Anotação de Conteúdo Multimídia em Repositórios com Interfaces Web baseadas em Conhecimento de Domínio

Wanderson Rigo Renato Fileto Roberto Willrich
Vilmar César Pereira Júnior Christiane Gresse von Wangenheim
Programa de Pós-Graduação em Ciência da Computação (PPGCC), Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC), Caixa Postal 476, CEP 88040-900, Florianópolis – SC, BRASIL
{wander | fileto | willrich | vilmar.pereira | gresse}@inf.ufsc.br

Vinícius de Araújo Oliveira Lina Sandra Barreto Brasil
Organização Pan-Americana da Saúde no Brasil (OPAS), Organização Mundial da Saúde (OMS)
Ministério da Saúde (MS), Esplanada dos Ministérios, Bloco G, CEP 70058-900 Brasilia – DF, BRASIL
{vinicius | linabarreto}@unasus.net

ABSTRACT
This paper presents a system to support semantic annotation and semantic browsing of contents from repositories on the Web, by using domain specific knowledge and visualization techniques. Empirical tests of this system, involving the annotation of multimedia learning objects from the health area, showed gains in annotation time, by using the proposed semantic support. These tests also indicated usability and viability of the proposal in the considered domain, and the users' predilection for an interface based on lexical and semantic auto-complete, instead of navigation in trees, to choose metadata values from a knowledge base for annotation purposes.

RESUMO
Este artigo apresenta um sistema para apoiar a anotação e a navegação semântica sobre o conteúdo de repositórios na Web, utilizando conhecimento específico de domínio e técnicas de visualização. Testes empíricos deste sistema, envolvendo a anotação de objetos de aprendizagem multimídia voltados para a área de saúde, mostraram 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 e semântico, ao invés de navegação em árvores, para escolher valores de metadados em uma base de conhecimento para fins de anotação.

Categories and Subject Descriptors
H.3.m [**Information Indexing and Retrieval**]: content tagging using domain knowledge.
H.5.2 [**User Interfaces**]: Graphical user interfaces (GUI).

General Terms
Design, Human Factors.

WebMedia'11: Proceedings of the 17th Brazilian Symposium on Multimedia and the Web. Full Papers.
October 3 -6, 2011, Florianópolis, SC, Brazil.
ISSN 2175-9642.
SBC - Brazilian Computer Society

Keywords
Graphical User Interfaces (GUI), Knowledge Visualization, Content Tagging, Information Retrieval.

1. INTRODUÇÃO
Um problema típico da nossa era é a sobrecarga de informação: à medida que cresce o volume de informações disponíveis, cresce também a dificuldade dos usuários encontrarem os objetos de informação (artigos, objetos multimídia, etc.) que procuram [1]. Uma das alternativas para minimizar este problema é criar repositórios de conteúdos que permitem descrever, organizar e recuperar as informações de forma mais eficiente [10][17][19].

Neste contexto, faz-se necessário um suporte adequado à anotação (*tagging*) de objetos de informação [15], pois objetos de informação mal descritos em decorrência das limitações das linguagens naturais (e.g., uso de sinônimos, homônimos, termos genéricos) podem ser difíceis de encontrar [21]. Por exemplo, um usuário pode anotar um objeto cujo conteúdo verse sobre "Acidente Vascular Cerebral" com qualquer um de uma variedade de termos distintos, mas que podem ser relacionados por sinonímia: "AVC", "derrame cerebral", "apoplexia", "ictus cerebral", entre outros. Posteriormente, se alguém usar um termo como palavra-chave de busca, somente os objetos anotados com o termo especificado serão recuperados por um sistema de recuperação de informação que não mantenha e explore relações semânticas entre termos.

Para que a recuperação de objetos tenha êxito, é preciso que o processo de anotação seja definido de modo a evitar que objetos sejam mal descritos e conseqüentemente "perdidos" no repositório. Para tanto, vocabulários controlados (VC) [22] 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. A anotação usando um VC e levando em consideração conceitos e instâncias que podem ser referenciados por diferentes termos semanticamente relacionados facilita a recuperação de informação.

Em [16] fazemos uma resenha de técnicas e ferramentas Web que permitem explorar interativamente conhecimento de domínio e propomos um sistema usando tais recursos para suportar

eficientemente anotação, navegação e buscas em repositórios de objetos de informação. Este trabalho concretiza tal proposta, descrevendo detalhes técnicos da implementação de um protótipo baseado em tal abordagem, além de apresentar resultados de sua avaliação com usuários reais em um estudo de caso na área de saúde. O sistema desenvolvido utiliza árvores hiperbólicas e hierárquicas para a visualização das relações semânticas entre termos de um grande VC, organizado como uma ontologia. Além disso, foi desenvolvida como parte deste sistema uma interface eficiente para completar expressões parcialmente digitadas pelo usuário com alternativas de termos deste VC que apresentam correspondências léxicas ou semânticas com o que é digitado. Constatou-se que o uso de conhecimento de domínio associado a interfaces gráficas para sua visualização contribui para a facilidade de uso, o entendimento e a produtividade do sistema de anotação semântica e navegação no conteúdo de um repositório de objetos de informação acessível via Web.

