UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

SANDRO JOSÉ RIGO

Integração de Recursos da Web Semântica e Mineração de Uso para Personalização de *Sites*

Tese apresentada como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação

Prof. Dr. José Palazzo Moreira de Oliveira
Orientador

Porto Alegre, agosto de 2008.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Rigo, Sandro José

Integração de Recursos da Web Semântica e Mineração de Uso para Personalização de Sites / Sandro José Rigo – Porto Alegre: Programa de Pós-Graduação em Computação, 2008.

166 f.:il.

Tese (doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2008. Orientador: José Palazzo Moreira de Oliveira.

1. Hipermídia Adaptativa. 2.Web Semântica. 3.Mineração do Uso da Web. I. Oliveira, José Palazzo Moreira. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. José Carlos Ferraz Hennemann
Vice-Reitor: Prof. Pedro Cezar Dutra Fonseca
Pró-Reitora de Pós-Graduação: Profa. Valquíria Linck Bassani
Diretor do Instituto de Informática: Prof. Flávio Rech Wagner
Coordenadora do PPGC: Profª Luciana Porcher Nedel
Bibliotecária-Chefe do Instituto de Informática: Beatriz Regina Bastos Haro

AGRADECIMENTOS

Em primeiro lugar, agradeço à minha família, pelo apoio e pela compreensão durante todo o processo do doutorado. Em especial, pelo entendimento e suporte durante as ausências durante estes quatro anos.

Agradeço à Renata Vieira, pelos exemplos e motivação para a pesquisa, em especial por ter indicado, mesmo que indiretamente, sugestões para este projeto.

Agradeço em especial ao meu orientador, professor Palazzo, por ter me orientado, trocado idéias, discutido possibilidades de trabalho e estimulado permanentemente.

Deixo agradecimentos às entidades que apoiaram o desenvolvimento do trabalho e a sua divulgação, FAPERGS e CNPQ. Também agradeço à UFRGS, pela possibilidade do curso.

Não poderia deixar de agradecer aos alunos da UNISINOS que acompanharam o trabalho e que implementaram etapas importantes, demonstrando interesse e comprometimento. Também deixo agradecimentos aos colegas da UNISINOS e da UFRGS que colaboraram durante o desenvolvimento deste trabalho, de diversas formas.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

CSS	Cascading Style Sheet
DAML	DARPA Agent Markup Language
DSL	Domain Specific Language
DTD	Document Type Definition
EAD	Educação a Distância
ER	Entity Relationship
HTML	Hypertext Markup Language
IP	Internet Protocol
JSP	JavaServer Pages
OIL	Ontology Inference Layer ou Ontology Interchange Language
OWL	Ontology Web Language
PHP	PHP Hypertext Preprocessor
PLN	Processamento de Linguagem natural
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RDQL	RDF Data Query Language
RST	Rethorical Structure Tool
SGBD	Sistema Gerenciador de Bases de Dados
SMIL	Synchronized Multimedia Integration Language
SQL	Structured Query Language
UML	Unified Modeling Language
WML	Wireless Markup language
XML	Extensible Markup Language
XSL	Extensible Stylesheet Language
XSLT	XSL Transformations

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Um dos motivos para o crescente desenvolvimento da área de mineração de dados encontra-se no aumento da quantidade de documentos gerados e armazenados em formato digital, estruturados ou não. A Web contribui sobremaneira para este contexto e, de forma coerente com esta situação, observa-se o surgimento de técnicas específicas para utilização nesta área, como a mineração de estrutura, de conteúdo e de uso. Pode-se afirmar que esta crescente oferta de informação na Web cria o problema da sobrecarga cognitiva. A Hipermídia Adaptativa permite minorar este problema, com a adaptação de hiperdocumentos e hipermídia aos seus usuários segundo suas necessidades, preferências e objetivos. De forma resumida, esta adaptação é realizada relacionando-se informações sobre o domínio da aplicação com informações sobre o perfil de usuários.

Um dos tópicos importantes de pesquisa em sistemas de Hipermídia Adaptativa encontra-se na geração e manutenção do perfil dos usuários. Dentre as abordagens conhecidas, existe um contínuo de opções, variando desde cadastros de informações preenchidos manualmente, entrevistas, até a aquisição automática de informações com acompanhamento do uso da Web. Outro ponto fundamental de pesquisa nesta área está ligado à construção das aplicações, sendo que recursos da Web Semântica, como ontologias de domínio ou anotações semânticas de conteúdo podem ser observados no desenvolvimento de sistemas de Hipermídia Adaptativa. Os principais motivos para tal podem ser associados com a inerente flexibilidade, capacidade de compartilhamento e possibilidades de extensão destes recursos.

Este trabalho descreve uma arquitetura para a aquisição automática de perfis de classes de usuários, a partir da mineração do uso da Web e da aplicação de ontologias de domínio. O objetivo principal é a integração de informações semânticas, obtidas em uma ontologia de domínio descrevendo o *site* Web em questão, com as informações de acompanhamento do uso obtidas pela manipulação dos dados de sessões de usuários. Desta forma é possível identificar mais precisamente os interesses e necessidades de um usuário típico. Integra o trabalho a implementação de aplicação de Hipermídia Adaptativa a partir de conceitos de modelagem semântica de aplicações, com a utilização de recursos de serviços Web, para validação experimental da proposta.

Palavras-Chave: Hipermídia Adaptativa, Web Semântica, Mineração do Uso da Web.

Integrating Semantic Web Resources and Web Usage Mining for Websites Personalization

ABSTRACT

One of the reasons for the increasing development observed in Data Mining area is the raising in the quantity of documents generated and stored in digital format, structured or not. The Web plays central role in this context and some specific techniques can be observed, as structure, content and usage mining. This increasing information offer in the Web brings the cognitive overload problem. The Adaptive Hypermedia permits a reduction of this problem, when the contents of selected documents are presented in accordance with the user needs, preferences and objectives. Briefly put, this adaptation is carried out on the basis of relationship between information concerning the application domain and information concerning the user profile.

One of the important points in Adaptive Hypermedia systems research is to be found in the generation and maintenance of the user profiles. Some approaches seek to create the user profile from data obtained from registration, others incorporate the results of interviews, and some have the objective of automatic acquisition of information by following the usage. Another fundamental research point is related with the applications construction, where can be observed the use of Web semantic resources, such as semantic annotation and domain ontologies.

This work describes the architecture for automatic user profile acquisition, using domain ontologies and Web usage mining. The main objective is the integration of usage data, obtained from user sessions, with semantic description, obtained from a domain ontology. This way it is possible to identify more precisely the interests and needs of a typical user. The implementation of an Adaptive Hypermedia application based on the concepts of semantic application modeling and the use of Web services resources that were integrated into the proposal permitted greater flexibility and experimentation possibilities.

Keywords: Adaptive Hypermedia, Semantic Web, Web Usage Mining.

INTRODUÇÃO

Este capítulo apresenta a motivação do trabalho, descreve o seu escopo e objetivos, identificando os temas tratados e a abordagem adotada. Nele também é descrita a organização geral do texto.

1.1 Motivação

Este trabalho descreve uma proposta para a aquisição automática de perfis de classes de usuários, a partir da Mineração do Uso da Web e da aplicação de ontologias de domínio. O objetivo principal é a integração de informações semânticas, obtidas em uma ontologia de domínio descrevendo o *site* Web em questão, com as informações de acompanhamento do uso obtidas pela manipulação dos dados de sessões de usuários. Desta forma é possível identificar mais precisamente os interesses e necessidades de um usuário típico e também é possível obter maior flexibilidade na geração de adaptações.

A integração de recursos de Mineração do Uso da Web, ontologias de domínio e tratamento de perfis de classes de usuários é utilizada para a geração de uma aplicação de Hipermídia Adaptativa, que possibilita a adaptação de estrutura em um *site* Web. Este tipo de aplicação possui como justificativa a melhoria da experiência dos usuários em um contexto de excesso de informações. Esta situação pode ser observada atualmente na Internet, onde a enorme quantidade de documentos disponibilizados é apontada como elemento gerador de dificuldades à tarefa de localização de informação relevante pelos usuários.

Pode-se afirmar que a crescente oferta de informação na Web cria o problema da sobrecarga cognitiva. Com a enorme quantidade de documentos disponíveis, o acesso e a coleta da informação desejada pode se tornar uma tarefa difícil e originar resultados de baixa qualidade. A adaptação de *sites* Web permite minorar este problema, quando seus conteúdos ou estrutura são apresentados de acordo com o perfil de uma classe de usuários.

O presente trabalho possui como foco principal a geração de adaptações em aplicações Web que se caracterizam por uma grande quantidade de informações publicadas frequentemente, com suporte para a criação e edição de conteúdos. Este comportamento dinâmico é aproveitado para a identificação de informações a serem utilizadas nos processos de adaptação.

A Hipermídia Adaptativa, segundo Brusilovsky (2001), possui como objetivo a adaptação de hiperdocumentos e hipermídia aos seus usuários, segundo suas necessidades, preferências e objetivos. A adaptabilidade pode estar relacionada com diversos aspectos de um *site* Web, tais como seu conteúdo e apresentação ou sua estrutura. De forma resumida, esta adaptação é realizada relacionando-se informações sobre o domínio da aplicação com informações sobre o perfil de usuários, empregando

conjuntos de regras específicas para a detecção de contextos e ações adequadas de adaptação.

Um dos tópicos importantes de pesquisa em sistemas de Hipermídia Adaptativa é a geração e manutenção do perfil dos usuários. Dentre as abordagens conhecidas para esta tarefa, existem opções como o uso de cadastros preenchidos manualmente, integração de informações parciais em bases de dados, entrevistas e registros em sistemas de informação, compartilhamento de dados de perfis em repositórios de dados e, por fim, aquisição automática de informações a partir de acompanhamento de uso. Esta última alternativa é bastante adequada para situações observadas na Web atualmente, tais como a grande quantidade de usuários e a diversidade de interesses demonstrados.

A identificação do perfil pode ser considerada para usuários ou para classes de usuários. Em alguns contextos é desejável ou necessário que o usuário do sistema seja identificado individualmente e assim, durante o processo de adaptação, as informações de seu perfil serão acessadas e atualizadas. Em outras situações esta identificação pode não ser desejada pelos usuários e pode não ser fundamental. Levando-se em conta o grande número de usuários em determinados contextos, o tratamento individualizado do perfil pode representar dificuldades práticas de implementação. Acredita-se que seja possível a obtenção de bons resultados, em determinadas situações, tratando-se a adaptação do *site* Web sob a perspectiva de uma classe de usuários e não de um usuário específico (Vasilyeva et al. 2007). Esta classe de usuários pode ser definida com base nos conteúdos acessados ou a partir de tarefas realizadas.

A Mineração de Dados é uma área na qual existe o importante envolvimento de uma extensa comunidade de pesquisa. Um dos motivos para seu crescente desenvolvimento encontra-se no aumento da quantidade de documentos gerados e armazenados em formato digital, estruturado ou não. A Web contribui sobremaneira para este contexto, com enormes quantidades de novos conteúdos sendo publicados diariamente. De forma coerente com esta situação, observa-se o surgimento de técnicas específicas para o tratamento de documentos originados na Web, que podem ser agrupadas em mineração de estrutura, conteúdo e uso, segundo Mobascher (2005). Na mineração de estrutura da Web são observadas as relações entre *sites* e partes destes, indicadas através de ligações com *hyperlinks*. Na mineração de conteúdo são analisados os componentes dos *sites* e documentos, normalmente com auxílio de recursos de Processamento de Linguagem Natural.

A Mineração do Uso da Web origina-se em trabalhos anteriores de Mineração de Dados e tem por objetivo a descoberta automática ou semi-automática de padrões de acesso gerados por usuários de *sites* Web, de tal forma que esta informação possa ser utilizada em sistemas de recomendação ou sistemas voltados à personalização (Mobasher e Dai, 2004). Tipicamente, ela é utilizada em diversas situações onde são necessárias a geração e manutenção de perfis de usuários. Uma das vantagens desta abordagem é a utilização de informações geradas pelos acessos de usuários como base para a mineração, o que favorece a obtenção automática e atualizada de padrões (Romero et al., 2007).

