanderson Rigo, Ronaldo Lima Rocha Campos, Vilmar César Pereira Júnior, Renato Fileto, Roberto Willrich

FLORIANÓPOLIS, 28 **de** Março **de** 2011.

Coordenador do CEPSH UFSC

ANEXO E – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO - TCLE

TCLE- TERMO DE CONSENTIMENTO LIVRE E ESCLARECIMENTO

Tecnologias de Informação aplicadas na Formação e Aperfeiçoamento de Profissionais da UnA-SUS - UFSC

(Universidade Aberta do Sistema Único de Saúde - Universidade Federal de Santa Catarina)

A presente pesquisa tem por objetivo analisar e avaliar os resultados tecnológicos aplicados nos cursos de especialização, modalidade a distância, enquadrados no contexto da Una-SUS/UFSC. Tendo como justificativa a necessidade de identificação por parte dos usuários dos pontos positivos e pontos a serem melhorados das tecnologias aplicadas no projeto Una-SUS/UFSC.

A pesquisa a ser realizada será a base para próximas implementações e modificações tecnológicas do projeto. Desta forma novos cursos poderão reutilizar o conhecimento adquirido até então.

A pesquisa tem como orientador o Dr. rer.nat. Aldo von Wangenheim e como participantes da pesquisa alunos de graduação, pós-graduação e professores do Curso de Ciências da Computação e Sistemas de Informação do departamento de Informática e Estatística (INE) da Universidade Federal de Santa Catarina.

Através deste termo são garantidos os direitos do participante de esclarecimento a dúvidas antes, durante e depois da pesquisa, assim como os diretos de sigilo e a privacidade de identidade dos participantes. Também é garantida a liberdade do participante de se recusar a participar, abandonar ou retirar seu consentimento, em qualquer etapa da pesquisa, sem penalidade e sem prejuízo ao seu cuidado.

DATA: _____

_____

Nome

_____

Assinatura

ANEXO F – QUESTÕES QUE REFINAM OS OBJETIVOS DOS TESTES DE USABILIDADE.

Objetivo	Questão	Descrição	Métrica
01	Q01	Os usuários gostam de usar as interfaces baseadas em conhecimento?	M1, M2, M3, M4
01	Q02	Os usuários julgam ser fácil usar as interfaces baseadas em conhecimento?	M1, M2, M3, M4
01	Q03	Os usuários preferem qual interface baseada em conhecimento?	M5, M6, M7
01	Q04	Os usuários julgam que as interfaces baseadas em conhecimento são claras e fáceis de entender?	M1, M2, M3, M4
02	Q05	Os usuários julgam que as interfaces baseadas em conhecimento podem ser utilizadas para o preenchimento de outros campos?	M1, M2, M3, M4
02	Q06	Os usuários julgam que as interfaces baseadas em conhecimento podem ser aplicadas a outros domínios de aplicação?	M1, M2, M3, M4
03	Q07	Os usuários julgam que as interfaces baseadas em conhecimento facilitam a execução das tarefas de catalogação?	M1, M2, M3, M4

continua na próxima página.

Tabela 14 – continuação da página anterior.

Objetivo	**Questão**	**Descrição**	**Métrica**
03	Q08	Os usuários julgam que os termos providos pelas interfaces baseadas em conhecimento são suficientes para a descrição dos recursos?	M1, M2, M3, M4
03	Q09	Os usuários conseguem realizar as tarefas propostas valendo-se das interfaces baseadas em conhecimento?	M12, M13, M14, M15
04	Q10	Quanto tempo cada um dos avaliadores levou para descrever completamente (todos os campos) cada recurso?	M8
04	Q11	Quanto tempo cada um dos avaliadores levou para preencher as palavras-chave de cada recurso?	M9
04	Q12	Os usuários julgam que qual interface é a mais trabalhosa de se usar?	M5,M6,M7
04	Q13	Qual é o esforço empreendido para a execução de cada tarefa?	M16, M17
05	Q14	Sem o apoio das interfaces baseadas em conhecimento, quais foram os termos usados pelos avaliadores para descrever os recursos?	M10

continua na próxima página.

Tabela 14 – continuação da página anterior.

Objetivo	**Questão**	**Descrição**	**Métrica**
05	Q15	Com o apoio das interfaces baseadas em conhecimento, quais foram os termos usados pelos avaliadores para descrever os recursos?	M11
05	Q16	Sem o apoio das interfaces baseadas em conhecimento, quantos foram os termos do vocabulário controlado usados pelos avaliadores para descrever os recursos?	M20
05	Q17	Com o apoio das interfaces baseadas em conhecimento, quantos foram os termos do vocabulário controlado usados pelos avaliadores para descrever os recursos?	M20
05	Q18	Sem o apoio das interfaces baseadas em conhecimento, quantos erros de grafia ocorreram na descrição de recursos?	M19
05	Q19	Com o apoio das interfaces baseadas em conhecimento, quantos erros de grafia ocorreram na descrição de recursos?	M19
05	Q20	Sem o apoio das interfaces baseadas em conhecimento, quantos termos foram usados na descrição de recursos?	M18

continua na próxima página.

Tabela 14 – continuação da página anterior.

Objetivo	**Questão**	**Descrição**	**Métrica**
05	Q21	Com o apoio das interfaces baseadas em conhecimento, quantos termos foram usados na descrição de recursos?	M18

Tabela 14: Questões que refinam os objetivos dos testes de usabilidade com catalogação.

ANEXO G – MÉTRICAS QUE OPERACIONALIZAM AS QUESTÕES.