O restante deste artigo é organizado com segue. A seção 2 contextualiza o tema abordado e o estudo de caso usado na avaliação da proposta. A seção 3 descreve o sistema para amparar a descrição de objetos de informação e a navegação no conteúdo de repositórios com interfaces baseadas em conhecimento. A seção 4 relata os testes de usabilidade efetuados para avaliar as técnicas de visualização e acesso a conhecimento na anotação de objetos de informação. A seção 5 analisa os resultados obtidos em tais testes. Finalmente, a seção 6 discute trabalhos relacionados e a seção 7 conclui o artigo, enumerando contribuições e indicando temas para pesquisas futuras.

2. CONTEXTUALIZAÇÃO

2.1 Anotação de Objetos de Informação

O processo de anotação de objetos de informação visa descrever as características relevantes dos objetos para suportar a sua recuperação (IR – *Information Retrieval* [1]). Isso é feito mediante a associação de metadados aos objetos, com o objetivo de municiar os sistemas de IR com informações que facilitem a recuperação de objetos que atendam às consultas dos usuários [21]. É fundamental definir e preencher adequadamente campos de metadados (e.g., título, autor, formato, descrição) para descrever os objetos. Diversos padrões definem conjuntos de metadados genéricos (e.g., Dublin Core [11]) ou para domínios específicos (e.g., *Learning Objects Metadata* – LOM [12]). Todavia, tão importante quanto definir os campos de metadados a utilizar em uma aplicação de IR é definir possíveis valores para esses campos e garantir o seu preenchimento correto, levando em consideração conhecimento específico do domínio.

2.2 A UnA-SUS

A implementação e a avaliação da nossa proposta de uso de conhecimento de domínio para a anotação de objetos de informação [16] dá-se no contexto do programa UnA-SUS[1] (Universidade Aberta do SUS). Tal programa mantém uma rede colaborativa de instituições acadêmicas, serviços de saúde e gestão do SUS (Sistema Único de Saúde) para atender as necessidades de formação e atualização permanente de profissionais da área médica através de ações focadas em *e-learning*, intercâmbio de experiências, compartilhamento de material instrucional e cooperação no desenvolvimento e implementação de novas tecnologias educacionais em saúde. Para tanto, foi elaborado um Ambiente Virtual de Ensino e Aprendizagem (AVEA), o qual abarca, entre outras, as contribuições advindas de nosso trabalho - funcionalidades baseadas em conhecimento para anotação, busca e gerência de coleções de objetos de aprendizagem (OAs) [12], i.e., objetos de informação, de diversos tipos (e.g., documentos, imagens, hipertextos, programas executáveis) usados para fins de ensino e aprendizagem via Web, na área de saúde na área de saúde.

2.3 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 repositório de conteúdo e um ambiente de gerenciamento de aprendizagem (SGA), que podem ser acessados por diversos tipos de usuários, como ilustrado na Figura 1. O repositório de conteúdo mantém os objetos de informação (OAs, no caso da UnA-SUS), devidamente descritos com metadados para permitir a sua recuperação. O *DSpace*[2] [23] foi a base para a nossa implementação do repositório. Ele foi personalizado e enriquecido com funcionalidades da Web 2 (Web Social) e interfaces gráficas baseadas em conhecimento para apoiar a anotação, a busca e a gerência de objetos de informação. O SGA usado é o *Moodle*[3] [20], que disponibiliza cursos previamente elaborados, possibilitando a aprendizagem colaborativa à distância usando os OAs que podem ser obtidos do repositório para reuso em diversos cursos.

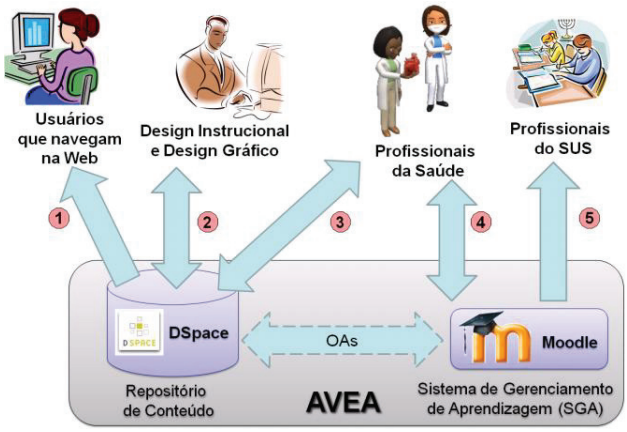

Figura 1: Tipos de usuários do AVEA.