Analisando-se as abordagens conhecidas para a geração de perfis de usuários com Mineração do Uso da Web, pode ser observado um padrão mais geral (Markellou et al., 2005; Woon et al., 2005), que envolve etapas bem definidas, resumidas a seguir. A primeira trata da aquisição dos dados de uso. A segunda etapa é dedicada ao pré-

processamento dos dados, com sua discretização, identificação de sessões de acesso, entre outras necessidades de ajustes em função de particularidades do ambiente Web (servidores *proxy*, *cookies* ou erros de acesso, por exemplo). Ao final desta etapa os dados são organizados em formatos adequados para a etapa de descoberta de padrões, onde podem ser geradas, por exemplo, regras de associação e agrupamentos, ou então identificados percursos freqüentes. Na etapa seguinte ocorre a análise e utilização destes padrões em aplicações ou contextos específicos.

As informações de uso consideradas neste trabalho foram os padrões seqüenciais de acesso. Cada visualização de uma página Web é considerada como uma ocorrência de acesso para estes padrões. Mais especificamente, foram tratados os padrões seqüenciais freqüentes, que consistem em um conjunto de acessos observados de forma repetitiva, dentro de limites desejados quanto ao número de ocorrências. Quando a adaptação é gerada apenas com a informação de uso, geralmente estes padrões freqüentes são utilizados a cada interação do usuário para verificar se o percurso atual do mesmo coincide parcialmente com algum dos padrões freqüentes. Caso exista coincidência, estima-se que o usuário apresentará interesse pelas demais páginas do padrão, que podem ser então sugeridas como adaptação de estrutura.

Portanto, a partir da análise e validação dos padrões obtidos é possível a sua utilização como informação complementar em sistemas de Hipermídia Adaptativa. Entretanto, deve ser ressaltado que os padrões gerados desta forma utilizam apenas as informações de acesso de uma sessão de usuário. Informações mais precisas poderiam ser obtidas relacionando-se estes dados de uso com a semântica associada às páginas que cada um representa. Ou ainda, relacionando-se os acessos entre si, buscando identificar relações entre as páginas. Supondo um conjunto de três páginas no qual o conteúdo da primeira seja o enunciado de um exercício, a segunda página contenha a resposta deste mesmo exercício e a terceira página apresente um exercício mais aprofundado sobre o mesmo tema. Este conjunto pode ser representado em um padrão de acessos freqüentes. Entretanto, no formato observado em grande parte dos trabalhos de mineração, este padrão de acesso seria utilizado sem diferenciação para com outro padrão que contenha três páginas com três tipos de exercícios diferentes, ou páginas com apresentação geral de três tópicos sem dependência entre si. Na primeira situação existe um comportamento indicando a busca de aprofundamento do conhecimento em um único tema. Nas demais existe o acesso a vários temas, sem aprofundamento. A integração de informações semânticas às informações de uso possibilita que sejam melhoradas as decisões de adaptação (Eirinaki et al., 2006).

A figura 1.1 identifica um dos problemas relacionados com esta abordagem. São ilustrados dois padrões de acesso e uma estrutura simplificada de um *site* Web. O primeiro padrão de acesso (a) é composto pelas páginas circuladas com traço contínuo. O segundo (b) está circulado pelo traço pontilhado. Utilizando-se apenas a informação de acesso não é possível identificar que o primeiro representa uma navegação em tópicos mais gerais (traço contínuo) e o segundo uma navegação em um tópico e nas opções mais detalhadas deste (traço pontilhado). Entretanto, esta identificação permite melhorar as decisões de adaptação, justamente por acrescentar ao padrão este significado.

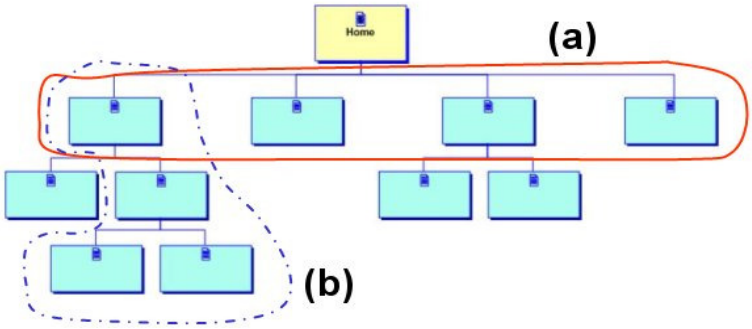


Figura 1.1: Padrões de acesso e sua interpretação

Para que seja possível esta forma de tratamento, torna-se fundamental que a aplicação Web possua uma descrição adequada. Isso pode ser realizado com a utilização de recursos da Web Semântica, que objetiva a disponibilização de documentos de modo que possam ser utilizados de forma automática por diversas aplicações e nos quais sejam superadas as limitações de interpretação e uso atualmente encontradas nas aplicações Web (Berners-Lee et al., 2001). O suporte necessário consiste em recursos para a descrição estruturada dos documentos, a publicação de ontologias e a utilização de mecanismos de inferência para o relacionamento destas informações. No contexto da Web Semântica, o termo "ontologia" está associado a formas para descrever precisamente conceitos e relações entre estes conceitos, a partir de um determinado domínio e de um consenso para uma determinada comunidade de usuários (Freitas, 2003). Assim podem ser superadas limitações, observadas atualmente na Internet, como a dificuldade de identificação do significado, de descrição precisa do formato, de indicação de origem ou de finalidade de documentos publicados (Fensel, 2002). Neste trabalho, são utilizadas as possibilidades de anotação semântica de documentos, o que facilita o tratamento dos conteúdos e seu ajuste, tanto em relação à interface como em relação ao conteúdo disponibilizado – características importantes para um sistema adaptativo. Além disso, ontologias são utilizadas para a descrição de estruturas e processos associados a *sites* Web e tarefas de adaptação.

As anotações semânticas, baseadas em ontologias, são utilizadas em diversos contextos de recuperação de informações, por exemplo, nos quais os conteúdos podem ser tratados a partir de seus significados e não apenas de sua sintaxe (Belew, 2000; Davies et al., 2004). Também são conhecidas outras situações onde recursos similares são utilizados para a anotação semântica de material didático (Aroyo et al., 2003; Nilsson et al., 2003) e para a descrição de Serviços Web e processos (Medjahed, 2003; Gangemi, 2003). Ontologias são utilizadas no apoio a personalização, em sistemas bastante diferenciados, tais como em sistemas de notícias (Aggarwal e Yu, 2002) ou no apoio a sistemas de recuperação de informações (Kim et al., 2003).

Podem ser encontradas diversas abordagens para a implementação de aplicações de Hipermídia Adaptativa (Bailey et al., 2002; Wu, 2002; De Bra et al., 2003; De Bra et al. 2007; Brusilovsky, 2007), sendo que recentemente observa-se o uso de recursos da Web Semântica, como ontologias de domínio ou anotações semânticas de conteúdo (Zimmerman et al., 2005; Schwabe et al., 2004; Oliveira e Muñoz, 2004; Tran et al., 2006; Aroyo et al., 2006). Os principais motivos para tal podem ser associados com a inerente flexibilidade, capacidade de compartilhamento e possibilidades de extensão destes recursos. As anotações semânticas de conteúdo permitem que estes sejam compartilhados entre aplicações e tratados com maior precisão em situações ligadas à

adaptação ou recuperação de informações. Já as ontologias de domínio permitem que seja definido um modelo conceitual para determinada aplicação, que pode ser tratado formalmente com mecanismos de inferência. Estes recursos também possibilitam a modelagem semântica de aplicações Web, envolvendo a descrição de modelos conceituais, modelos de navegação e modelos de apresentação. Uma das vantagens a ser destacada nestas iniciativas, inclusive com bastante importância para a área de Hipermídia Adaptativa, é o fato de existir uma descrição semântica associada aos conteúdos disponibilizados, permitindo assim diversas ações específicas de adaptação.

Diversos métodos que empregam recursos da Web Semântica na construção de aplicações de Hipermídia Adaptativa são conhecidos. O método XWMF (Klapsin, 2001), utiliza RDF[1] e uma linguagem Orientada a Objetos para modelagem de aplicações, que podem ser manipuladas por uso de recursos de inferência em Prolog. Já na metodologia do OntoWebber (Jin et al., 2001) pode ser observada a separação da aplicação em camadas para integração, composição e geração de resultados. A estas camadas estão associadas ontologias para descrição de navegação, conteúdo, apresentação e personalização. O mecanismo de inferência TRIPLE[2] é usado para a manipulação das instâncias de dados, mantidas em RDF. A metodologia do SEAL (Maedche et al., 2003) também é definida a partir do uso de ontologias e utiliza o framework KAON[3] para a implementação das aplicações. O método SHDM (Lima, 2003) trata o projeto de uma aplicação Web a partir de cinco etapas (coleta de requisitos, projeto conceitual, projeto navegacional, projeto de interface abstrata e implementação), sendo que em diversas destas etapas podem ser utilizadas ontologias na representação dos componentes. A flexibilidade atribuída ao uso desta forma de desenvolvimento pode ser interessante para sistemas adaptativos, permitindo que o modelo de domínio, o modelo de usuário, modelo de adaptação, encontrados na maioria dos sistemas adaptativos, sejam associados com modelos de apresentação e navegação (Schwabe, 2004) ou com componentes específicos voltados para a adaptação e meta-adaptação (Assis, 2005). O portal semântico Liferay (Tran et al., 2007) também utiliza o framework KAON e baseia a construção das opções de adaptação em ontologias e regras.

Atualmente observa-se o rápido crescimento de pesquisas relacionadas com o desenvolvimento de sistemas de Hipermídia Adaptativa. Alguns autores associam este movimento ao desenvolvimento da Web, que traz consigo o desafio de atendimento das necessidades de um número muito grande de diferentes usuários, utilizando por vezes diferentes dispositivos. Ao mesmo tempo em que estes desafios são postos pelo desenvolvimento da Internet, esta possibilita uma plataforma real para a experimentação e avaliação destes sistemas (Magoulas e Chen, 2005).

De forma alinhada a este contexto, novas técnicas de Mineração de Dados são desenvolvidas para atender às características específicas de sistemas desenvolvidos para a Web. A análise e utilização dos dados de uso favorecem situações tais como Recuperação de Informação, Sistemas de Recomendação ou sistemas de Hipermídia

---

[1] http:// www.w3.org/RDF/

[2] http:// triple.semanticweb.org/

[3] http://kaon.semanticweb.org/

Adaptativa (Scime, 2005). Por fim, recursos adicionais, descritos pela iniciativa da Web Semântica, são cada vez mais utilizados para proporcionar que uma maior quantidade de informações possa ser associada aos documentos e serviços disponíveis na Web. Esta iniciativa já possibilita que sejam tratadas com maior eficiência algumas tarefas de localização e descrição de recursos (Belew, 2000) e também outras tarefas como a descrição e implementação de aplicações (Assis et al., 2006).

Apesar de serem observadas iniciativas no sentido de acompanhamento de uso da Web e também iniciativas de modelagem semântica de aplicações Web, iniciativas integrando estas abordagens ainda são pouco conhecidas. Entretanto a integração destes recursos permite que sejam obtidas informações para atender ao desenvolvimento de um dos tópicos importantes de pesquisa na área de Hipermídia Adaptativa, que é a geração e manutenção do perfil de usuários. Também a modelagem semântica das aplicações permite que sejam disponibilizadas informações adequadas para a tarefa de adaptação. Portanto este trabalho trata de etapas importantes no processo de adaptação de *sites* Web, como o acompanhamento de uso e a modelagem da aplicação, vindo a contribuir com o estudo e definição de um mecanismo para a aquisição automática de dados de uso da Web, com a análise de possibilidades de tratamento destes dados e a obtenção de padrões de acesso, com a exploração de opções de descrição semântica de conteúdos de *sites* e de aplicações Web, com a utilização e integração destas informações para a geração de classes de usuários e, por fim, com a experimentação e análise da aplicação destas informações na geração de adaptações.

Duas áreas fundamentais para os sistemas de Hipermídia Adaptativa são a obtenção de informações para a composição do perfil de usuários e a descrição das informações das aplicações para que exista a flexibilidade necessária às tarefas de adaptação. Acredita-se que estas tarefas possam ser realizadas de forma eficiente com a utilização conjunta de tecnologias associadas à Web semântica e Mineração do Uso da Web.

A figura 1.2 busca resumir tópicos do trabalho realizado, dentro de uma perspectiva geral, indicando as principais áreas tratadas e a sua função dentro do escopo de desenvolvimento. Para isso são identificadas as áreas de conhecimento, os temas envolvidos no trabalho e o subsídio proporcionado por cada um, além de sua forma de integração com os demais. Nos capítulos seguintes são apresentadas brevemente características destas tecnologias e é realizada uma análise geral da área de Hipermídia Adaptativa, com a apresentação de exemplos deste tipo de sistema. Também são descritas características de metodologias para especificação de aplicações Web e apresentam-se sistemas relacionados, identificando-se pontos positivos e fraquezas em iniciativas similares ou complementares. Procura-se, com a abordagem destes conteúdos, o estabelecimento de uma visão geral que possibilite a identificação das possibilidades de utilização conjunta dos recursos indicados. A seguir, relacionado com os conteúdos citados, são descritos os objetivos, abordagem e escopo das questões de pesquisas tratadas neste trabalho.