Métrica	Coletada via	Descrição
M01	Questionário	Número de avaliadores que escolheram “Discordo fortemente”.
M02	Questionário	Número de avaliadores que escolheram “Discordo”.
M03	Questionário	Número de avaliadores que escolheram “Concordo”.
M04	Questionário	Número de avaliadores que escolheram “Concordo fortemente”.
M05	Questionário	Número de avaliadores que escolheram “Autocompletar”.
M06	Questionário	Número de avaliadores que escolheram “Hierárquica”.
M07	Questionário	Número de avaliadores que escolheram “Hiperbólica”.
M08	Logging, Morae	Para cada recurso: (Horário Final Catalogação - Horário Início Catalogação)
M09	Logging, Morae	Para cada recurso: (Horário Final Descrição Palavra-Chave - Horário Início Descrição Palavra-Chave)
M10	Ficha Catalográfica	Termos usados para descrever os recursos manualmente.
M11	Logging, Morae	Termos usados para descrever os recursos valendo-se das interfaces baseadas em conhecimento.
M12	Questionário	Número de avaliadores que escolheram “Não consegui realizar”.
M13	Questionário	Número de avaliadores que escolheram “Realizei parcialmente”.
M14	Questionário	Número de avaliadores que escolheram “Realizei com dificuldade”.

continua na próxima página.

Tabela 15 – continuação da página anterior.

Métrica	**Coletada via**	**Descrição**
M15	Questionário	Número de avaliadores que escolheram "Realizei com facilidade".
M16	Morae	Número de cliques do mouse.
M17	Morae	Movimentos do mouse (distância percorrida).
M18	Logging, Ficha Catalográfica, Morae	Número de termos usados para descrever um dado recurso.
M19	Logging, Ficha Catalográfica, Morae	Erros de grafia encontrados nos termos usados pelos avaliadores para descrever um dado recurso. A grafia correta é a definida pelo vocabulário controlado.
M20	Logging, Ficha Catalográfica, Morae	Número de termos oriundos do vocabulário controlado que são usados para descrever um dado recurso.

Tabela 15: Métricas que operacionalizam as questões.

ANEXO H – QUESTIONÁRIO

Avaliação da Catalogação no repositório Una-SUS - UFSC

Em cada questão, assinale a alternativa que melhor define sua opinião:

1. Você gostou de usar as interfaces baseadas em conhecimento?
   [ ] Discordo fortemente
   [ ] Discordo
   [ ] Concordo
   [ ] Concordo fortemente

2. Você julga ser fácil usar as interfaces baseadas em conhecimento?
   [ ] Discordo fortemente
   [ ] Discordo
   [ ] Concordo
   [ ] Concordo fortemente

3. Qual interface baseada em conhecimento você prefere?
   [ ] Autocompletar
   [ ] Hierárquica
   [ ] Hiperbólica
   [ ] Nenhuma

4. As interfaces baseadas em conhecimento são claras e fáceis de entender?
   [ ] Discordo fortemente
   [ ] Discordo
   [ ] Concordo
   [ ] Concordo fortemente

5. Você gostaria que as interfaces baseadas em conhecimento fossem utilizadas para o preenchimento de outros campos?
   [ ] Discordo fortemente
   [ ] Discordo
   [ ] Concordo
   [ ] Concordo fortemente

   Em caso de concordância na questão anterior, especifique os campos:_____

6. Você julga que as interfaces baseadas em conhecimento podem ser aplicadas a outros domínios?
   [ ] Discordo fortemente
   [ ] Discordo
   [ ] Concordo
   [ ] Concordo fortemente

7. As interfaces baseadas em conhecimento facilitaram a execução das tarefas?
   [ ] Discordo fortemente
   [ ] Discordo
   [ ] Concordo
   [ ] Concordo fortemente

8. Os termos providos pelas interfaces baseadas em conhecimento foram suficientes para a descrição dos recursos?
   [ ] Discordo fortemente
   [ ] Discordo
   [ ] Concordo
   [ ] Concordo fortemente

9. Você conseguiu realizar as tarefas propostas valendo-se das interfaces baseadas em conhecimento?
   [ ] Discordo fortemente
   [ ] Discordo
   [ ] Concordo
   [ ] Concordo fortemente

10. Qual interface você julga mais trabalhosa de se usar?
    [ ] Autocompletar
    [ ] Hierárquica
    [ ] Hiperbólica

11. Você tem alguma crítica, sugestão ou observação a respeito das interfaces baseadas em conhecimento e/ou repositório UnA-SUS - UFSC?

    _____

    _____

    _____

ANEXO I – AJUDA PARA PREENCHIMENTO DA FICHA CATALOGRÁFICA

Formulário para Catalogação dos Objetos de Aprendizagem (OA)

Os descritores marcados com "*" são obrigatórios.

Ajuda ao Preenchimento

Descritor	Descrição
Título*	Título do OA (podendo ser mais de um, ex. título traduzido)
Autores*	Autores do OA (como presente no Lattes, ou se não for o caso, nome completo)
Editores	Editores do OA (como presente no Lattes, ou se não for o caso, nome completo)
Ilustradores	Designers Gráficos que participaram da implementação do OA . (como presente no Lattes, ou se não for o caso, nome completo)
Colaboradores	Demais pessoas que colaboraram no projeto do OA (como presente no Lattes, ou se não for o caso, nome completo).
Data*	Data da criação ou produção do OA (podendo ser apenas ano, ano e mês, ou ano, mês e dia.
Descrição/resumo*	Descrição do OA
Palavras-chaves livres	Palavras chaves livres, que complementariam os descritores DECs.
Palavras-chaves do vocabulário controlado*	Lista de descritores DECs que descrevem o OA. Estes descritores são disponíveis em http://decs.bvs.br/. Esta lista será estendida com descritores apropriados para a farmácia.
Tipo de recurso*	Seleção de tipos de recursos educacionais que descrevem o OA. O vocabulário IEEE LOM inclui: Exercício, Simulação, Questionário, Diagrama, Figuras, Grafos, Índice, Slide, Tabela, Texto narrativo, prova, experimento, autoavaliação, palestra.
Versão	Versão do OA.
Tipo de Interatividade:*	• ativa: aprendizagem ativa (por exemplo, aprendendo fazendo) é suportado pelo conteúdo que induz a ação produtiva diretamente pelo aluno. • expositiva: aprendizagem expositiva (por exemplo, aprendizagem passiva) ocorre quando o trabalho do aluno consiste principalmente de absorver o conteúdo exposto aos mesmos. • mista: Uma mistura de tipos de interativade ativa e expositive..
Objetivos educacionais*	Objetivos educacionais do OA.
Público alvo*	A entidade para quem o recurso é direcionado ou útil: Dentista, Enfermeiro, farmacêutico, Médico. Outros termos podem ser acrescidos.
Nível educacional*	Uma classe de entidade, definida em termos de progressão através de um contexto educacional ou de treinamento, para que o recurso descrito é direcionado. As opções são Técnico, Superior, Especialista
idioma	Idioma principal do objeto.