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

3. A ABORDAGEM PROPOSTA

O nosso trabalho propõe o uso de interfaces Web baseadas em conhecimento para amparar a descrição de objetos de informação em repositórios e facilitar a recuperação de tais objetos. Tais funcionalidades se apóiam em uma arquitetura de sistema baseada em conhecimento que é mostrada na Figura 2. Um vocabulário

[1] http://www.universidadeabertadosus.org.br

[2] http://dspace.org

[3] http://moodle.org

[4] http://repositorio.unasus.ufsc.br

controlado (VC) de domínio e as anotações de objetos de informação usando tal VC constituem a base de conhecimento da nossa abordagem e são representados em um grafo. Cada nó deste grafo representa um conceito ou uma instância de conceito (que pode ser referenciado por diferentes termos, isto é, através de sinônimos). Cada aresta representa uma relação semântica entre termos ou uma anotação de um objeto de informação armazenado no repositório com o conceito ou instância ligado ao objeto através de tal aresta. Técnicas de visualização auxiliam o gerente do sistema a analisar o conteúdo do repositório, 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 e tipos de conceitos e relações semânticas a serem utilizados na anotação e recuperação dos objetos de informação. 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 conhecimento e expansão semântica no grafo contido na base de conhecimento do sistema.

O processo para a operacionalização de sistemas baseados em conhecimento e técnicas de visualização para anotação, busca e gerenciamento do conteúdo de repositórios é composta de 4 passos, os quais estão enumerados na Figura 2. As seções que seguem descrevem cada um destes passos, os quais estão sendo gradativamente implementados no repositório UnA-SUS UFSC.

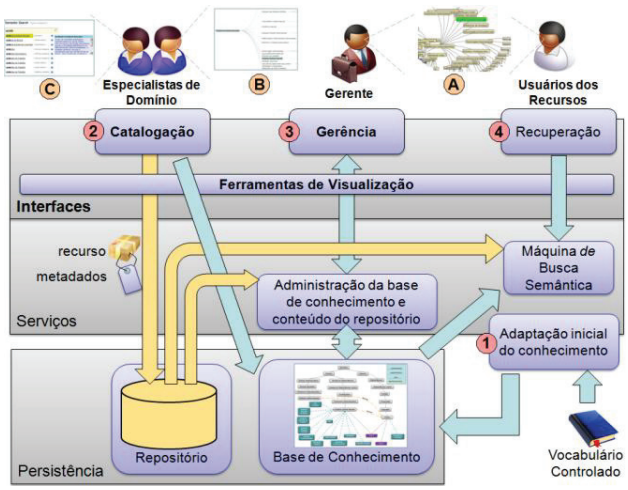

Figura 2: Arquitetura da proposta.

3.1 Adaptação inicial do conhecimento

A adaptação do conhecimento (Passo 1) para uso na anotação e recuperação da informação 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 a 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). Um trecho da hierarquia de doenças adaptada do DeCS [3] (Descritores em Ciências da Saúde, um VC muito difundido na área de saúde), é ilustrado na porção superior esquerda da Figura 3. Outro exemplo é a hierarquia de termos referentes à **anatomia**, ligados através de relações do tipo *PART_OF* (composição). Um trecho da hierarquia referente à anatomia também adaptado do DeCS é ilustrado na porção superior direita da Figura 3. 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, AVC) são representadas por linhas tracejadas na Figura 3.

3.2 Uso de conhecimento de domínio na anotação de objetos

O processo de anotação de um objeto se dá durante a catalogação do mesmo no repositório (Passo 2). É neste momento que especialistas de domínio devem definir os valores para o conjunto de metadados que descrevem o novo objeto a ser incluído no repositório. Diversos metadados, como título, autor e outros, não exigem VCs. Mas outros, como palavras-chaves, podem utilizar VCs. Nestes casos, a **anotação** de um objeto de informação 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 3 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".

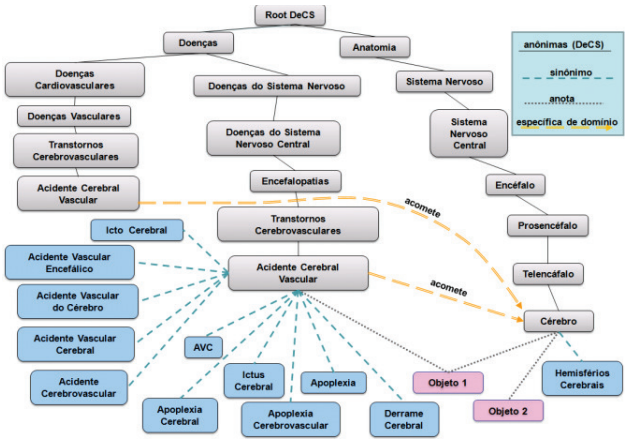

Figura 3: Estrutura de conhecimento: termos ligados por relações semânticas.