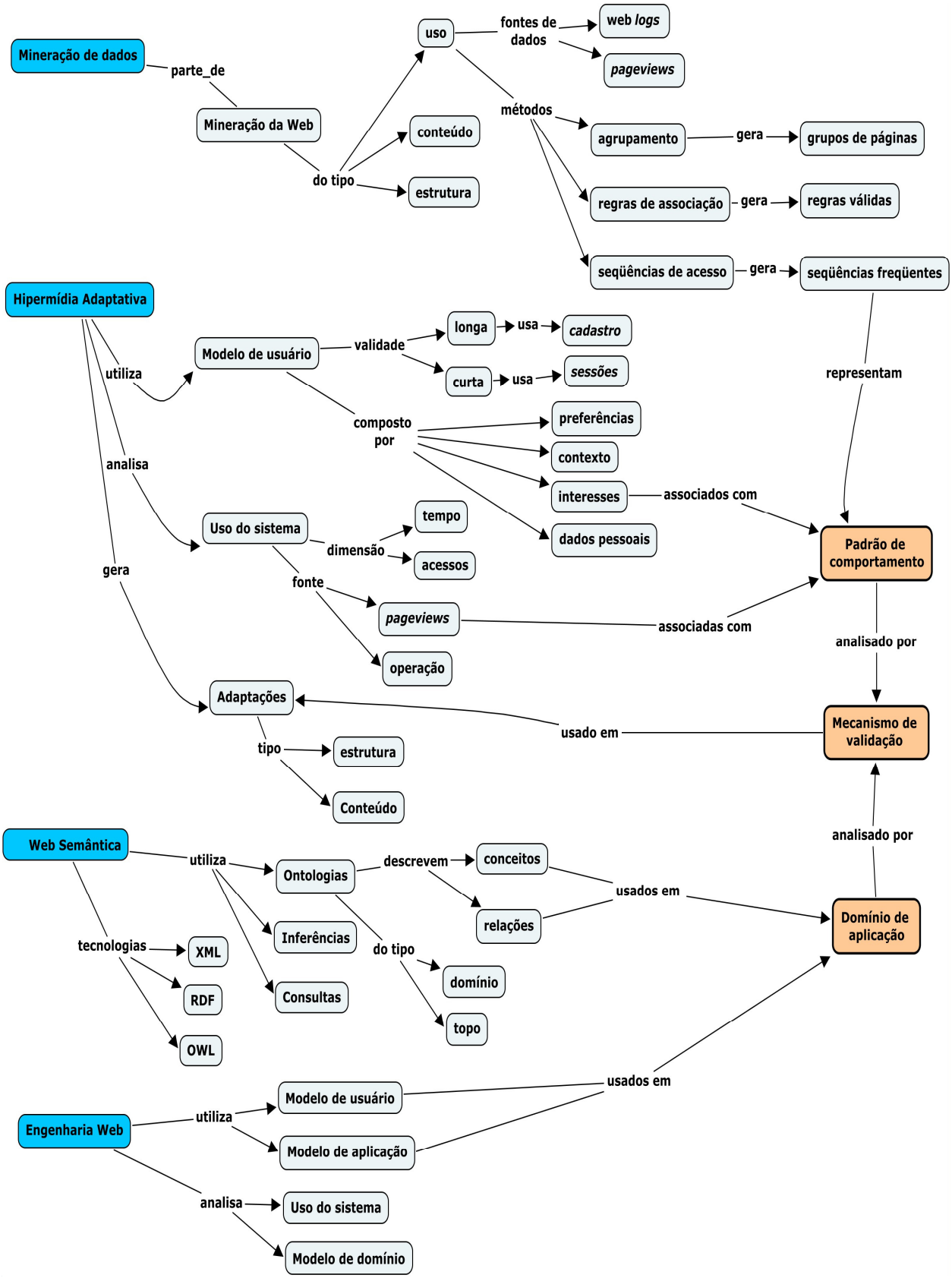


Figura 1.2: Visão geral das áreas e objetos envolvidos neste trabalho

1.2 Objetivos

Este trabalho descreve uma metodologia para a melhoria de sistemas de Hipermídia Adaptativa que consiste na aquisição automática de perfis de usuários, a partir da Mineração do Uso da Web e da utilização de recursos da Web Semântica, tais como ontologias de domínio. O objetivo principal é a integração de informações semânticas, obtidas em uma ontologia de domínio descrevendo o *site* Web em questão, com as informações de acompanhamento de uso obtidas pela manipulação dos dados de sessões de usuários. Desta forma as possibilidades de identificação de interesses dos usuários aumentam em função do acréscimo de informações contextuais aos padrões obtidos com as tarefas de mineração. O foco do trabalho está associado à questão de aquisição de informações que permitam a identificação de classes de usuários e ao uso destas informações em situações de adaptação.

Com este propósito, foram estudados recursos de ontologias e anotação semântica de documentos como forma de proporcionar uma descrição mais rica dos componentes de *sites* Web. Foram também estudadas as possibilidades existentes com a mineração de uso da Web, que proporciona o acesso a dados importantes para a identificação de atividades de usuários.

Parte-se do pressuposto de que as informações de perfis de usuários possuem componentes válidos por longo prazo, mas também por outros componentes que indicam interesses de curto prazo, sendo que estes últimos apresentam maior dificuldade para a sua declaração por parte do usuário. Serão pesquisados mecanismos de identificação automática dos contextos de interesses de curto prazo. Para esta identificação é especificada uma integração de recursos de Mineração do Uso da Web e Web Semântica.

1.3 Abordagem e estrutura

Para atingir os objetivos acima citados são identificadas a seguir as questões de pesquisa consideradas e a forma adotada para a abordagem de cada uma, dentro da estrutura geral do trabalho.

Objetivo: Mineração de seqüências temporais de navegação em *sites* Web.

Hipótese: é possível identificar seqüências padrão para gerar modelos pré-definidos de navegação.

Comentários: A primeira etapa para o encaminhamento deste tópico está relacionada com a aquisição de dados do uso da Web, para um contexto de um *site* Web específico. São analisadas formas conhecidas de tratamento para os dados de uso. É proposto um processo geral de acompanhamento de Uso da Web. A abordagem sugere a utilização de recursos adicionais de programação associados à geração das páginas Web, uso de recursos como *cookies* e manipulação de dados em formato XML. Para seu estudo e validação são propostos experimentos, sendo que o resultado dos mesmos está sumarizado em publicação (Oliveira e Rigo, 2006; Rigo e Oliveira, 2006) que trata deste tópico e da integração destes dados com mecanismos que proporcionem a geração de padrões.

A partir do conjunto de dados originados no acompanhamento do uso podem ser realizadas diversas ações para que sejam detectados padrões de uso e que estes possam ser utilizados como informações para etapas de adaptação de

sites Web. Para o encaminhamento desta questão foi implementado o mecanismo de detecção de percursos freqüentes, baseado no algoritmo SPADE e melhorias deste (Leleu et al., 2003; Zaki, 2001). A implementação de procedimentos flexíveis para a adaptação dos dados originais a formatos de entrada para sistemas existentes de mineração de dados possibilita sua utilização para a geração de conjuntos de regras de associação ou de agrupamentos, como forma complementar.

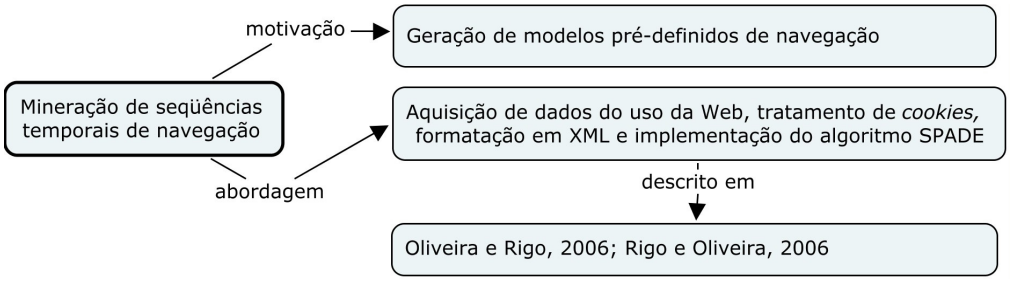


Figura 1.1: Abordagem: mineração de seqüências de navegação em sites Web

A figura 1.3 resume esta etapa do trabalho. A seguir é descrito o próximo tema de pesquisa considerado.

Objetivo: Descrição semântica da estrutura de um *site* Web, com recursos da Web Semântica.

Hipótese: é possível desenvolver uma descrição da estrutura e do relacionamento de componentes de um *site* Web em uma ontologia de domínio que permita a sua manipulação por linguagens de consulta ou mecanismos de inferência.

Comentários: Para melhorar a possibilidade de adaptação foram adotados recursos de anotação semântica para a descrição do conteúdo do *site* Web. Também foi realizada a descrição da estrutura do *site* e de algumas relações significativas em uma ontologia do domínio da aplicação. Uma ontologia permite a definição de conceitos e de relações entre estes, sendo possível descrever conceitos e relações mais gerais ou mais específicos. No presente trabalho foi adotada a abordagem de descrição mais específica, sendo que uma vantagem neste caso é a indicação das relações e dos conceitos de interesse para cada aplicação. Entretanto, esta decisão implica em revisão da ontologia a cada domínio de aplicação diferente, para adaptação e descrição de relações significativas.

A implementação de estudos de caso e validação da abordagem proposta foi realizada com aplicações na área educacional. Assim a ontologia de domínio possui como objetivo descrever conceitos da área educacional. Entretanto a metodologia pode ser aplicada a diversas outras áreas. Esta ontologia foi descrita manualmente por especialistas no domínio da aplicação, com uso do editor de ontologias Protégé[4], tendo sido utilizada a linguagem OWL[5] (*Ontology Web Language*) para sua representação e manipulação. Esta forma de representação proporciona vantagens para etapas posteriores por facilitar

[4] http://protege.stanford.edu

[5] http:// www.w3.org/2004/OWL

a sua manipulação e integração com os dados tratados nos processos implementados. Os resultados desta etapa estão descritos em publicações relacionadas (Rigo e Oliveira, 2006b; Rigo e Oliveira 2007).

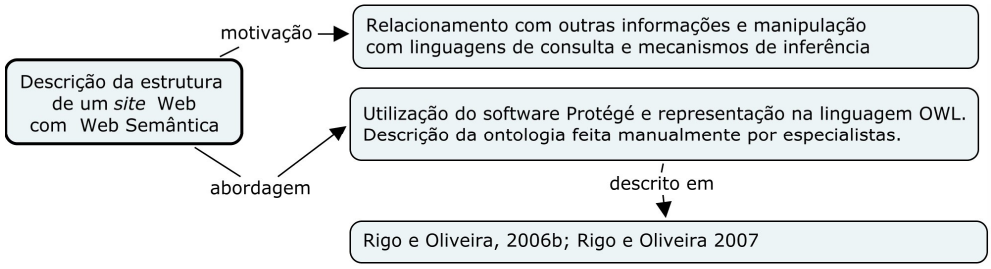


Figura 1.2: Abordagem: descrição semântica de site Web

Uma visão resumida desta etapa pode ser observada na figura 1.4 acima. A seguir é descrito o próximo tema de pesquisa considerado.

Objetivo: Associação das seqüências padrão com a estrutura do *site* Web, para auxiliar processos de adaptação.

Hipótese: é possível a geração de estereótipos para adaptar a estrutura do *site* Web a diferentes padrões de navegação.

Comentários: Para a verificação do uso efetivo destes estereótipos para a adaptação de *sites* Web foi gerado um experimento utilizando como base um sistema de gerenciamento de conteúdo Web disponibilizado em regime de código aberto. Realizaram-se alterações no mesmo para que as etapas de geração de páginas Web levassem em conta a informação de perfis genéricos de navegação, originados da integração do acompanhamento de uso e descrição semântica do *site* Web. A adaptação destas páginas tratou a estrutura de navegação exibida. A modelagem das aplicações Web também está associada a esta questão, sendo que foram analisadas metodologias para descrição de aplicações Web com ênfase no uso de recursos da Web Semântica, pois assim proporciona-se mais flexibilidade e possibilidades de tratamento. Também são analisadas formas de implementação destas aplicações, sendo estudados recursos como serviços Web e linguagens específicas de domínio. Para qualificar as informações obtidas com o acompanhamento do uso estudaram-se as possibilidades de integração entre as informações de uso e as informações semânticas descritas no modelo da aplicação. Neste contexto é identificado o foco principal do trabalho, na integração entre estas informações para a geração de padrões de acesso. Além desta descrição é tratada a forma de utilização destes padrões, bem como sua comparação com padrões gerados com abordagens diversas. Os resultados desta etapa estão descritos em publicações relacionadas (Rigo e Oliveira 2007; Rigo e Oliveira, 2007b; Rigo, Schneider e Oliveira, 2008).