Os descritores marcados com "*" são obrigatórios.

ANEXO J – FICHA CATALOGRÁFICA EM BRANCO

Formulário para Catalogação dos Objetos de Aprendizagem (OA)

Descritor	Descrição
Título*:	
Autores*:	
Editores:	
Ilustradores:	
Colaboradores:	
Data*:	/ /
Descrição/ Resumo*:	
Palavras-chaves do vocabulário controlado*:	Lista de descritores DECs que descrevem o OA. Estes descritores estão disponíveis em http://decs.bvs.br/. Favor, indicar pelo menos três palavras-chave.
Palavras-chaves livres:	Palavras chaves livres separadas por ";"
Tipo de recurso*:	☐ Exercícios ☐ Simulações ☐ Questionários ☐ Diagramas ☐ Figuras ☐ Grafos ☐ Índices ☐ Slides ☐ Tabelas ☐ Textos Narrativos ☐ Provas ☐ Experimentos ☐ Autoavaliações ☐ Palestras
Versão:	
Tipo de Interatividade*:	○ Ativa ○ Expositiva ○ Mista
Objetivos educacionais*:	
Público alvo*:	☐ Dentistas ☐ Enfermeiros ☐ Farmacêuticos ☐ Médicos Outros:
Nível educacional*	☐ Médio ☐ Técnico ☐ Especialista ☐ Superior
Idioma	○ Português ○ Espanhol ○ Inglês ○ Francês ○ Alemão

ANEXO K – FICHA CATALOGRÁFICA PREENCHIDA

Formulário para Catalogação dos Objetos de Aprendizagem (OA)

Descritor	Descrição
Título*:	Infecções Respiratórias Agudas Mais Comuns nas Crianças
Autores*:	Marcela Döhms; Leandro Pereira Garcia; Dr. Luiz Roberto Agea Cutolo
Editores:	
Ilustradores:	
Colaboradores:	
Data*:	/ 05 / 2011
Descrição/ Resumo*:	Discorre sobre os principais problemas infecciosos respiratórios agudos mais comuns nas crianças.
Palavras-chaves do vocabulário controlado*:	Resfriado Comun; Pneumonia; Sinuzite; Influenza Humana; Bronquite; Criança Otite Media; Infecções Respiratória; Laringitis Lista de descritores DECs que descrevem o OA. Estes descritores estão disponíveis em http://decs.bvs.br/. Favor, indicar pelo menos três palavras-chave.
Palavras-chaves livres:	Palavras chaves livres separadas por ";"
Tipo de recurso*:	☑ Exercícios ☐ Simulações ☐ Questionários ☐ Diagramas ☑ Figuras ☐ Grafos ☐ Índices ☐ Slides ☐ Tabelas ☑ Textos Narrativos ☐ Provas ☐ Experimentos ☐ Autoavaliações ☐ Palestras
Versão:	
Tipo de Interatividade*:	○ Ativa ○ Expositiva ◉ Mista
Objetivos educacionais*:	Apoiar a formação continuada de profissionais do SUS
Público alvo*:	☐ Dentistas ☐ Enfermeiros ☐ Farmacêuticos ☑ Médicos Outros:
Nível educacional*	☐ Médio ☐ Técnico ☐ Especialista ☑ Superior
Idioma	○ Português ○ Espanhol ○ Inglês ○ Francês ○ Alemão

ANEXO L – BUSCA SEMÂNTICA VIA WEB SERVICE

A Figura 60 ilustra como o módulo de busca semântica poderia ser explorado numa arquitetura baseada em *Web-Service*. Cada aplicação poderia acessar o *Web service* via interfaces disponibilizadas e fornecer a (1) consulta, suas (2) anotações semânticas em formato RDF e baseadas na ontologia definida pelo DeCS e (3) eventualmente os parâmetros de configuração específicos para a técnica de processamento SA.

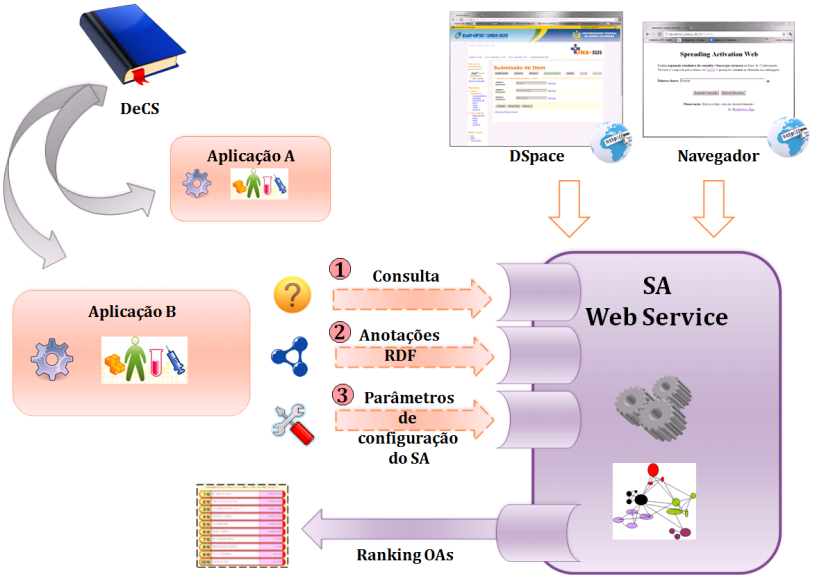


Figura 60: Módulo de busca semântica via Web service.