O uso de técnicas de visualização facilita a navegação em visões da base de conhecimento e a escolha dos termos relevantes dos VCs para a descrição dos objetos inseridos no repositório. Em [16] fizemos uma resenha sobre tais técnicas, enfatizando que visualizações hierárquicas e hiperbólicas são formas de se explorar VCs e assim inquirir novos conhecimentos. Técnicas de visualização (indicadas pelas letras A, B e C na Figura 2 e apresentadas em mais detalhes na Figura 4, Figura 5 e Figura 6, respectivamente) apresentam visões do VC para apoiar a seleção dos termos a serem utilizados como valores de metadados na anotação e busca de objetos de informação armazenados no repositório. A Figura 4 e a Figura 5 retratam nossas implementações, onde trechos do DeCS são exibidos com recursos de visualização hiperbólica e hierárquica providos pelas bibliotecas *Treebolic*[5] e *Prefuse*[6], respectivamente. Além dessas

[5] http://treebolic.sourceforge.net/en/index.htm

[6] http://prefuse.org

técnicas de visualização, um componente de interface capaz de completar e sugerir termos à medida que o usuário digita uma palavra-chave é útil para apoiar a anotação ágil e precisa quando o usuário já tem conhecimento sobre o domínio e o VC utilizado. A Figura 6 ilustra tal interface, que chamamos de Autocompletar. Ela utiliza eficientes mecanismos de indexação e busca do *Apache Lucene*[7] para pesquisar o VC a cada letra inserida pelos usuários e então sugere conceitos do domínio relacionados ao que está sendo digitado.

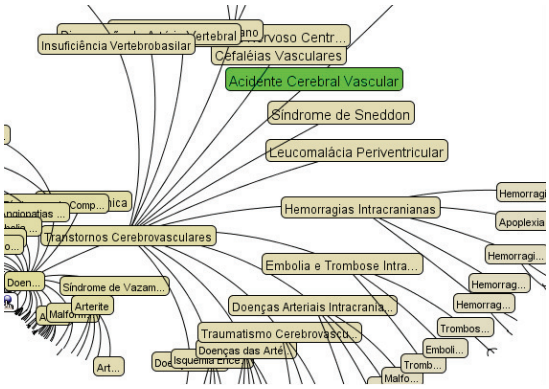

Figura 4: Visualização Hiperbólica de trecho do DeCS.

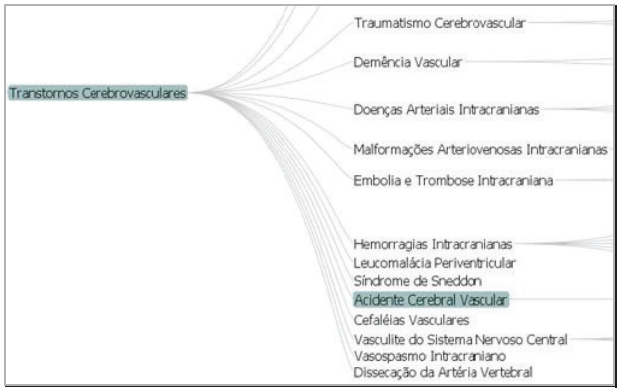

Figura 5: Visualização Hierárquica de trecho do DeCS.

Figura 6: Interface Autocompletar.

Este componente suporta buscas por sinônimos e pela descrição textual de cada conceito. Junto a cada termo são exibidas a categoria e a definição do mesmo, de modo a contribuir para a desambiguação e mostrar em que contexto tal conceito se encontra. O usuário também pode acessar informação adicional sobre um conceito através de um link que o conduz diretamente à sua definição no DeCS (em destaque na Figura 6).

3.3 Funcionalidades de Gerência

3.3.1 Gerência de Conteúdo

Em grandes repositórios alimentados dinâmica e colaborativamente é importante identificar o número de objetos disponíveis para cada assunto específico do domínio. Tal informação pode embasar gestores de repositórios a tomarem decisões sobre investimentos na produção de objetos de informação sobre determinados temas. Este trabalho propõe a visualização desse tipo de informação de maneira sintética sobre hierarquias de termos da base de conhecimento, como ilustrado na Figura 7. Os termos mais usados na anotação de objetos catalogados no repositório são destacados dos demais pelo tamanho, pelo tom da cor e pelo rótulo, que exibe o número de objetos do repositório anotados com aquele termo. Na Figura 7, o termo "Cérebro" é o mais destacado, por ser usado para anotar 3 objetos, enquanto o termo "AVC" é usado para anotar 2 objetos.

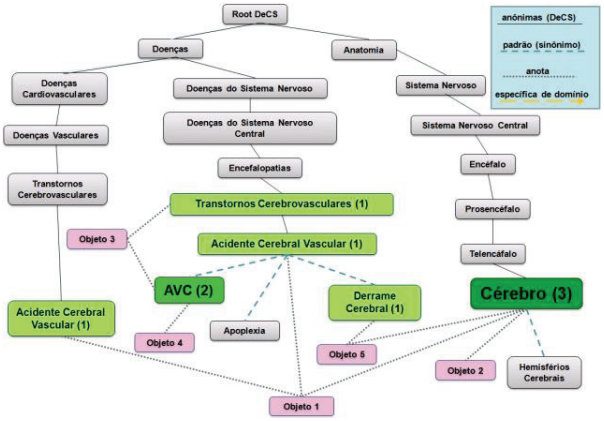

Figura 7: Objetos catalogados no repositório.