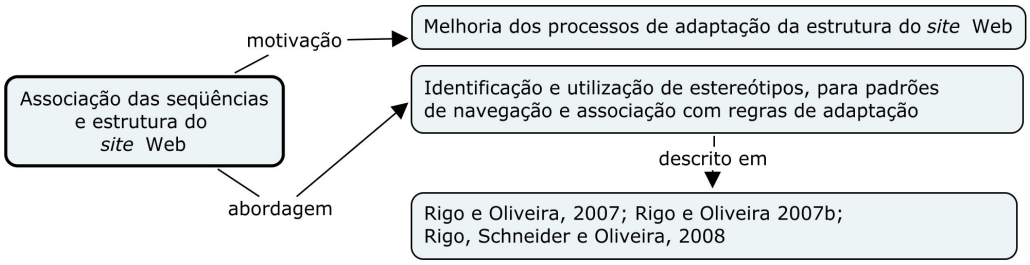


Figura 1.3: Abordagem: Integração semântica com dados de uso

A figura 1.5 resume esta etapa. A seguir é descrito o próximo tema de pesquisa considerado.

Objetivo: Acompanhamento e avaliação da utilização e resultados do processo de adaptação.

Hipótese: a adaptação baseada em integração de semântica e mineração de uso possibilita a melhoria dos resultados do processo e facilita a navegação de usuários.

Comentários: Abordagens para sistemas de Hipermídia Adaptativa apresentam algumas dificuldades quanto à sua avaliação, tanto para avaliação do desempenho como da qualidade final. O sistema proposto pode ser avaliado nos dois quesitos. Com o objetivo de identificar melhorias na qualidade das adaptações, baseadas nas informações de classes de usuários, alguns testes foram realizados para monitorar a quantidade e qualidade das adaptações geradas e a quantidade de acessos a estas sugestões. Experimentos foram realizados ao longo de períodos diversos (seis meses, dez meses) onde o material esteve disponível para acesso, com as informações de adaptação sendo geradas. O contexto de uso foi a aplicação do sistema como suporte para a interação e disponibilização de material em disciplinas de curso de graduação. Alguns testes específicos foram feitos inicialmente para a geração de uma base de dados sintética (isolada) que permitisse uma validação preliminar com respeito às expectativas de resultados. Alguns exemplos destes resultados são a identificação de usuários buscando material sobre um determinado tópico, de usuários acessando o *site* para obter uma visão geral dos conteúdos, acessando material descritivo sobre tarefas, acessando exercícios ou conteúdos complementares Os resultados desta etapa e a visão geral do procedimento proposto estão descritos em publicações relacionadas (Rigo e Oliveira 2007, Rigo e Oliveira, 2008).

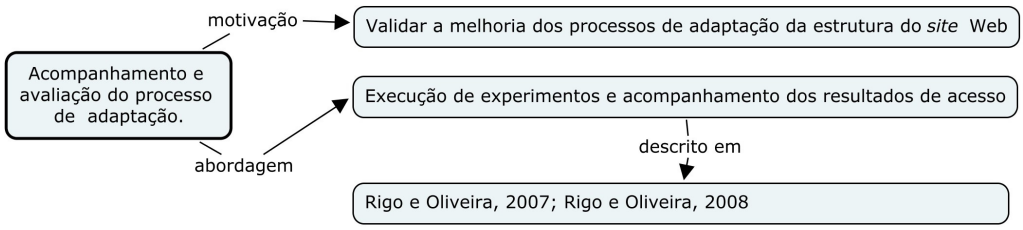


Figura 1.4: Abordagem: Avaliação dos resultados das adaptações

1.4 Contribuições identificadas

As contribuições identificadas nesta tese estão relacionadas com diversos tópicos na área de Hipermídia Adaptativa, Web Semântica e Mineração do Uso da Web. A partir do foco principal do trabalho, a integração de informações de uso e informações semânticas para a geração de classes de usuários, pode-se destacar também alguns tópicos relacionados, tratados ao longo do trabalho e previstos em sua continuação. Segue um breve relato das contribuições observadas:

- Contribuição no estudo, definição e teste de um mecanismo capaz de integrar conhecimento prévio ao processo de Mineração de Dados. No caso, trata-se da integração das informações que constam da descrição de aplicações Web com recursos de Web Semântica e das informações de uso.
- Estudo e identificação de características importantes no tratamento de informações de perfil de usuários, que neste trabalho são considerados a partir da perspectiva de classes de usuários e associados a intenções ou tarefas descobertas implicitamente, a partir de suas ações e do modelo da aplicação.
- Estudo, definição e teste de um mecanismo para o tratamento de adaptações de *sites* Web, que permita a aquisição de informações de uso e também esteja associado a um modelo do *site* Web, permitindo maior flexibilidade nestas duas etapas.
- Definição e implementação do mecanismo de aquisição de dados de uso e seu processamento, para fornecer subsídios a mecanismos já existentes de mineração de dados.
- A implementação de método de identificação de percursos freqüentes, gerados a partir do acompanhamento do uso da Web, observando-se sessões de usuários.

Os trabalhos desenvolvidos e publicados, relacionados a cada uma das etapas são relacionados a seguir, de acordo com o seu foco principal.

Mineração do Uso da Web

Em relação à Mineração do Uso da Web foram publicados dois artigos identificados abaixo, nos seguintes eventos: CLEI 2006 e ERBD 2007. Descrevendo em maiores detalhes a abordagem utilizada no trabalho foi publicado um capítulo de livro ("Handbook of Web Log Analysis"), editado pela editora IGI Global.

Rigo, S. J., Oliveira, J. P. M. d., Wives, L. K. Identifying users stereotypes for dynamic Web pages customization. In B. J. Jansen, A. Spink & I. Itaksa (Eds.), Handbook of Web log analysis. Hershey, PA: IGI. 2007.

Schwertzner, M. A. ; Rigo, S. J. ; Oliveira, J. P. M. Mineração de uso em sistema de informação. In: ERBD - Escola Regional de Banco de Dados, 2007, Caxias do Sul. ERBD, 2007.

Rigo, S. J., Oliveira, J. P. M. Mineração de uso em sites Web para a descoberta automática de classes de usuários. In: Conferência Internacional de la Sociedad Chilena de Ciência de la Computacion, 2006, Santiago, Chile. Conferência Latinoamericana de informática, 2006.

Tratamento e integração de Semântica

A descrição e análises da abordagem para a integração das informações de semântica e de uso foram realizadas em artigos abordando o sistema implementado nos seguintes eventos: WEBMEDIA 2006, WAAMD 2006, WEBMEDIA 2007, SBIE

2007, WISM 2008. Também foram publicados trabalhos abordando possibilidades relacionadas e complementares. Um exemplo é a adaptação de conteúdos, com recursos de Processamento de Linguagem Natural, no evento TIL, junto ao Congresso da SBC em 2007. Outro exemplo é a descrição de uma arquitetura para implementação de aplicações de Hipermídia Adaptativa na Web, baseada em recursos da Web semântica, no evento SBSI 2008.

Rigo, S. J. ; Oliveira, J. P. M. . Aquisição automática de classes de usuários integrando mineração de uso da Web e ontologias. In: WAAMD, 2006, Florianópolis, Brasil. III Workshop em Algoritmos e Aplicações de Mineração de Dados, 2006. v. 2. p. 65-72.

Rigo, S. J. ; Oliveira, J. P. M. . Aquisição Automática de classes de usuário por mineração do uso da Web. In: Webmedia 2006, 2006, Natal, Brasil. Webmedia 2006, 2006

Barbieri, C. ; Rigo, S. J. ; Oliveira, J. P. M. . Classificação de textos baseada em ontologias de domínio. In: TIL 2007 Workshop de Tecnologia da Informação e da Linguagem Humana, 2007, 2007, Rio de Janeiro. Proceedings do Congresso Nacional da SBC, 2007. p. 1640-1649.

Rigo, S. J. ; Oliveira, J. P. M. . Personalização de sites Web integrando mineração de uso e ontologias de domínio. In: WebMedia 2007, the XIII Brazilian Symposium on Multimedia and the Web, 2007, Gramado, RS, Brasil. WebMedia 2007, the XIII Brazilian Symposium on Multimedia and the Web, 2007.

Rigo, S. J. ; Oliveira, J. P. M. . Uso de semântica e mineração de uso Web para identificação de classes de usuários em sites educacionais. In: SBIE - Simpósio Brasileiro de Informática na Educação, 2007, São Paulo - SP. Workshop de Web Semântica e Educação, 2007.

.

Rigo, S. J.; ; Oliveira, J. P. M.; Schneider, E.E., Arquitetura baseada em Web Semântica para aplicações de Hipermídia Adaptativa. In: SBSI, 2008, Rio de Janeiro, RJ. 2008.

Rigo, S. J.; ; Oliveira, J. P. M.; Identifying users stereotypes in educational Websites with Semantic Web resources and Web Usage Mining. Revista Scientia. ISSN 0104-1770. São Leopoldo – Brasil. A ser publicado no segundo semestre de 2008.

Rigo, S. J., Oliveira, J. P. M., Identifying Users Stereotypes with Semantic Web Mining Fifth International Workshop on Web Information Systems Modeling (WISM 2008), ER 2008. A ser publicado.

Na etapa final do trabalho foram escritos alguns trabalhos que encontram-se ainda em revisão. Um deles encontra-se em etapa de avaliação pelo corpo de revisores, para a revista Brasileira de Informática na Educação (RBIE).

Outras iniciativas

A interação com projetos e iniciativas de pesquisa foi realizada sempre que possível, a partir da expectativa de compartilhamento de resultados ou de possibilidades de continuidade do trabalho aqui relatado. Desta forma foram realizados contatos com projetos ligados à Hipermídia Adaptativa e PLN (Renata Vieira – UNISINOS); Acompanhamento de Uso e RI (Paulo Quaresma - ÉVORA); Acompanhamento de Uso e Adaptação de *sites* Web (Guilherme Liberali - Unisinos/MIT); Recursos de adaptação para o contexto de portadores de necessidades Especiais (Édina Fagundes – UNISINOS); Recomendação Semântica e Bibliotecas Digitais (José Palazzo Oliveira – UFRGS).

Também foi realizada a aplicação de subsídios gerados pelo trabalho da tese em situações de implementação ou de adaptação de sistemas de informação para a Web, como no caso do sistema de recuperação de informações e no acompanhamento de uso no *site* da Universidade do Vale do Rio dos Sinos - Unisinos.

Prevendo a continuação do trabalho e a aplicação de recursos de adaptação em sistemas voltados para portadores de necessidades especiais, foram desenvolvidos trabalhos explorando a utilização de mecanismos de síntese de voz. Visando expandir o escopo das adaptações, estes trabalhos exploram possibilidades do mecanismo de adaptação utilizar estes recursos, de forma produtiva, tal como

utilizando mecanismos de síntese de voz para a geração de áudio equivalente ao texto disponibilizado na interface das aplicações. Abaixo segue a relação de três artigos publicados sobre este tema.

Brukshen, M. ; Rigo, S. J. ; Fagundes, E. . Uma solução de acessibilidade baseada em software livre para pessoas com deficiência visual. In: Segundo simpósio Nacional de Tecnologia e Sociedade, 2007, Curitiba - Paraná, Brazil. Uma solução de acessibilidade baseada em software livre para pessoas com deficiência visual, 2007.

Brukshen, M. ; Rigo, S. J. ; Fagundes, E. . Desenvolvimento de software educacional livre e inclusão de alunos com deficiência visual. In: X Ciclo de palestras: Inovações em Tecnologia na Educação: Processos e Produtos, 2007, Porto Alegre. Revista Novas Tecnologias na Educação (ISSN 1679-1916), 2007.

Brukshen, M.; Rigo, S. J. Uma Aplicação Educacional Livre e Acessível para o Ensino de Matemática Fundamental. Fórun internacional de Software livre, Porto Alegre. 2008. Workshop de Software Livre. 2008.