ANEXO M – BUSCA SEMÂNTICA VIA *FRONT-END* WEB

A Figura 61 apresenta a inteface Web que dispara buscas semânticas baseadas nos termos oriundos do DeCS, os quais são selecionados via interfaces gráficas baseadas em conhecimento. Tais termos iniciam a expansão semântica sobre uma RS que abriga OIs anotados com termos do VC DeCS e que estão catalogados em um SRI hipotético.

Figura 61: *Front-end* acessível via Web que utiliza o módulo de busca semântica desenvolvido.

ANEXO N – INTEGRAÇÃO DO MÓDULO DE BUSCA SEMÂNTICA AO DSPACE.

Os módulos do CIBELE foram desenvolvidos com o objetivo de prover um mecanismo de busca semântica a SRIs e manter um fraco acoplamento com tais sistemas. Este acoplamento é realizado via funções JavaScript que realizam *HTTP Request* invocando os módulos desenvolvidos, como mostra a Figura 62. Ela também ilustra os 8 passos compreendidos entre a consulta do usuário e a obtenção da resposta:

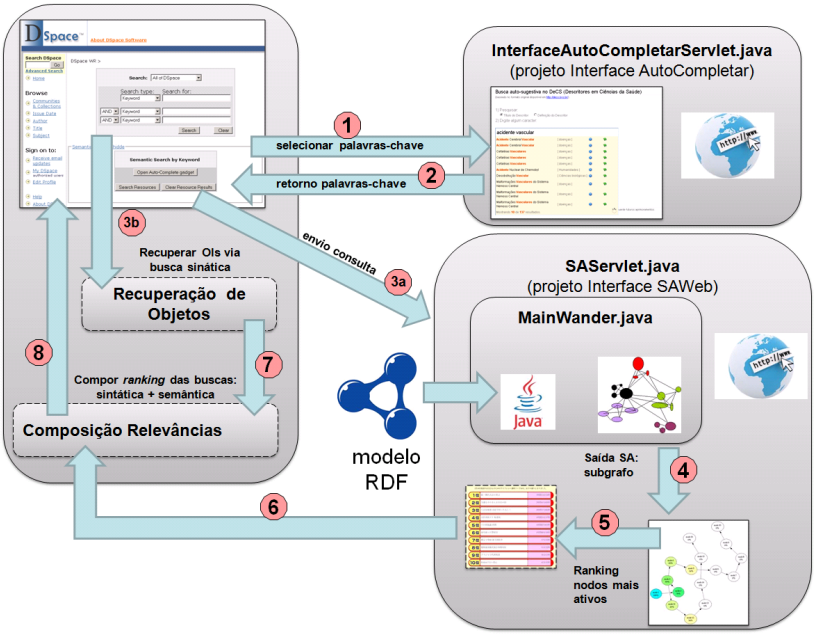


Figura 62: Integração módulo de busca semântica ao DSpace.

1. via interface Autocompletar, o usuário seleciona palavras-chave visando compor uma consulta $c_i \in C$;

2. o conjunto de palavras-chave $P(c_i)$ que compõe a consulta é enviado para o DSpace para possível verificação e confirmação do usuário;

3. o usuário então dispara a busca híbrida que atua em paralelo e:

   3a. no DSpace opera busca sintática baseada em campos de metadados que possam ter sido especificados (e.g., autor, data de publicação);

   3b. no módulo de recuperação semântica, via SA, opera busca semântica a partir do conjunto $P(c_i)$;

4. a execução do SA resulta num subgrafo RS' contendo os nodos ativados e as respectivas relevâncias semânticas;

5. a partir de RS', uma listagem de OIs ordenados por relevância semântica é montada;

6. tal listagem é enviada ao DSpace;

7. após o processamento da busca sintática, o DSpace retorna uma listagem de resultados ordenados por relevância sintática, segundo os campos de metadados especificados na busca;

8. as listagens obtidas nos passos 5 e 7 são combinadas segundo a Definição 10 e então uma listagem resultante é apresentadas ao usuário.

ANEXO O – DETALHES SOBRE O DECS

Estrutura

Os conceitos que compõem o DeCS são organizados em uma estrutura hierárquica permitindo a execução de pesquisa em termos mais amplos ou mais específicos ou todos os termos que pertençam a uma mesma estrutura hierárquica. O DeCS é um vocabulário dinâmico totalizando 30.369 descritores, sendo destes 25.671 do MeSH e 4698 exclusivamente do DeCS. Ele é atualizado anualmente e em 2010 apresentava 20 categorias:

(A) Anatomia

(B) Organismos

(C) Doenças

(D) Compostos Químicos e Drogas

(E) Técnicas Analíticas, Diagnósticas e Terapêuticas e Equipamentos

(F) Psiquiatria e Psicologia

(G) Fenômenos e Processos

(H) Disciplinas e Ocupações

(HP) Homeopatia

(I) Antropologia, Educação, Sociologia e Fenômenos Sociais

(J) Tecnologia, Indústria, Agricultura

(K) Ciências Humanas

(L) Ciência da Informação

(M] Denominações de Grupos

(N) Assistência à Saúde

(SH) Ciência e Saúde

(SP) Saúde Pública

(V) Características de Publicações

(VS) Vigilância Sanitária

(Z) Denominações Geográficas

As 4 categorias em negrito foram especialmente desenvolvidas no Brasil visando melhor representar a literatura brasileira gerada. Os conceitos do vocabulário DeCS, na versão 2010, estão assim distribuídos:

- 25,8% referem-se a **Compostos Químicos e Drogas** (categoria D), entendendo aqui tanto as drogas exógenas como as endógenas;
- 20,4% do total são da área de **Anatomia** (categoria A), de **Organismos** (categoria B) e de **Fenômenos e Processos** (categoria G);
- 12,9% do total são referentes a **Doenças** (categoria C);
- 21,6% são representados pelas áreas de **Técnicas e Equipamentos** (categoria E), ciências afins (categorias F, H, I, J, K, L, M, N), **Características de publicações** (categoria V) e **Denominações Geográficas** (categoria Z)
- 10,2% do total de conceitos referem-se a área de **Saúde Pública** (categoria SP)
- 5,7% do total de conceitos referem-se a **Homeopatia** (categoria HP)
- 2,4% do total de conceitos referem-se a **Vigilância Sanitária** (categoria VS)
- 0,6% do total de conceitos referem-se a **Ciência e Saúde** (categoria SH)

Em (BOCCATO; FUJITAI, 2006), os autores fazem uma síntese de diversos trabalhos mostrando que o DeCS foi objeto de estudos de alguns pesquisadores latino-americanos na verificação de sua performance e de sua estrutura na descrição e recuperação de informação em sistemas de informação. A partir desses estudos, recomendações foram propostas à BIREME, enfatizando que o vocabulário possui carências e tende a continuar evoluindo.

Aquisição

Graças ao trabalho do pesquisador Dr. Divino Ignácio Ribeiro Júnior, o DeCS foi obtido a partir do Serviço DeCS/XML disponibilizado em http://decs.bvsalud.org/vmx.htm. O processo de captura e conversão foi realizada em duas etapas:

a) desenvolvimento de um *crawler* para captura dos descritores, sinônimos, definições, taxonomia e relacionamentos horizontais nos seus 3 idiomas, com o qual cada descritor é salvo em um arquivo XML. A validação da captura é realizada por amostragem das categorias, examinando-se o conteúdo obtido e a informação correspondente fornecida no site do Serviço DeCS/XML. O *crawler* captura os descritores guiando-se pelos códigos hierárquicos de tais descriores.

b) conversão dos arquivos XML salvos, por meio de um *script* que popula um banco de dados MySQL previamente preparado para este fim.

Tratamento

A partir do DeCS armazenado em formato relacional é que iniciou-se o desenvolvimento dos artefatos expostos neste trabalho. O DeCS forneceu os insumos básicos que alimentam as interfaces de visualização de conhecimento. Através de algoritmos e *parsers* desenvolvidos programaticamente por nós, o DeCS foi convertido para os formatos exigidos (XML) pelas ferramentas de visualização, assim como para a geração da espinha dorsal da Rede Semântica em RDF. Tais ferramentas podem ser vistas em detalhes nos Anexos A.1, A.2 e A.3.

ANEXO P – PROTÓTIPO PARA EXECUÇÃO DO SA

A interface do protótipo *desktop* desenvolvido em Java é mostrada na Figura Q.

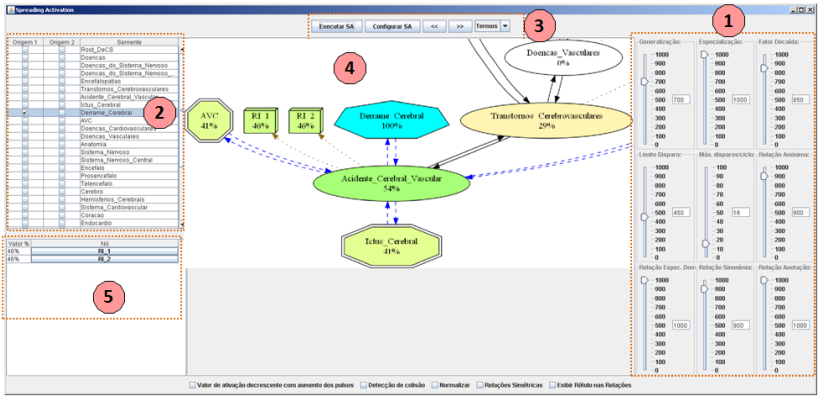
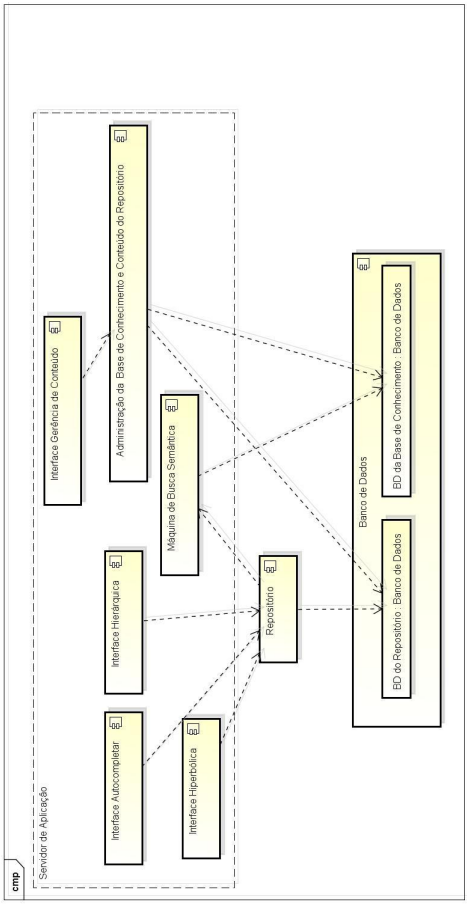


Figura 63: Interface do protótipo para execução do SA.