3.3.2 Gerência de Conhecimento

Segundo [4], palavras--chave que ocorrem freqüentemente juntas (co-ocorrência na descrição ou busca de objetos) não ocorrem por acaso, mas sim porque há algum tipo de relação entre elas. Pode-se explorar esta característica aproveitando-se a colaboração dos usuários que anotam objetos de informação, utilizando essas co-ocorrências para identificar possíveis relações semânticas entre termos do VC que sejam relevantes para o processo de anotação e recuperação de informação. O usuário pode eventualmente inserir novas relações semânticas dentre as relações de um conjunto de tipos pré-especificados ou elas podem ser inferidas a partir de suas interações com o sistema (e.g., ao usar dois termos para anotar um mesmo objeto). A Figura 3 representa a relação denominada "acomete" através de uma linha com traços longos que liga o termo "Acidente Cerebral Vascular" ao termo "Cérebro". Uma relação semântica como esta pode ser explicitamente inserida pelo usuário ou inferida pelo sistema, através da análise dos descritores dos termos envolvidos ou das co-ocorrências dos mesmos em anotações efetuadas pelos usuários. O gerente define os tipos de relações que podem ser inseridos na base de conhecimento. Ele também valida as relações inseridas e configura aquelas que podem ser utilizadas no processo de expansão semântica e recuperação de informação. Detalhes sobre os algoritmos e as políticas utilizadas na inserção e

[7] http://lucene.apache.org

validação de relações semânticas para o enriquecimento de VCs visando apoiar a recuperação de informação são deixados para trabalhos futuros.

3.4 Recuperação dos objetos de informação

A estrutura em grafo da base de conhecimento permite expandir semanticamente as buscas, a partir dos termos usados como palavras-chaves na especificação das mesmas, valendo-se da técnica de *Spreading Activation* [5][6]. Detalhes da estrutura de representação de conhecimento e dos algoritmos utilizados para processar as buscas transcendem o escopo e o espaço disponível neste trabalho, donde serão tratados em trabalhos futuros.

4. TESTES DE USABILIDADE

Visando aferir a aceitação e o potencial do sistema desenvolvido a partir de nossa proposta, testes de usabilidade explorando a Catalogação de OAs foram conduzidos junto a alguns usuários do repositório UnA-SUS UFSC.

4.1 Avaliadores

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

- 18 Profissionais e alunos da área de Tecnologia da Informação (TI).
- 04 Profissionais da área de Saúde.
- 03 Profissionais da área de *Design* Instrucional (DI) e *Design* Gráfico (DG).

Os profissionais 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 Design Instrucional 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.

4.2 Ambiente

Os testes foram realizados nos laboratórios de informática da UFSC (Universidade Federal de Santa Catarina), onde os avaliadores, através de navegadores Web acessaram o repositório de objetos de aprendizagem UnA-SUS UFSC disponível em http://repositorio.unasus.ufsc.br. O repositório tem como base o *DSpace* versão 1.7.1, que é uma aplicação Web desenvolvida em *Java 6*[8] e acessa um banco de dados *PostgreSQL 9.0.4*[9]. O *DSpace* foi hospedado em um *container* de servlets *Tomcat*[10], versão 6. As extensões do *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 Toda esta infra-estrutura executa sobre um servidor *FreeBSD*[11] versão 8.1. As máquinas disponibilizadas aos avaliadores contam com *Windows XP*, processador *Intel Core 2 Duo E7500* de 2.93 GHz, 2 GB de memória RAM e navegadores Web *Firefox* e *Internet Explorer*.

[8] http://www.oracle.com/technetwork/java/index.html

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

[10] http://tomcat.apache.org

[11] http://www.freebsd.org

4.3 Objetivos

Os objetivos dos testes de usabilidade das funcionalidades de anotação de objetos de informação são:

1) Avaliar a **aceitação** das soluções propostas e a **satisfação** dos usuários ao fazerem uso de tais soluções.

2) Avaliar o **potencial,** a **eficácia** e a **eficiência** das soluções propostas.

3) 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.

4) Avaliar a **eficiência** para se acessar, navegar e explorar as funcionalidades do AVEA UnA-SUS UFSC.

5) Detectar possíveis **falhas** decorrentes da falta de escalabilidade do sistema.

6) Colher **observações**, críticas e sugestões junto aos avaliadores.

4.4 GQM

Pela aplicação do método GQM (*Goal, Question, Metric*) [2], questões e métricas de avaliação foram esmiuçadas a partir do refinamento dos objetivos. Sucintamente, a seguir, esboçamos os subprodutos resultantes do método aplicado: questões e métricas.