Uma vez evidenciada a necessidade de tratamento de semântica também junto às ferramentas de autoria, foi desenvolvido um protótipo que permite a utilização de padrões homologados por consórcios como o W3C[6] para descrição de aplicação multimídia e que será ampliado, em trabalhos futuros, para a inclusão de características de adaptação. Segue a citação de um artigo desenvolvido neste contexto.

Fonseca, F. M. B.; Souza, v. c. ; Rigo, S. J. . SCOMIL: uma ferramenta livre para criação de conteúdo multimídia baseada nos padrões SCORM e SMIL. In: Simpósio Brasileiro de Informática na Educação, 2007, São Paulo, SP. SCOMIL: uma ferramenta livre para criação de conteúdo multimídia baseada nos padrões SCORM e SMIL, 2007.

O trabalho desta tese foi desenvolvido também no âmbito de orientações de trabalhos de conclusão de curso de graduação na área de Informática e de orientações de monografias de especialização na área de Administração e de desenvolvimento em Software Livre. Durante seu desenvolvimento algumas etapas foram realizadas por alunos de graduação, dentro do escopo de seus projetos de Trabalhos de Conclusão de Curso de Graduação ou Monografias de Especialização. Desta forma, cabe destacar a existência de cinco trabalhos de conclusão de curso de graduação sendo orientados no momento da escrita deste texto (sendo dois em co-orientação), em temas relacionados direta ou indiretamente com o trabalho da tese. Estes trabalhos estão relacionados abaixo e abordam temas como o desenvolvimento de ferramentas para a integração de semântica já no processo de autoria de aplicações Web (um trabalho), a aplicação de recursos de Processamento de Linguagem Natural em textos, provendo condições para a geração de futuras adaptações de conteúdo (dois trabalhos, em co-orientação), a aplicação de recursos de Hipermídia Adaptativa na Educação Especial (um trabalho) e a integração de recursos de Mineração de Uso com Web Semântica (um trabalho).

Trabalhos de Conclusão de Curso de Graduação em andamento

Rodrigo Pereira. Uma ferramenta para criação de modelos de páginas Web baseada em XSLT e CSS. Início: 2007. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos.

Mirian Brukshen. Reconhecimento automático de relações entre entidades mencionadas em textos de língua portuguesa. Início: 2008. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos.

Maico Dionisio Lehmen da Silva. Extração automática da descrição de ontologias em textos da língua portuguesa. Início: 2008. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos.

Ezequiel Pienegonda. Uma proposta de desenvolvimento de aplicações visuais personalizáveis. Início 2008. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) - Universidade do Vale do Rio dos Sinos.

---

6 http://www.w3c.org

Ronan Vargas. Arquitetura para Hipermídia Adaptativa integrando Mineração de Uso da Web e Web Semântica. Início 2008. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) - Universidade do Vale do Rio dos Sinos.

Durante o período da elaboração da tese foram concluídas doze orientações de trabalhos de conclusão de curso de graduação, com temas relacionados direta ou indiretamente com o assunto tratado. Ainda versando sobre assuntos relacionados, foram orientadas quatro monografias de conclusão de cursos de especialiação (MBA Administração de Tecnologia da Informação - UNISINOS e Especialização em Desenvolvimento em Software Livre - UNISINOS). Estes trabalhos estão relacionados a seguir.

Monografias de Conclusão de Curso de Graduação concluídos

Rudi Urigh neto. Análise de adaptatividade e personalização para sistemas de ensino à distância.. 2004. Trabalho de Conclusão de Curso. (Graduação em Informática) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Alexandre Maciel. Análise do uso de ontologias como apoio ao desenvolvimento de um sistema de localização em uma área urbana. 2005. 130 f. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Tales Kunz Cabral. Tecnologia Push associada com a web semântica. 2005. 86 f. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Ari Stopassola Junior. Geração de roteiros turísticos personalizados com o uso de ontologias. 2006. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Igor dos Santos Correa. Estudo sobre Análise de Uso Web e Hipermídia Adaptativa. 2006. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Vinicius Ferla. Acompanhamento de uso como forma de melhoria da relevância na recuperação de informações em sites web. 2006. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Cristiano Barbieri. Classificação de textos baseada em ontologias de domínio. 2006. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Miriam Bruckshen. Solução para interação por voz em sistemas de janelas. Início: 2007. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos. (Orientador).

Everton Schneider. Framework para desenvolvimento de aplicação para a Web Semântica. Início: 2007. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) - Universidade do Vale do Rio dos Sinos. (Orientador).

Carlos Alberto Bastos do Nascimento. Análise de possibilidades: Integração de Mineração de Uso da Web e Ontologias de Domínio. 2007. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Felipe José Nardi Gomes. Estudo sobre aquisição de dados de uso da web e sua aplicação em Hipermídia Adaptativa. 2007. Trabalho de Conclusão de Curso. (Graduação em Sistemas de Informação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Marco Antônio Schwertner. Aplicação de recursos de Hipermídia Adaptativa em um sistema de informação voltado para a áea da Saúde. 2007. Trabalho de Conclusão de Curso. (Graduação em Sistemas de Informação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Monografias de Conclusão de cursos de Especialização concluídas

João Carlos Mindorf. Uso de webservices para geração automática de formulários HTML. 2004. Monografia. (Especialização Desenvolvimento Em Software Livre) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Eliane Rebechi. Integração de tarefas em ambiente web. 2004. Monografia. (Especialização Desenvolvimento Em Software Livre) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

Eduardo da Rosa Devens. Comércio eletrônico: integração de usabilidade, interatividade e marketing nas práticas de negócios on-line. 2007. Monografia. (MBA Administração de Tecnologia da Informação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

James de Assis Silva. Análise de usabilidade e funcionalidades dos sistemas de gerenciamento de projetos: active collab, dot-project e open-project. 2007. Monografia. (MBA Administração de Tecnologia da Informação) - Universidade do Vale do Rio dos Sinos. Orientador: Sandro José Rigo.

1.5 Visão geral do texto

Esta tese está estruturada conforme a descrição a seguir. No capítulo inicial foram descritos a motivação, os objetivos e as etapas da abordagem proposta. Estas consistem na aquisição de dados de uso, seu processamento, a descrição da aplicação, integração de informações de uso e do modelo da aplicação. Neste capítulo estão especificadas as questões que o trabalho se propõe a tratar, dentro do escopo desta pesquisa. Por fim, são descritas as contribuições identificadas.

No capítulo dois são revisados brevemente os principais fundamentos da área da Web Semântica. Como esta é uma área de pesquisa recente, com diversas tecnologias em definição e desenvolvimento, são relacionados os principais recursos empregados no trabalho. Em especial são identificadas informações sobre a linguagem XML, linguagens de manipulação de documentos XML e linguagens para descrição de ontologias, como RDFS e OWL. Também são descritas linguagens de consulta (SPARQL e RDQL). Por fim, são destacados alguns exemplos de aplicações nas áreas de recuperação de informações, anotação semântica e personalização com uso de recursos da Web Semântica.

No capítulo três apresenta-se um resumo dos principais conceitos de Mineração de Dados, sendo descritas brevemente suas etapas principais e também detalhadas características de técnicas de Mineração de Dados voltadas para aplicação na Web, tais como mineração de estrutura, conteúdo e uso da Web. São apresentadas características de algoritmos de Mineração de Uso da Web, para os casos de mineração de percursos freqüentes.

A área de Hipermídia Adaptativa é descrita no capítulo quatro, sendo que são resumidas suas características gerais e traçado um breve histórico de sua evolução, o que pode ser identificado e acompanhado nos exemplos de sistemas destacados e descritos.

No capítulo cinco, são identificadas e analisadas algumas metodologias para projeto de sistemas de informação na Web. Esta área está em desenvolvimento, sendo que as análises e comparações possíveis são de interesse para o trabalho proposto, especialmente aquelas relacionadas com o uso de informações semânticas.

No capítulo seis são descritos brevemente trabalhos relacionados com o trabalho apresentado. Procura-se identificar, nos exemplos destacados, situações onde são empregadas tecnologias associadas com Mineração de Dados ou com recursos da Web Semântica. Estas situações são avaliadas, de modo a contribuir para a descrição de fatores comparativos relacionados com este trabalho.

No capítulo sete são descritos os experimentos realizados para validação da abordagem desenvolvida. Nestes experimentos são abordadas com maiores detalhes as questões de pesquisa propostas. São analisadas as situações que permitem a comprovação da abordagem sugerida.

No capítulo oito são identificados alguns temas para análise e descritas as conclusões do trabalho realizado até o momento. Também estão relacionadas neste capítulo algumas informações a respeito de etapas futuras possíveis.

2 WEB SEMÂNTICA

O objetivo deste capítulo é a apresentação do contexto geral da Web Semântica e a identificação das tecnologias de interesse para o presente trabalho. São apresentadas resumidamente as linguagens utilizadas no trabalho desenvolvido e exemplos de seu uso.

2.1 Visão geral

A iniciativa conhecida como Web Semântica tem como objetivo a resolução de deficiências observadas e a implementação de uma série de melhorias em relação às possibilidades atuais da Internet. Nessa proposta busca-se uma forma de descrição dos documentos a partir de informações mais precisas e também a utilização de terminologias adotadas por comunidades de usuários. Neste sentido, objetiva-se a disponibilização de documentos que possam ser utilizados de forma automática por diversas aplicações e a superação das limitações de interpretação e uso atualmente encontradas. Apesar desta abordagem estar prevista já na proposta original para a organização de documentos na Internet, apenas recentemente vem sendo desenvolvido o suporte necessário, como comentado por Berners-Lee et al. (2001), que consistiria resumidamente em mecanismos mais robustos para a descrição dos documentos, mecanismos para a descrição de ontologias e de mecanismos de inferência para a utilização destas informações.

No momento de escrita deste texto, a iniciativa da Web Semântica encontra-se consolidada em diversos exemplos de aplicações e iniciativas envolvendo não apenas grupos de pesquisa, mas também empresas em um volume significativo. O desenvolvimento de padrões necessários ocorre a partir de consórcio de entidades interessadas, favorecendo a sua consolidação.

O atendimento destes objetivos da Web Semântica pressupõe o atendimento de alguns princípios. Um deles é a identificação universal dos recursos disponíveis, o que pode ser alcançado com o uso da padronização já definida, conhecida como URI (*Universal Resource Identifier*)[7], a partir da qual um determinado recurso pode ser designado e acessado sem limites regionais. Outro princípio a ser observado é a codificação dos documentos, que pode ser feita de forma a assegurar compatibilidade, por exemplo, com o uso da codificação Unicode[8]. Além disso, também é possível o uso

[7] http://www.w3.org/Addressing

[8] http://www.unicode.org

de linguagens de marcação, conforme comentado a seguir, para que os documentos disponibilizados estejam definidos com separações claras entre estrutura e conteúdo, o que facilita o seu tratamento automático (Freitas, 2003). A partir destas possibilidades de identificação inequívoca de documentos disponíveis na Web, codificação de forma a manter a interoperabilidade entre plataformas, passando pela descrição de sua estrutura e conteúdo, configura-se um das visões iniciais da Web Semântica, que seria a de possibilitar a localização inequívoca e também o processamento automático dos documentos.

A seguir, com a descrição de informações diversas sobre o conteúdo dos documentos ou, como colocado por Horrocs (2003), com a descrição de metadados, torna-se possível a correta identificação de seu autor, assuntos tratados e data de criação, para citar algumas possibilidades. O relacionamento dos metadados acerca deste documento com outros contidos em ontologias (comentadas a seguir) permite a implantação de processos mais eficientes de tratamento destes documentos, seja para fins de recuperação de informações, para o apoio ao comércio eletrônico, educação a distância, ou a integração de dados em diferentes repositórios de dados (Hendler, 2002; Nilsson, 2002). Esta identificação de metadados pode ser utilizada para a associação de informações descritoras bem definidas, reconhecidas por uma determinada comunidade de usuários. Para este processo existem mecanismos que permitem a associação de um documento com informações diversas tais como formato, origem, validade, autoria, relacionamentos, entre outros. Estes mecanismos favorecem o surgimento de diversas aplicações que poderão utilizar as informações descritas em diferentes contextos. Linguagens de marcação específicas, comentadas a seguir, possibilitam que sejam implementados estes mecanismos de associação de informações aos documentos, de forma bem definida, permitindo o processamento automático.