Os sliders situados do lado direito da interface (1) configuram os valores dos parâmetros do SA. Em (2), o usuário seleciona o conjunto de termos sementes do SA. A execução é iniciada pelos comandos definidos em (3). Em (4) são exibidas imagens que retratam cada passo de execução (pulso) do SA. Ao final da execução, uma listagem de OIs ordenada por relevância semântica é apresentada em (5). Tal protótipo usa o mesmo motor de busca utilizado pelo módulo de recuperação semântica.

ANEXO Q – DIAGRAMA UML DOS COMPONENTES DO CIBELE



REFERÊNCIAS

AFONSO, M. da C. L. *Banco Internacional de Objetos Educacionais (BIOE): normas para a definição dos metadados*. Brasília: [s.n.], 2010.

ANDERSON, J. R. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, v. 22, p. 261–295, 1983.

ASWATH, D. et al. Boosting item keyword search with spreading activation. In: *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*. [S.l.: s.n.], 2005. p. 704 – 707.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval*. New York: [s.n.], 1999. 511 p.

BASILI, V. R.; ROMBACH, H. D. The tame project: Towards improvement-oriented software environments. *IEEE Trans. Software Eng.*, p. 758–773, 1988.

BERNERS-LEE, J. H. T.; LASSILA, O. *The Semantic Web*. May 2001. 29-37 p.

BERTHOLD, M. R. et al. Pure spreading activation is pointless. In: *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2009. p. 1915–1918. ISBN 978-1-60558-512-3.

BIREME. *Descritores em Ciências da Saúde (DeCS)*. 2010. Acessado em 3 de Maio de 2010. Disponível em: <http://decs.bvs.br>.

BOCCATO, V. R. C.; FUJITAI, M. S. L. Estudos de avaliação quantitativa e qualitativa de linguagens documentárias: uma síntese bibliográfica. *Perspectivas em Ciência da Informação*, scielo, v. 11, p. 267 – 281, 08 2006. ISSN 1413-9936.

BORGO, S. *Classifying (Medical) Ontologies.* Trento-Roma, Italy, 2004.

BURKHARD, R. A.; MEIER, M. Tube map: Evaluation of a visual metaphor for interfunctional communication of complex projects. In: *Proceedings of I-Know '04 - 4th International Conference on Knowledge Management*. Graz, Austria: [s.n.], 2004.

CARDOSO, O. N. P. *Recuperação de Informação*. Lavras: Infocomp - Revista de Computação da Universidade Federal de Lavras - UFLA, 2000. 33-38 p. [Online; acessado em 27 Julho de 2010]. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos% -/v2.1/art07.pdf>.

CHEN, M.; QIN, J. Deriving ontology from folksonomy and controlled vocabulary. In: UNIVERSITY OF CALIFORNIA. Los Angeles: iConference 2008, 2008.

COLLINS, A. M.; LOFTUS, E. F. A spreading-activation theory of semantic processing. *Psychological Review*, v. 82, n. 6, p. 407–428, 1975.

COOPER, G. F. The computational complexity of probabilistic inference using bayesian belief networks (research note). *Artif. Intell.*, Elsevier Science Publishers Ltd., Essex, UK, v. 42, p. 393–405, March 1990. ISSN 0004-3702. Disponível em: <http://portal.acm.org/citation.cfm?id=77754.77762>.

CRESTANI, F. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, v. 11, p. 453–482, 1997.

CRESTANI, F.; LEE, P. L. Webscsa: Web search by constrained spreading activation. In: *ADL'99*. [S.l.: s.n.], 1999. p. 163–170.

CRESTANI, F.; RIJSBERGEN, C. J. V. Modelling adaptive information retrieval. *In: Journal of Intelligent Information Systems*, v. 8, p. 29–56, 1993.

D'AGOSTINI, C. S.; FILETO, R. Capturing users' preferences and intentions in a semantic search system. In: *SEKE*. [S.l.]: Knowledge Systems Institute Graduate School, 2009. p. 587–591. ISBN 1-891706-24-1.

DAHL, D.; VOSSEN, G. Evolution of learning folksonomies: social tagging in e-learning repositories. *International Journal of Technology Enhanced Learning*, v. 1, n. 1/2, p. 35–46, 2008. Disponível em: <http://dx.doi.org/10.1504/IJTEL.2008.020229>.

DEAN, M.; SCHREIBER, G. *OWL Web Ontology Language Reference*. 2004. W3C Recommendation. Disponível em: <http://www.w3.org/TR/2004/REC-owl-ref-20040210>.

DECAMPOS, L.; FERNáNDEZ-LUNA, J. M.; HUETE, J. F. Bayesian networks and information retrieval: an introduction to the special

issue. *Information Processing & Management*, v. 40, n. 5, p. 727–733, set. 2004. ISSN 03064573. Disponível em: <http://dx.doi.org-/10.1016/j.ipm.2004.03.001>.

DIAS, M. P.; CARVALHO, J. O. F. de. A visualização da informação e a sua contribuição para a ciência da informação. DataGramaZero - Revista de Ciência da Informação, v. 8, 10 2007. [Online; acessado em 3 de Maio de 2010]. Disponível em: <http://dgz.org.br/out07/Art_02-.htm>.

DUPRIEZ, C.; SCHUBNEL, J. Windmusic, example of the new possibilities for dspace when adding skos thesaurus and authority lists management. In: DSPACE USER GROUP MEETING 2009. Gothenburg, Suécia, 2009.

ECHARTE, F. et al. Ontology of folksonomy: A new modeling method. In: HANDSCHUH, S. et al. (Ed.). [S.l.: s.n.], 2007.

EUZENAT, J. Eight Questions about Semantic Web Annotations. *IEEE Intelligent Systems*, IEEE Computer Society, Los Alamitos, CA, USA, v. 17, n. 2, p. 55–62, 2002.