4.4.1 Questões

As questões desdobram os conceitos abstratos relatados nos objetivos em subfatores, facilitando o entendimento de tais conceitos.

4.4.2 Métricas

Algumas métricas devem ser definidas para ajudar a responder as questões. Essas métricas 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*[12], além da instrumentação do código (*logging*).

4.5 Tarefas

Visando avaliar a utilização das funcionalidades para anotação de objetos de informação do repositório 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, assistiram a um vídeo sobre o repositório e se familiarizaram com o DeCS e 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.
- **T2**: Valendo-se da ficha catalográfica do OA2 previamente preenchida por especialistas de domínio, catalogar o OA2 no repositório.
- **T3**: Valendo-se da ficha catalográfica preenchida manualmente pelo avaliador em T1, catalogar o OA1 no repositório.

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

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 o DeCS em seu formato original [3] 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.

4.6 Questionário

Após a execução das tarefas descritas na seção anterior, foi aplicado um questionário aos usuários. As alternativas de resposta são baseadas na escala de Likert [14]. O respondente indica o seu grau de concordância ou discordância, selecionando para cada item do questionário um valor como Discordo fortemente, Discordo, Concordo e Concordo fortemente.

5. ANÁLISE DOS RESULTADOS

5.1 Análise Quantitativa

A análise quantitativa utiliza dados colhidos pela instrumentação do código do AVEA e pelo *Morae*. Ela compara o tempo gasto pelos avaliadores para realizarem as anotações propostas nas tarefas T2' e T4, as quais foram descritas na seção 4.5. A Figura 8 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 iniciaram o teste de usabilidade, 4 o abandonaram). Na maioria dos casos o ganho ultrapassou os 50% e em dois deles os 70%.

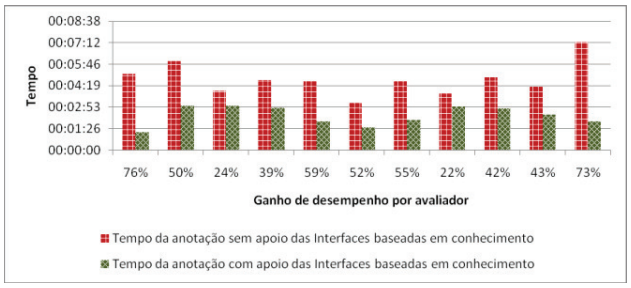

Figura 8: Comparação do tempo de anotação de T2' e T4.

A Figura 9 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.

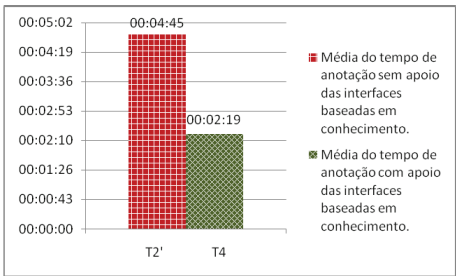

Figura 9: Média do tempo de anotação de T2' e T4.

5.2 Análise Qualitativa

A análise qualitativa avalia a percepção que os avaliadores tiveram ao usar o sistema proposto com dados coletados através do questionário descrito na seção 4.6. A Figura 10 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.

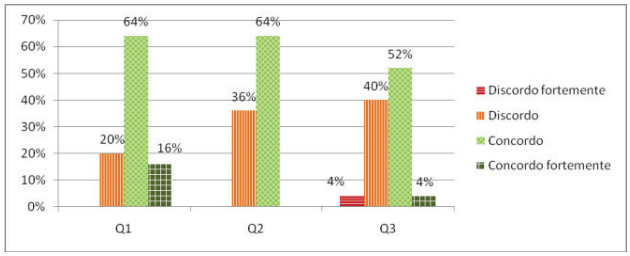

Figura 10: Satisfação do Usuário.

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 11 mostra que 60% dos avaliadores julgaram a interface Hiperbólica como a mais trabalhosa de usar, seguida pela Hierárquica com 32%.

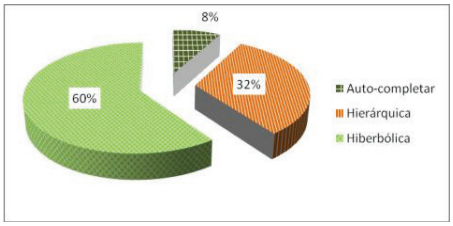

Figura 11: Interface mais trabalhosa.

A Figura 12 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%.

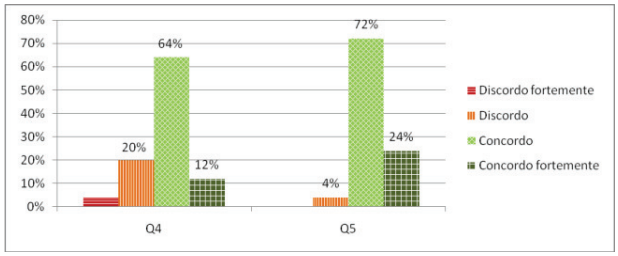

Figura 12: Potencial de uso.