Dado que os documentos disponíveis na Internet podem ser livremente gerados e possuirem diversas origens, é importante a possibilidade de verificação, tanto das informações disponibilizadas como do processamento aplicado às mesmas. Neste sentido são previstos alguns recursos para que seja possível o tratamento de provas para resultados obtidos em operações de inferência e de verificação da confiança possível de ser associada às informações. De forma similar às explicações obtidas em Sistemas Especialistas sobre as inferências realizadas para a obtenção de um determinado resultado, existem trabalhos que realizam estas possibilidades, como o projeto Inference Web[9] (McGuinness, 2003).

A padronização de diversas tecnologias e linguagens que colaboram na implementação da Web Semântica vem sendo efetuada no âmbito do W3C (*World Wide Web Consortium*[10]*)*, criado em outubro de 1994 e atualmente contando com a participação de aproximadamente trezentas e cincoenta organizações afiliadas. Diversas tecnologias fundamentais para a concretização dos objetivos da Web Semântica são desenvolvidas pelo consórcio, em diferentes grupos de trabalho, com a participação de entidades diversas. Algumas destas tecnologias, relacionadas ao contexto em questão, são a linguagem de marcação XML (*eXtensible Markup Language*), as linguagens para descrição de recursos na Internet RDF (*Resource Description Framework*) e RDFS (*Resource Description Framework Schema*), a linguagem OWL (*Ontology Web*

9 Disponível em http://www.ksl.stanford.edu/software/iw

10 Disponível em http://www.w3.org

Language) voltada para a descrição de ontologias na Internet e ainda a linguagem SPARQL[11] utilizada para realização de consultas em grafos RDF.

A linguagem XML (Yergeau, 2004) possibilita a descrição mais precisa do conteúdo e de estrutura de documentos, facilitando o seu processamento automático. Recursos como o DTD (*Document Type definition*) e o XML *Schema* permitem que os documentos sejam associados com estruturas pré-definidas, possibilitando a validação automática, o que facilita a implementação de tarefas como a troca de informações e a integração de dados de fontes diversas. Como os marcadores de documentos descritos na linguagem XML podem ser livremente definidos, esta fornece um mecanismo para a descrição da estrutura de documentos de acordo com diferentes contextos de aplicações. Um destes exemplos é justamente a linguagem de descrição de recursos na Web, RDF (descrita a partir de Miller et al., 2004), que implementa um padrão para a descrição de recursos, associação de propriedades a estes recursos e descrição de valores para estas propriedades.

A linguagem de ontologias na Internet (OWL), permite que sejam disponibilizadas informações que descrevem termos ou processos em determinados domínios, de forma precisa e com riqueza de possibilidades para a utilização automática, sendo este um dos suportes básicos necessários para a implementação da Web Semântica. Segundo Heflin (2004), a partir do documento de requerimentos para a OWL, esta linguagem deve prover um suporte consistente para a descrição de ontologias, prevendo seu uso por mecanismos de inferência e também deve facilitar operações de integração entre ontologias diversas. A partir de sua descrição (Smith et al., 2003) pode ser verificada a riqueza de construtores disponibilizados (que não serão descritos neste documento), o que vem facilitando sua adoção como padrão de fato.

A seguir são descritos em maiores detalhes os tópicos da Web Semântica que apresentam relevância para o trabalho desenvolvido. Na medida do possível os exemplos utilizados estão relacionados com etapas desenvolvidas. Não são, contudo, detalhados completamente estes tópicos, em função das características do presente trabalho. O material referenciado contém detalhes adicionais.

2.2 XML (*eXtensible Markup Language*)

Um componente central da Web Semântica é a linguagem de marcação XML[12], sendo que um de seus objetivos iniciais é a promoção da interoperabilidade entre sistemas. Para isso a linguagem XML permite descrever a estrutura dos documentos, a partir de categorias de elementos que o próprio usuário pode definir. Estes elementos são então descritos com marcadores criados pelo usuário, que possuem significado específico dentro de um determinado contexto. Isto permite que os dados sejam descritos com significado específico, abrindo caminho para a associação de semântica aos documentos da Web.

Um exemplo desta forma de aplicação da linguagem XML pode ser observado na figura 2.1 abaixo. Nela pode ser visto um trecho de documento XML com a identificação de seus elementos, tais como o elemento "*acesso*". Este elemento, por sua vez, é composto pelos seguintes elementos: "*ip*", "*page*", "*parametro*", "*agent*",

---

11 Disponível em http://www.w3.org/TR/rdf-sparql-query/

12 http://www.w3.org/XML/

"*data*", "*horario*" e "*userid*". Cada elemento pode ainda estar associado com atributos, como pode ser visto no elemento "*userid*", que possui o atributo "*tipo*" associado ao valor "padrao". No caso deste exemplo os elementos servem para a anotação de um acesso a um *site* Web, sendo identificadas separadamente as informações que descrevem a origem do acesso ("*ip"*), o alvo do acesso ("*page"* e "*parâmetro"*), o momento no qual o acesso ocorreu ("*data"* e *"horário"*) e por fim a identificação do usuário que realizou o acesso ("*userid"*).

```
<acesso>
        <ip>10.20.202.142</ip>
        <page>/congresso/sbc2005/index.php</page>
        <parametro>apresentacao</parametro>
        <agent>Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)</agent>
        <data>26/11/2004</data>
        <horario>18:20:39</horario>
        <userid tipo="padrao">bdd71aabe50937116671e787de3364ed</userid>
</acesso>
```

Figura 2.1: Trecho de arquivo em formato XML

Para oportunizar a interoperabilidade de dados existem mecanismos de verificação padrão associados ao manuseio dos documentos XML. Estes mecanismos de validação possibilitam assegurar que o documento encontra-se de acordo com o formato definido previamente Segundo estes mecanismos, um documento XML pode ser considerado bem-formado quando observa todas as regras sintáticas definidas na sua recomendação, tais como possuir um elemento raiz ou apresentar sempre elementos com a marcação de abertura e finalização definidas. Já um documento considerado válido é aquele que, além de ser bem-formado, possui também conformidade com esquemas de validação como o DTD ou o XML *Schema*, descritos brevemente no item a seguir.

2.3 DTD (Document Type Definition) e XML Schema

Um DTD[13] define a estrutura e sintaxe de um documento XML, sendo usado pelas aplicações na etapa de validação dos documentos, verificando se os mesmos estão em conformidade com a estrutura definida. Este foi o primeiro esquema de validação definido e utilizado em conjunto com o XML, porém possui limitações, tais como a possibilidade de descrição apenas da sintaxe dos elementos e não de alguma semântica. Também existe pouco suporte para a definição de restrições como cardinalidade ou para a descrição de tipos associados aos elementos. Além disso, o documento é descrito em uma sintaxe própria que não segue a sintaxe de documentos XML, fatores estes que motivaram a criação de um esquema mais robusto.

O XML *Schema*[14] têm a mesma função do DTD, mas possui uma maior quantidade de recursos para a definição da estrutura de documentos. Pode-se definir o

---

13 http://www.w3.org/TR/REC-xml/#dt-doctype

14 http://www.w3.org/XML/Schema

tipo e formato exato dos atributos, ou o número de instâncias permitidas um aninhamento. Existem mecanismos de inclusão e derivação que proporcionam o reuso. Estas definições podem ser estruturadas em um documento de estrutura e outro de tipos, fomentando o reuso, conforme comentado. Estes documentos são descritos a partir da sintaxe de documentos XML, o que facilita sua manipulação.

De forma geral, a utilização do padrão XML possibilita maior flexibilidade às aplicações e, quando associada a mecanismos de validação como o XML Schema ou DTD, permite que o tratamento dos documentos seja realizado de forma automatizada e mais efetiva, sem problemas de interoperabilidade observados em outras alternativas. Além da validação de documentos, este mesmo mecanismo permite às aplicações o uso de outros recursos, como a indicação de documentos XSL (*Extensible Stylesheet Language*), contendo regras de transformação a serem aplicadas aos dados dos documentos. No próximo item este formato é descrito brevemente.

2.4 XSL (eXtensible Stylesheet Language)

O XSL[15] é uma linguagem de transformação de documentos XML. Ela permite a descrição de transformação dos dados de um documento XML para a geração de outro documento em formato textual ou HTML (*HyperText Markup Language*), por exemplo. Além deste uso, talvez o caso mais frequente, esta linguagem pode ser utilizada em qualquer outra situação de transformação de dados em XML, servindo, por exemplo, para aplicações de processamento de linguagem natural, compartilhamento de documentos e mineração de dados, entre outras. As transformações podem ser descritas com elementos da linguagem XSL que permitem a identificação de partes específicas dos documentos XML, a utilização de condições de tratamento e a definição de mecanismos recursivos, entre outras características.

Um exemplo de uso conjunto de XML, DTD e XSL, ilustrando a aplicação destas tecnologias no caso de tratamento de dados de acesso obtidos com o monitoramento do uso de um *site* Web, poderia ser esboçado considerando-se o documento XML contendo os dados de uso, como ilustrado na figura 2.1. A partir deste documento é possível a descrição de um documento de validação (em formato DTD ou XML *Schema*) que pode ser associado ao documento original e com isso as aplicações de software terão condições de realizar a validação do documento quanto ao seu formato. As regras de manipulação descritas em um documento XSL podem ser associadas a este documento XML, de forma a indicarem as modificações e manipulações de dados desejadas. A integração indicada está ilustrada na figura 2.2 a seguir. Na primeira linha, está indicado que o documento é descrito no formato XML, na segunda linha está indicado o arquivo com a codificação das regras de transformação em XSL ("manipula_acesso.xsl") e na terceira linha está a referência do documento de validação em formato DTD ("estrutura_acesso.dtd"). Logo após observa-se o conteúdo do documento, com o elemento "acesso", iniciando na linha de número quatro e finalizando na linha de número doze.

---

15 http://www.w3.org/TR/xsl/

```
1   <?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
2   <?xml-stylesheet type="text/xsl" href="manipula_acesso.xsl"?>
3   <!DOCTYPE acesso SYSTEM "estrutura_acesso.dtd">
4   <acesso>
5           <ip>10.20.202.142</ip>
6           <page>/congresso/sbc2005/index.php</page>
7           <parametro>apresentacao</parametro>
8           <agent>Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)</agent>
9           <data>26/11/2004</data>
10          <horario>18:20:39</horario>
11          <userid>bdd71aabe50937116671e787de3364ed</userid>
12  </acesso>
```

Figura 2.2: Integração de dados em XML com DTD e XSL

Esta forma de composição das tecnologias permite a manipulação dos documentos e o seu tratamento para uso em situações específicas. Esta necessidade aumenta com a disponibilização de documentos para a exibição em dispositivos diversos, que utilizam codificação específica. Por exemplo, o mesmo conteúdo descrito em um documento XML pode ser necessário em dois formatos, tais como HTML para navegadores utilizados em microcomputadores de mesa e como WML (Wireless Markup Language) utilizado em dispositivos móveis. A linguagem XSL permite a descrição de dois processos diferentes de manipulação, um para cada formato final desejado. O mesmo conjunto de dados em XML será transformado pelas regras descritas nos dois documentos XSL, gerando os resultados adequados a cada situação.

Entretanto, para que ocorra esta forma de uso, torna-se necessário que as aplicações envolvidas utilizem os documentos a partir de um conhecimento prévio e de um acordo em relação à semântica dos elementos. Não existe, neste contexto, um mecanismo para diferenciar conjuntos de dados descritos com os mesmos marcadores, ou o inverso, para identificar o mesmo conjunto de dados descrito com marcadores de nome diferenciado. A definição do XML deixa livre a criação dos marcadores, como forma de facilitar a sua adoção em situações específicas, sendo que deste modo existe a necessidade de utilização de recursos adicionais para que possa ser adotada uma identificação semântica dos elementos identificadores, em alguns contextos. Recursos como o RDF e o RDF Schema permitem melhores condições de uso neste sentido e serão descritos resumidamente a seguir.

2.5 RDF (Resource Description Framework)

O RDF[16] é uma linguagem para a descrição de metadados sobre recursos disponibilizados na Web. Estes recursos são disponibilizados a partir do uso de espaços de nomes (*namespaces*), descritos com o uso de URIs. Deste modo é viável o compartilhamento de descrições em RDF por um grande número de aplicações, dada a possibilidade de acesso inequívoco. A cada recurso identificado podem ser associados

16 http://www.w3.org/RDF/

um ou mais marcadores descritos em RDF. Assim fica identificado um mecanismo padrão de anotação de metadados, sendo usada a sintaxe do XML como forma de serialização.