FALOUTSOS, C.; OARD, D. W. *A survey of information retrieval and filtering methods*. College Park, MD, USA, 1995.

FILETO, R. et al. Performance evaluation and tuning of spreading activation for associative information retrieval. In: *ICWI*. Rio de Janeiro, RJ, Brazil: IADIS, 2011.

FREITAS, C. M. D. S. et al. *Introdução à Visualização de Informações*. Porto Alegre: Revista de Informática Teórica e Aplicada, 2001. 143-158 p. [Online; acessado em 27 Julho de 2010]. Disponível em: <http:/-/graphs.ucpel.tche.br/luzzardi/Rita.pdf>.

GOMES, H. E.; MOTTA, D. F. da; CAMPOS, M. L. de A. *Elaboração de Tesauro Documentário - Glossário*. 2009. Acessado em 2 de Agosto de 2010. Disponível em: <http://www.conexaorio.com/biti/tesauro-/glossario.htm>.

GONçALVES, V. M. B. *A Web Semântica no Contexto Educativo: Um sistema para a recuperação de objectos de aprendizagem baseado nas*

tecnologias para a Web Semântica, para o e-Learning e para os agentes. Tese (Doutorado) — Faculdade de Engenharia da Universidade do Porto - FEUP, 2007.

GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, v. 5, n. 2, p. 199–221, 1993. Disponível em: <http://tomgruber.org/writing/ontolingua-kaj-1993.pdf>.

HILLMANN, D. *Using Dublin Core*. 2005. Acessado em 3 de Maio de 2010. Disponível em: <http://dublincore.org/documents/usageguide>.

HODGINS, W. *Learning Object Metadata (LOM)*. 2002. Acessado em 3 de Maio de 2010. Disponível em: <http://ltsc.ieee.org/wg12>.

HUSSEIN, T.; NEUHAUS, S. *Explanation of spreading activation based recommendations*. Duisburg, 2010. Disponível em: <http://semais.org-/papers/Hussein2.pdf>.

JANSEN, B. J. The effect of query complexity on web searching results. *Information Research*, v. 6, n. 1, p. –1–1, 2000.

JENSEN, F. V. *An introduction to Bayesian networks*. Springer, 1996. ISBN 9780387915029. Disponível em: <http://books.google.com-/books?id=g8hlQgAACAAJ>.

JUDELMAN, G. B. *Knowledge visualization: Problems and Principles for Mapping the Knowledge Space*. Dissertação (Mestrado) — International School of New Media, Germany, 06 2004. [Online; acessado em 11 de Dezembro de 2010]. Disponível em: <http://www.gregjudelman.com/media/judelmanThesis2004.pdf>.

KATIFORI, A. et al. Ontology visualization methods-a survey. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 39, November 2007. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/1287620-.1287621>.

KIRYAKOV, A. et al. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics*, v. 2, p. 49–79, 2004.

LAMPING, J.; RAO, R.; PIROLLI, P. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co.,

1995. (CHI '95), p. 401–408. ISBN 0-201-84705-1. Disponível em: <http://dx.doi.org/10.1145/223904.223956>.

LEVY, D. M. Cataloging in the digital order. In: *DL*. [s.n.], 1995. p. 0–. Disponível em: <http://www.csdl.tamu.edu/DL95/papers/levy/levy-.html>.

LIKERT, R. A technique for the measurement of attitudes. *Archives of Psychology*, v. 22, n. 140, p. 1–55, 1932.

MANGOLD, C. A survey and classification of semantic search approaches. *Int. J. Metadata Semant. Ontologies*, Inderscience Publishers, Inderscience Publishers, Geneva, SWITZERLAND, v. 2, n. 1, p. 23–34, 2007. ISSN 1744-2621.

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology. *Decis. Support Syst.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 15, p. 251–266, December 1995. ISSN 0167-9236. Disponível em: <http://dx-.doi.org/10.1016/0167-9236(94)00041-2>.

MATHES, A. Folksonomies - cooperative classification and communication through shared metadata. In: . [s.n.], 2004. Disponível em: <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>.

MAXIMINO, A.; MARTINS, P. *Web Cataloging*. Covilhã, Portugal, 2004. Disponível em: <http://www.di.ubi.pt/\˜api/web_cataloging.pdf>.

MENG, F.; CHU, W. W. *Database Query Formation from Natural Language using Semantic Modeling and Statistical Keyword Meaning Disambiguation*. [S.l.], 1999.

NILAS, N.; NILAS, P.; MASAKUL, K. A spreading activation approach for e-commerce site selection system. Proceedings of the International Conference on e-Business 2007, 2007.

NOVAK, J. D. *The Theory Underlying Concept Maps and How to Construct Them*. Pensacola, 2006. Disponível em: <http://cmap.ihmc.us/Publications/ResearchPapers/TheoryCmaps/TheoryUnderlyingConceptMaps.htm>.

O'REILLY, T. *What Is Web 2.0*. 2005. [Online; acessado em 9 de Agosto de 2010]. Disponível em: <http://oreilly.com/pub/a/web2/archive/what-is-web-20.html>.

OREN, E. et al. *What are Semantic Annotations*. 2006.

PEFFERS, K. et al. A design science research methodology for information systems research. *J. Manage. Inf. Syst.*, M. E. Sharpe, Inc., Armonk, NY, USA, v. 24, p. 45–77, December 2007. ISSN 0742-1222. Disponível em: <http://portal.acm.org/citation.cfm?id=1481765.1481768>.

PELLIZZON, R. de F. Pesquisa na área da saúde: 1 - base de dados decs (descritores em ciências da saúde). *Acta Cirurgica Brasileira*, scielo, v. 19, p. 153 – 163, 04 2004. ISSN 0102-8650.