A Figura 13 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]. A Figura 14 mostra que 72% dos avaliadores julgaram que os termos oriundos do DeCS e providos pelas interfaces foram suficientes para descrever os OAs. A Figura 15

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

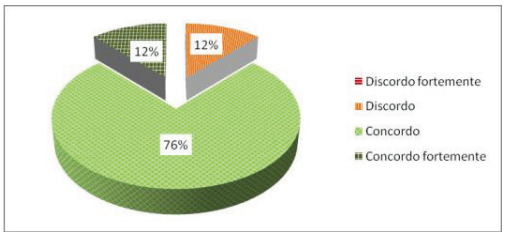

Figura 13: Facilitação na execução das tarefas.

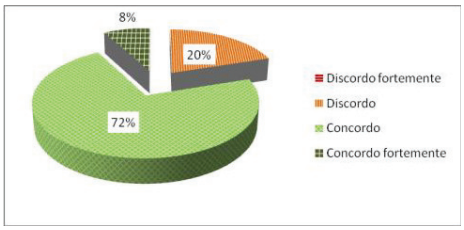

Figura 14: Completude do vocabulário.

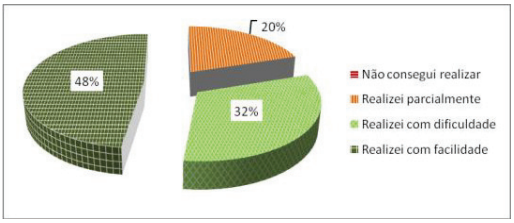

Figura 15: Realização das tarefas.

5.3 Discussão

Muitos avaliadores tiveram dificuldade em pesquisar termos no sítio do DeCS [3] 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 decorrência de tais observações coletadas nos testes de usabilidade, melhorias serão implementadas.visando facilitar a operação e o entendimento do sistema proposto.

5.3.1 Considerações quanto a validade dos Testes

Na última etapa de aplicação dos testes de usabilidade, dos 15 avaliadores da área de TI 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.

6. TRABALHOS RELACIONADOS

Em [9] são relatados esforços para dotar o *DSpace* de recursos similares aos que propomos neste trabalho, para catalogação e busca de objetos musicais, especificamente partituras e gravações. O preenchimento de certos campos de metadados foi implementado em [9] 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 [9], porém pode ser expandido para acomodar tal característica, inclusive usando versões do DeCS disponíveis em inglês e espanhol. Outros trabalhos, como [18] abordam somente a expansão de funcionalidades do *DSpace* como navegação em coleções e buscas, sem usar conhecimento específico de domínio.

7. CONCLUSÃO

Este trabalho mostra como técnicas para acesso e visualização de conhecimento de domínio podem apoiar a anotação e o gerenciamento de conteúdo multimídia em repositórios disponíveis na Web. Suas principais contribuições são: (i) definição de uma arquitetura de sistema Web baseado em conhecimento para anotação, busca e gerenciamento de grandes coleções de objetos de informação multimídia; (ii) uso de árvores hierárquicas e hiperbólicas para a visualização de conhecimento; (iii) desenvolvimento de um componente com autocompletar dinâmico para operar buscas por valores de metadados em VCs e (iv) avaliação de um sistema integrando tais recursos na anotação de objetos de informação em um estudo de caso com relevante aplicação na área da saúde. Esta avaliação foi possível graças à colaboração voluntária de avaliadores de diferentes segmentos, os quais representam um substrato considerável dos reais usuários do AVEA UnA-SUS.

O apoio das interfaces baseadas em conhecimento propiciou ganhos de até 76% no tempo de anotação de objetos, frente a uma abordagem sem tal apoio. Na média este ganho foi de 51%. O gosto pelo uso e a facilidade na operação das interfaces baseadas

em conhecimento foi manifestado por 64% dos avaliadores, sendo que 76% deles julgaram que elas facilitaram a execução das tarefas. Esta aprovação deve-se principalmente à interface de autocompletar baseada em correspondências léxicas e semânticas, a preferida por 100% dos avaliadores, os quais julgaram-na ágil e fácil de operar. O sistema desenvolvido foi aplicado na área de saúde, mas também pode operar em outros domínios para os quais haja conhecimento formalizado disponível para inserção na sua base de conhecimento.

7.1 Trabalhos Futuros

O módulo para a recuperação de objetos de informação está em desenvolvimento. Tal módulo utiliza a técnica de *Spreading Activation* [5][6] e está sendo testado sob diferentes configurações de parâmetros, buscando-se maximizar a precisão e a cobertura dos resultados retornados. Esta pesquisa também desvelou outras idéias promissoras a serem exploradas em trabalhos futuros:

- Desenvolver interfaces baseadas em conhecimento na forma de componentes reusáveis. Hoje elas são módulos Web que se integram ao *DSpace* via Javascript[13] mas podem se implementadas como *Web Services* para minimizar o acoplamento com a aplicação.
- 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 [4], 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 de buscas [7].