Toda a descrição de recursos em RDF está estruturada a partir de um formato padrão composto por expressões (triplas) contendo os seguintes elementos: o sujeito, o predicado e o objeto. O primeiro elemento indica a entidade sobre o qual se deseja descrever algo. Esta entidade é identificada com um URI e pode, portanto, referenciar qualquer elemento que esteja publicado na Web. O predicado é utilizado para indicar a relação ou propriedade associada a esta entidade. Este predicado pode ser uma relação ou propriedade padrão, definida e utilizada por uma comunidade de usuários, ou pode ser algo específico de uma determinada aplicação. Por fim o objeto é o valor associado ao sujeito a partir da relação indicada, podendo tratar-se de um valor literal ou de um outro objeto descrito a partir de um URI.

A figura 2.3 exibe um exemplo de arquivo codificado com o padrão RDF, onde são utilizadas algumas terminologias padrão, que constituem um recurso importante para que os metadados anotados sejam realmente utilizados por um conjunto grande de usuários, além de favorecerem o reuso de definições. O primeiro é o "Dublin Core", voltado para a descrição de publicações. Na figura 2.3 ele está indicado com o prefixo "dc", associado ao seguinte documento: http://purl.oclc.org/dc. O segundo é o "Virtual Card", utilizado para descrição de informações pessoais e indicado na figura pelo prefixo "vc", associado ao seguinte documento: "http://www.w3c.org/2001/vcard-rdf/3.0". No caso deste exemplo estas terminologias auxiliam na descrição de uma pessoa, autor de um trabalho ("dc:Creator") e dados complementares como nome ("v:Name") e email ("v:email"). Com uma terminologia padrão, diferentes aplicações podem utilizar o mesmo documento e os mesmos dados, com um significado descrito de forma explícita.

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3/org/1999/02/22-rdf-syntax-ns"
xmlns:dc="http://purl.oclc.org/dc#"
xmlns:v="http://www.w3.org/2001/vcard-rdf/3.0#">
    <rdf:Description about="http://ww.inf.unisinos.br/~rigo">
        <dc:Creator>Sandro Rigo</dc:Creator>
        <v:Name>Sandro José Rigo</v:Name>
        <v:Email>rigo@unisinos.br</v:Email>
    </rdf:Description >
</rdf:RDF >
```

Figura 2.3: Exemplo de código RDF

Uma declaração RDF pode ser representada através de grafos rotulados, ou diagramas de nós e arcos. Ilustrando esta representação, a figura 2.4 apresenta um grafo resultante da interpretação dos dados descritos na figura 2.3. Observa-se que o elemento de origem dos arcos está identificado com o atributo "about" do elemento

"rdf:Description". Nestes exemplos, os objetos para as três relações indicadas consistem de valores literais, mas poderiam ser outros recursos, indicados por URIs.

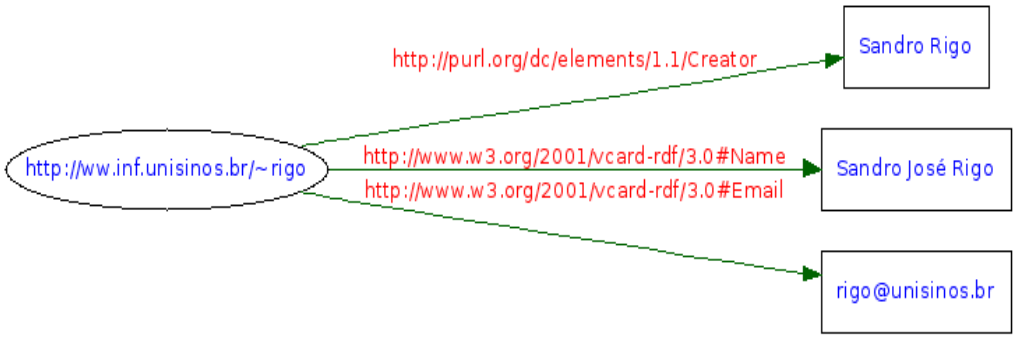


Figura 2.4: Grafo com representação de documento RDF

O RDF permite a descrição e o tratamento de metadados que podem ser associados a recursos na Web, porém não possibilita a descrição de relações mais complexas, o que pode ser obtido com o uso de recursos como ontologias. A seguir são apresentados conceitos gerais sobre ontologias e descritas linguagens para sua especificação.

2.6 Ontologias

O termo "Ontologia" foi utilizado originalmente na Filosofia para designar o estudo das coisas do mundo e suas relações. Quando aplicado ao contexto da Web Semântica e associado com técnicas de Inteligência Artificial, vem descrever formas para permitir que sejam nomeados precisamente conceitos e relações entre estes conceitos, a partir de um determinado domínio e de um consenso para uma determinada comunidade de usuários. Diversos autores tais como Guarino (1997), Sowa (2003) ou Gruber (2003) descrevem estas possibilidades associadas ao uso de ontologias neste contexto, sendo equivalente a uma modelagem conceitual expandida e semanticamente rica.

Uma vez estabelecida esta especificação acerca dos conceitos principais de um domínio, dentro de uma conceitualização aceita por uma comunidade de usuários, podem ser superadas limitações, observadas atualmente na Internet, como a dificuldade de identificação do significado, de descrição precisa do formato, de indicação de origem ou de finalidade de documentos publicados. Este processo, descrito por Fensel (2001, 2002), inicia com a criação de associações entre documentos e conceitos indicados em ontologias, etapa na qual pode ser observada uma das aplicações da linguagem XML e de outras linguagens (como o RDF, por exemplo). Esta associação consiste, resumidamente, na anotação semântica, de modo que os conceitos indicados possam ser empregados para apoio em etapas posteriores, como a recuperação de informações.

Entretanto esta anotação semântica pode não ser suficiente para a realização de operações mais complexas. Para estas situações podem-se utilizar as linguagens de ontologias. A partir da definição de classes, relações de herança e outras relações específicas, ampliam-se as possibilidades de tratamento automático. Operações como inferência, validações diversas ou consultas podem ser executadas sobre os dados relacionados. Ontologias possibilitam que sejam criadas e compartilhadas bases de conhecimento descrevendo entidades importantes para determinados domínios,

estabelecendo relacionamentos diversos entre estas entidades. Deste modo, a utilização de ontologias possibilita a geração de uma nova classe de aplicações na qual amplia-se a possibilidade de tratamento automático de informações. Ou, de outra forma, permite que aplicações já estabelecidas sejam melhoradas, com acréscimo de informações semânticas.

As ontologias podem ser classificadas de acordo com o escopo definido na sua construção. As caracterizações mais conhecidas são de ontologias de topo e ontologias de domínio, representando objetivos opostos. Nas ontologias de topo objetiva-se a descrição para termos de propósito geral, sendo, portanto, bastante ampla a sua utilização Estes termos podem eventualmente não auxiliar em situações específicas, dentro de áreas de conhecimento altamente especializadas. Alguns exemplos de ontologias de topo são encontrados nos projetos Cyc[17,18] ou SUMO[19] (*Suggested Upper Merged Ontology*). O número de termos e relações é normalmente bastante grande neste tipo de ontologia, situando-se na ordem de milhares. Por exemplo, a ontologia SUMO possui aproximadamente vinte mil termos e setenta mil axiomas. Já a ontologia Cyc, indica a quantia de trezentos mil conceitos e três milhões de relações.

Quando uma ontologia é definida com vistas à descrição de uma área de conhecimento específica, recebe a denominação de ontologia de domínio. Uma das principais vantagens desta abordagem é a possibilidade de adequação ao contexto e necessidades específicas, permitindo que os conceitos e relações sejam especificados de forma bastante otimizada. São conhecidos diversos exemplos de ontologias de domínio, tanto em projetos largamente conhecidos e com aplicação ampla, como em situações particulares e restritas. Uma dos casos de ontologia de domínio voltada para a lingüística é o Wordnet[20], que serve para ilustrar também que uma ontologia de domínio pode ser descrita com um conjunto extenso de termos, sendo neste caso composta de aproximadamente cento e vinte mil termos. Áreas como Direito ou Geografia representam situações típicas adequadas à utilização de ontologias de domínio. Exemplos são as ontologias LKIF[21] para termos legais ou a ontologia "*Geographic Information - Metadata* (ISO 19115:2003)"[22] para informações geográficas.

A criação de ontologias pode ser feita de forma manual, a partir da experiência e conhecimento de especialistas em determinados domínios, ou de forma automática ou semi-automática, como pode ser observado em algumas iniciativas. Para qualquer destas formas de geração de ontologias, são importantes alguns princípios, descritos já por Grubber (1993), que favorecem sua qualidade. Tais princípios estão associados com a clareza e legibilidade dos termos utilizados, com a necessidade de coerência e de previsão de extensão da ontologia e, por fim, com o trabalho no sentido de implementar descrições que sejam independentes de formas de codificação, resumidas de modo a consistirem a informação fundamental necessária. Para a utilização de ontologias em

---

17 http://www.opencyc.org/

18 http://www.cyc.com/

19 http://www.ontologyportal.org/

20 http://www.cogsci.princeton.edu/~wn

21 http://www.estrellaproject.org/lkif-core/

22 http://loki.cae.drexel.edu/~wbs/ontology/iso-19115.htm

aplicações, existem alguns recursos como mecanismos de inferência e linguagens de consulta, que podem ser integrados com a tecnologia específica utilizada para o desenvolvimento da aplicação.

A seguir são descritas algumas possibilidades de linguagens para descrição de ontologias. Apesar de existirem outras linguagens, não existe o objetivo de relacionamento e comparação mais ampla e assim este texto trata apenas de duas, escolhidas pela sua grande utilização. Também são comentados a seguir alguns mecanismos de inferência e linguagens de consulta, com o mesmo objetivo de servir apenas como uma referência para algumas possibilidades.

2.7 RDFS (Resource Description Framework Schema)

O RDFS[23] descreve alguns tipos básicos que podem ser utilizados para a descrição de esquemas específicos. Conforme mencionado, desta forma é facilitada a descrição, o compartilhamento e a extensão de vocabulários controlados. Esta extensão pode ser obtida com a criação de novos esquemas que referenciem objetos já identificados anteriormente como recursos de outro esquema. Neste ponto particular temos um exemplo de utilização necessária de URIs e espaços de nomes como forma de garantir que o objeto desejado esteja sendo referenciado.

A partir da especificação do RDFS, encontram-se algumas classes e algumas propriedades cuja semântica está dada, sendo que desta forma podem ser utilizadas na criação de vocabulários compartilhados. Algumas destas são citadas na tabela 2.1 abaixo. Utilizou-se o prefixo "rdfs:" como forma de indicar o uso das definições do RDFS, descritas na seguinte URL: "`http://www.w3.org/2000/01/rdf-schema#`".

Tabela 2.1: Relação de alguns componentes do RDFS

Classes/propriedades	**Descrição**
rdfs:domain	Restrições globais de domínio
rdfs:range	Restrições globais de valores
rdfs:Resource	Classe geral utilizada para descrição de recursos
rdfs:Class	Permite atribuir a um recurso o tipo de classe
rdfs:Property	Possibilita associar a um elemento o tipo de propriedade
rdfs:SubClassOf	Indicação de relação de subclasse
rdfs:SubPropertyOf	Utilizado para descrição de subpropriedades
rdfs:seeAlso	Indica o relacionamento de um recurso com sua definição

Com este conjunto de elementos é possível a descrição de relacionamentos, porém, em função das definições existentes, não é possível a descrição de algumas relações mais expressivas, sendo que para isso são conhecidas outras linguagens, que utilizam os elementos do RDFS, porém acrescentam outras possibilidades de descrição. Uma destas linguagens para a descrição de ontologias, de forma mais expressiva, é descrita a seguir.

---

[23] http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

2.8 OWL (Ontology Web Language)

A linguagem OWL[24] é desenvolvida no contexto de um grupo de trabalho do W3C e deriva de experiências prévias com outras linguagens para a descrição de ontologias, que são a linguagem OIL[25] (*Ontology Inference Layer ou Ontology Interchange Language*) e a linguagem DAML[26] (*DARPA Agent Markup Language*). A linguagem OIL foi a primeira destas linguagens para descrição de ontologias e teve como principal requisito a facilidade de adoção pelos desenvolvedores, servindo principalmente à comunidade ligada ao desenvolvimento de aplicações no âmbito da Web Semântica. Em uma iniciativa de padronização e ampliação de possibilidades de representação, a linguagem DAML-OIL[27] foi criada integrando as duas linguagens referidas (DAML e OIL).

A partir das experiências e da necessidade de ampliação das possibilidades conhecidas, a linguagem OWL[28] foi criada como sugestão de linguagem padrão. A linguagem OWL permite que sejam disponibilizadas informações que descrevem termos ou processos em determinados domínios, de forma precisa e com riqueza de possibilidades para a utilização automática. Existem três sub-linguagens da OWL, diferenciadas de acordo com o grau de expressividade, sendo que a escolha da sub-linguagem a ser utilizada depende da definição da ontologia. Segue um resumo de suas características:

- OWL *Lite*: permite classificar hierarquias de classes e definições simples de propriedades;
- OWL DL: fornece a máxima expressividade que se pode conseguir mantendo-se completude computacional e decidibilidade. Possui todas as construções da linguagem OWL, mas estas devem respeitar algumas restrições. Por exemplo, uma classe pode ser subclasse de muitas outras, mas não pode ser instância de outra classe;
- OWL *Full*: além de possuir todas as construções da linguagem OWL, permite uma maior flexibilidade de tipos. Esta característica fornece máxima expressividade, porém não garante completude computacional e decidibilidade.

Para definir uma ontologia em OWL é preciso, em primeiro lugar, especificar onde estão localizadas, na Web, as classes primitivas, para que se possam definir novas classes como subclasses destas. Também é necessário determinar um espaço de nomes para a nova ontologia. Isto é codificado como no exemplo da figura 2.5 a seguir. As classes a serem definidas estarão localizadas no espaço de nomes da definição, que consta na linha seis da figura. A definição na linha sete serve para que ontologias externas possam referenciar a ontologia que está sendo definida. As definições restantes

---

24 http://www.w3.org/2004/OWL/

25 http://www.ontoknowledge.org/oil/

26 http://www.daml.org/

27 http://www.w3.org/TR/daml+oil-reference

28 http://www.w3.org/TR/owl-ref/

localizam as definições primitivas de OWL (linha cinco), RDF (linha dois), RDFS (linha quatro) e XML Schema (linha três).