PEREIRA, V. C. J. *Recuperação Associativa de Recursos de Informação Utilizando Spreading Activation*. Florianópolis: [s.n.], 2011.

PETERS, I.; BECKER, P. *Folksonomies : indexing and retrieval in Web 2.0*. Berlin: De Gruyter/Saur, 2009.

QUILLIAN, M. R. Semantic memory. In: MINSKY, M. (Ed.). *Semantic Information Processing*. Cambridge, MA, USA: MIT Press, 1968. p. 216–270.

RIGO, W. et al. Interfaces web baseadas em conhecimento para anotação de recursos de informação e gerenciamento de repositórios. In: SOCIEDADE BRASILEIRA DE COMPUTAçãO. *XXI Simpósio Brasileiro de Informática na Educação (SBIE 2010)*. João Pessoa, PB, 2010. Disponível em: <http://www.proativa.virtual.ufc.br/sbie2010-/Anais_do_XXI_SBIE//Artigos_Completos_files/76456_1.pdf>.

RIGO, W. et al. Anotação de conteúdo multimídia em repositórios com interfaces web baseadas em conhecimento de domínio. In: *WebMedia'11*. Florianópolis, SC: [s.n.], 2011.

ROCHA, C.; SCHWABE, D.; ARAGãO, M. P. de. A hybrid approach for searching in the semantic web. In: *WWW'04*. [S.l.: s.n.], 2004. p. 374–383.

ROCHA, T. R. da et al. *Anotações Semânticas em Bibliotecas Digitais Voltadas ao Ensino.* Fortaleza, CE: [s.n.], 2008.

RODRIGUES JR., J. F. et al. The visual expression process: Bridging vision and data visualization. In: *Proceedings of the 9th international symposium on Smart Graphics*. Berlin, Heidelberg: Springer-Verlag, 2008. (SG '08), p. 207–215. ISBN 978-3-540-85410-4. Disponível em: <http://dx.doi.org/10.1007/978-3-540-85412-8_19>.

SAVOY, J. Bayesian inference networks and spreading activation in hypertext systems. *Inf. Process. Manage.*, p. 389–406, 1992.

SCHIEL, U. Abstractions in semantic networks: axiom schemata for generalization, aggregation and grouping. *SIGART Bull.*, ACM, New York, NY, USA, n. 107, p. 25–26, 1989. ISSN 0163-5719.

SHAH, U. et al. Information retrieval on the semantic web. In: . ACM Press, 2002. p. 461–468. Disponível em: <http://semais.org/papers-/Hussein2.pdf>.

4TH INTERNATIONAL CONFERENCE ON OPEN REPOSITORIES. *Depth Customization of DSpace: Best Practices and Techniques of Institutional Repository at IIT Kanpur, India*. Atlanta, GA, USA: Georgia Institute of Technology, 2009.

SILVA, M. F. *Estudo comparativo entre interfaces hipertextuais de softwares para a representação do conhecimento*. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, Belo Horizonte, 2007. [Online; acessado em 3 de Maio de 2010].

SILVA, R. S. da. *Moodle para autores e tutores.* [S.l.]: Novatec Editora, 2010.

SOUZA, A. B. de et al. *Recuperação Semântica de Objetos de Aprendizagem: Uma Abordagem Baseada em Tesauros de Propósito Genérico*. Simpósio Brasileiro de Informática na Educação (SBIE), 2008. [Online; acessado em 9 de Agosto de 2010]. Disponível em: <http://br-ie.org/pub/index.php/sbie/article/viewFile/749/735>.

SOUZA, A. S. de. *Avaliação de Técnicas de Visualização de Informações na Web: estudo de caso - mapa do site do UniRitter*. Porto Alegre: Centro Universitário Ritter dos Reis, 2007.

SOUZA, R. R. *Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências*. Belo Horizonte: Perspectivas em Ciência da Informação, 2006.

SVENONIUS, E. *Unanswered Questions in the Design of Controlled Vocabularies*. 1986. 331-340 p. Disponível em: <http://polaris.gseis.ucla.edu/gleazer/260_readings/Svenonius.pdf>.

SVENONIUS, E. Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, v. 37, n. 5, p. 331–340, set. 1986. ISSN 00028231.

TANSLEY, R. et al. The dspace institutional digital repository system: current functionality. In: *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*. Washington, DC, USA: IEEE Computer Society, 2003. (JCDL '03), p. 87–97. ISBN 0-7695-1939-3. Disponível em: <http://portal.acm.org/citation.cfm?id=827140.827151>.

TSATSARONIS, G.; VAZIRGIANNIS, M.; ANDROUTSOPOULOS, I. I.: Word sense disambiguation with spreading activation networks generated from thesauri. In: *20th Int. Joint Conf. in Artificial Intelligence*. [S.l.: s.n.], 2007. p. 1725–1730.

TUCK, M. *The Real History of the GUI*. SitePoint, 2001. Disponível em: <http://articles.sitepoint.com/article/real-history-gui>.

WAL, T. V. *Folksonomy Coinage and Definition*. 2007. Acessado em 3 de Maio de 2010. Disponível em: <http://vanderwal.net/folksonomy.html>.

WAL, T. V. *Folksonomy Coinage and Definition*. February 2007.

WARNER, A. J. *A Taxonomy Primer*. 2002. Acessado em 3 de Maio de 2010. Disponível em: <http://www.ischool.utexas.edu/~i385e/readings/Warner-aTaxonomyPrimer.html>.

YAMAOKA, E. J. *Padrão de Metadados do Governo Eletrônico e-PMG*. 2009. Acessado em 2 de Agosto de 2010. Disponível em: <http://www.governoeletronico.gov.br/anexos/padrao-de-metadados-do-governo-eletronico-e-pmg>.

YU, L. *Introduction to the Semantic Web and Semantic Web Services*. Boca Raton, Flórida: Chapman & Hall/CRC, 2007.