AGRADECIMENTOS

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

8. REFERÊNCIAS

[1] Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval. New York: ACM Press, 1999. 511p.

[2] Basili, V. R.; Rombach, H. D. The TAME Project: Towards Improvement-Oriented Software Environments. IEEE Trans. on Software Engineering, v.14, n.6, p. 758-773, 1988.

[3] Centro Latino-Americano e do Caribe de Informação em Ciência da Saúde (BIREME). http://decs.bvs.br

[4] Chen , M. and Qin, J. Deriving Ontology from Folksonomy and Controlled Vocabulary. iConference 2008, University of California, Los Angeles, 2008.

[5] Crestani, F. Application of Spreading Activation Techniques in Information Retrieval, Artificial Intelligence Review, 11(6), pp.453-482, 1997.

[6] Crestani, F. Retrieving documents by constrained spreading activation on automatically constructed hypertexts. Proc. of 5th European Congress on Intelligent Techniques and Soft Computing, Aachen, Germany, pp. 1210-1214, 1997.

[7] D'Agostini, C. S. and Fileto, R. Capturing users' preferences and intentions in a semantic search system. In Proc. 21st. Intl. Conf. on Software Engineering & Knowledge Engineering (SEKE), Boston, MA, USA, pp. 587–591, 2009.

[8] Dias, M. P. and Carvalho, J. O. F. A Visualização da Informação e a sua contribuição para a Ciência da Informação. DataGramaZero 8 (5), 2007.

[9] Dupriez, C. and Schubnel, J. WindMusic, example of the new possibilities for DSpace when adding SKOS thesaurus and authority lists management. DSpace User Group Meeting 2009, Gothenburg, Suécia, 2009, http://gupea.ub.gu.se/bitstream/2077/21341/1/gupea_2077_21341_1.pdf (acessado em 18 de Abril de 2011).

[10] Faloutsos, C. and Oard, D.W. A survey of information retrieval and filtering methods, Technical. Report CS-TR-3514, University of Maryland, College Park, MD, 1995.

[11] Hillmann, D. Using Dublin Core. Dublin Core Metadata Initiative, 2005. http://dublincore.org/documents/usageguide (acessado em 3 de Maio de 2010).

[12] IEEE Learning Technology Standards Comitte (LTSC). Learning Object Metadata (LOM), 2002. http://ltsc.ieee.org/wg12 (acessado em 3 de Maio de 2010).

[13] Lassila, O. and Swick, R. R. Resource Description Framework (RDF): Model and syntax specification, 1999. http://www.w3.org/TR/1999

[14] Likert, R. A Technique for the Measurement of Attitudes. Archives of Psychology, n.140, p. 1-55, 1932

[15] Peters, I. Folksonomies: Indexing and Retrieval in Web 2.0. Walter de Gruyter GmbH & Co., Berlin, 2009.

[16] Rigo, W.; Fileto, R.; Júnior, D. I. R.; Oliveira, V. de A.; Pereira, V. C. J.; Silveira, R. A.. Interfaces Web baseadas em Conhecimento para Anotação de Recursos de Informação e Gerenciamento de Repositórios. XXI Simpósio Brasileiro de Informática na Educação (SBIE), João Pessoa, PB, 2010.

[17] Shah, U., Finin, T., Joshi, A., Cost, R. S. and Mayfield, J. Information Retrieval on the Semantic Web. 10th Intl. Conf. on Information and Knowledge Management, 2002.

[18] Shrivastava, V. D.; Shukla, Gaurav; Vijaianand, S. K: Depth Customization of DSpace: Best Practices and Techniques of Institutional Repository at IIT Kanpur, India. In: 4th Intl. Conf. on Open Repositories, Atlanta, GA, USA, 2009.

[19] Silva, M. F and Lima, G. A. B. O. Estudo comparativo entre interfaces hipertextuais de softwares para a representação do conhecimento. Dissertação (Mestrado em Ciência da Informação) - Escola de Ciência da Informação, UFMG, Belo Horizonte, 2007.

[20] Silva, R. S. Moodle para autores e tutores. São Paulo: Novatec, 2010.

[21] Souza, A. B. et al. Recuperação Semântica de Objetos de Aprendizagem: Uma Abordagem Baseada em Tesauros de Propósito Genérico. XIX Simpósio Brasileiro de Informática na Educação (SBIE). 2008.

[22] Svenonius, E. Unanswered questions in the design of controlled vocabularies. Journal of the American Society for Information Science, 37(5), 331-340, 1986.

[23] Tansley, R.; Bass, M.; Stuve, D.; Branschofsky, M.; Chudnov, D.; McClellan, G.; Smith, M. DSpace: An Institutional Digital Repository System: Current Functionality. In Proc. JCDL, Houston, Texas, 87-97, 2003.

[13] http://www.w3schools.com/js/js_intro.asp