```
1   <rdf:RDF
2           xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3           xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
4           xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
5           xmlns:owl="http://www.w3.org/2002/07/owl#"
6           xmlns="http://www.site.com.br/cms01/adaptacao/curso.owl#"
7
    xml:bd="http://www.site.com.br/cms01/adaptacao/curso.owl">
```

Figura 2.5: Identificação de espaços de nomes em uma ontologia descrita em OWL

As classes podem ser construídas de várias formas: por herança, união, interseção, complemento, pela enumeração de instâncias ou por restrições de propriedades. No exemplo seguinte, descrito na figura 2.6, a classe "Revisao_online" é subclasse de "Material_de_Apoio", e o atributo "DescritoEm" pode ser usado para indicar instâncias da classe "Documentos".

```
<owl:Class rdf:ID="Revisao_online">
    <rdfs:subClassOf rdf:resource="#Material_de_Apoio"/>
</owl:Class>
<rdf:Property rdf:ID="DescritoEm">
    <rdfs:domain rdf:resource="#Material_de_Apoio"/>
    <rdfs:range rdf:resource="#Documentos"/>
</rdf:Property>
```

Figura 2.6: Exemplo de construção de classes em OWL

As referências às superclasses primitivas owl:Class e rdf:ID conseguem ser localizadas pelas aplicações de software devido aos espaços de nomes indicados. Outro fato a ser observado é que, em lógica de descrição (*Description Logics*[29]), a definição dos atributos não têm de estar junto com a classe. A classe "Revisão_online" pode ser definida como a subclasse de "Material_de_apoio" que está descrito em ("DescritoEm") arquivos do tipo HTML ("HTMLFile"), ou seja, a classe é definida com o auxílio de uma restrição. Este fato está exemplificado na figura 2.7 a seguir.

---

29 http://dl.kr.org/

```
<owl:Class rdf:ID="Revisao_online">
    <rdfs:subClassOf rdf:resource="#Material_de_Apoio"/>
    <rdfs:subClassOf>
            <owl:Restriction>
                    <owl:onProperty rdf:resource="#DescritoEm"/>
                    <owl:allValuesFrom rdf:resource="#HTMLFile"/>
            </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>
```

Figura 2.7: Descrição de restrições em OWL

A linguagem OWL possui uma maior expressividade do que as linguagens anteriores, sendo que sua adoção com padrão também possui o objetivo de favorecer o compartilhamento e reuso de informações descritas em ontologias. A seguir, na tabela 2.2, são sumarizadas as principais propriedades descritas pela linguagem OWL, segundo sua especificação[30]. Esta tabela possui como objetivo relacionar algumas das características importantes da linguagem para ilustrar sua expressividade e não pretende esgotar o assunto. Entretanto é suficiente para demonstrar a riqueza de opções disponíveis.

Tabela 2.2: Sumarização de propriedades OWL

Propriedade	**Breve descrição**
Symmetric	Se P(x, y) então P(y, x)
hasValue	P(x,y) and y=hasValue(v)
cardinality	cardinality(P) = N
minCardinality	minCardinality(P) = N
maxCardinality	maxCardinality(P)= N
equivalentProperty	P1 = P2
Transitive	if P(x,y) and P(y,z) then P(x, z)
Functional	if P(x,y) and P(x,z) then y=z
InverseOf	if P1(x,y) then P2(y,x)
unionOf	C = unionOf(C1, C2, …)
disjointWith	C1 != C2
equivalentClass	C1 = C2
oneOf	C = one of(v1, v2, …)
sameIndividualAs	I1 = I2
complementOf	C = complementOf(C1)
AllDifferent	I1 != I2, I1 != I31, I2 != I3, …
differentFrom	I1 != I2

As descrições possíveis com a linguagem OWL podem ser serializadas em RDF, possibilitando que sejam utilizadas linguagens de consulta para o aproveitamento

30 http://www.w3.org/2004/OWL/#specs

das informações neste formato de grafo. Esta exploração do esquema do RDF é viabilizada a partir de um conjunto de linguagens de consulta específicas, das quais serão descritas a seguir a RDQL e SPARQL. Estas mesmas informações podem ser utilizadas por mecanismos de inferências, sendo que a análise das características de cada aplicação pode resultar na indicação da forma mais adequada (o uso de linguagens de consulta ou mecanismos de inferência).

2.9 Mecanismos de inferência

As possibilidades de descrição de conceitos e relações de uma ontologia podem ser melhor utilizadas com mecanismos de inferência que geralmente implementam versões de Lógica de Descrição (Baader, 2002), permitindo assim que tarefas como validação dos conceitos e relacionamentos ou inferência sobre propriedades sejam executadas. Com o uso de máquinas de inferência o conhecimento descrito em ontologias e demais recursos da Web Semântica pode ser empregado para a dedução de novas informações, ampliando as possibilidades de aplicações de software. Apesar de existirem como área de pesquisa há pelo menos duas décadas, as lógicas de descrição tomaram um grande impulso com o surgimento de Web Semântica e da possibilidade de sua aplicação em um amplo conjunto de áreas.

Algumas implementações de mecanismos de inferência com Lógicas de Descrição são conhecidos e utilizados em diversas situações voltadas para a Web Semântica, tais como o RAcer (http://www.racer-systems.com/), Triple (http://triple.semanticweb.org/), FACT (http://www.cs.man.ac.uk/~horrocks/FaCT/). Outros são disponibilizados em plataformas comerciais, como na caso da TopQuadrant (http://www.topquadrant.com/) e da IntelliDimension (http://www.intellidimension.com/).

2.10 Linguagens de consulta

Em alguns contextos os dados descritos em formato RDF podem ser utilizados satisfatoriamente com linguagens de consulta. Estas constituem um campo de pesquisa em desenvolvimento, desde a adoção mais expressiva de padrões como XML e RDF. A seguir são descritas informações resumidas sobre as linguagens de consulta RDQL e SPARQL. No presente trabalho estas duas linguagens foram utilizadas em etapas específicas.

2.10.1 RDQL (RDF Data Query Language)

RDQL[31] é uma implementação da linguagem de consulta a documentos RDF, desenvolvida para a API Jena[32], um *framework* na linguagem Java de código aberto para construção de aplicações de Web Semântica. Para que as consultas possam ser realizadas e as informações possam ser extraídas, a linguagem RDQL disponibiliza uma sintaxe que, embora possua algumas particularidades, funciona de maneira similar à linguagem SQL, uma linguagem de consulta para Banco de Dados relacionais. Assim oferece suporte para consultas com utilização de expressões genéricas que envolvem variáveis para nós e arestas, permitindo a utilização de documentos descritos em RDF,

---

31 http://www.w3.org/Submission/RDQL/

32 http://www.hpl.hp.com/semweb/

ou seja, no formato de grafos. A linguagem também permite a utilização integrada de esquemas RDF e suas instâncias, favorecendo o reuso de esquemas.

A utilização da linguagem se dá com um conjunto de recursos que permitem agrupamentos de primitivas, suporte a esquemas RDFS, ações aritméticas, funções de agregação e facilidades para uso e manipulação de espaços de nomes. Por exemplo, na figura 2.8 é ilustrada uma consulta simples em RDQL. O resultado deste exemplo é a recuperação de todas as triplas de determinado documento, dado que não é indicada nenhuma restrição. Esta consulta usa a cláusula "SELECT", que permite selecionar quais as variáveis que serão retornadas como resultado da sua execução. Para a identificação das variáveis é utilizado o sinal de interrogação ("?") antes de cada nome. A cláusula "FROM" permite especificar qual documento RDF será pesquisado, sendo que para isso é utilizado um URI para referenciá-lo.

```
SELECT ?x

FROM <http://www.site.com/tcc/rdf/arquivo.rdf>
```

Figura 2.8: Exemplo de consulta simples utilizando RDQL

A cláusula "WHERE" permite especificar as restrições para a realização das consultas. Essas restrições seguem o formato de tripla (sujeito, predicado e objeto), que podem ser formadas tanto por um objeto quanto por um valor literal. Para os predicados, como forma de abreviação ou simplificação, pode ser utilizado o prefixo determinado na cláusula "USING" para referenciar as propriedades contidas em diversos espaços de nomes, ao invés de utilizar todo o URI.

```
SELECT ?curso
FROM <http://www.site.com/tcc/rdf/exemplo.rdf>
WHERE (?curso, <exemplo:atividadeNoCurso>, 'Ensino'),
      (?y, <exemplo:tipoPessoa>, 'Docente'),
      (?z, <exemplo:cargo>, 'coordenador'),
      (?w, <dc:creator>, 'Joao Silva')
USING arquivo FOR <http://www.site.com/tcc/exemplo/esquema#>,
dc FOR <http://purl.org/dc/elements/1.0/>
```

Figura 2.9: Exemplo de utilização das cláusulas RDQL

Na figura 2.9 está descrito um exemplo de consulta RDQL onde são recuperadas triplas contendo alguns relacionamentos específicos, como "tipoPessoa", "cargo", "creator". Deve se chamar a atenção para o fato de que as duas primeiras relações indicadas são relações descritas em uma ontologia específica e que a terceira é uma relação indicada a partir de uma terminologia bastante conhecida, o Dublin Core.

2.10.2 SPARQL

A linguagem SPARQL[33] incorpora duas características interessantes, como linguagem de consulta para dados em RDF e também como protocolo. É desenvolvida no âmbito de um grupo de trabalho do W3C. Como linguagem de consulta para grafos

[33] http://www.w3.org/TR/rdf-sparql-query/

RDF, ela provê facilidades para tratar informações diversas, tais como URIs ou valores literais. Também permite o tratamento de partes dos grafos e a construção de novos grafos a partir das informações obtidas com as consultas.

A proposta de um protocolo para acesso aos dados em RDF permite que sejam concebidos e acessados serviços Web para a consulta remota a conjuntos de grafos e descreve um protocolo de transporte para o acesso a estes serviços.

A utilização da linguagem, conforme proposta, está baseada na consulta aos grafos RDF e no casamento de padrões, de forma similar ao RDQL. Considerando-se que uma expressão RDF é composta por três componentes (sujeito, predicado e objeto), é possível a utilização de uma variável contendo valores desejados para quaisquer dos três componentes destas triplas. A consulta retorna como resultado todos os padrões encontrados, dada a variável e o componente indicado. Este mecanismo possibilita que sejam realizadas consultas explicitando algumas das relações expressas na ontologia ou buscando valores específicos associados com estas relações.

2.11 Exemplos do uso de tecnologias da Web Semântica

Neste item são relacionadas características de sistemas ou iniciativas empregando informações semânticas para a área de recuperação de informações, anotação semântica e também para aplicações de personalização. O objetivo do item é ilustrar a possibilidade de integração destes recursos permitindo comparações com o trabalho desenvolvido nesta tese.

2.11.1 Aplicações na Recuperação de Informações

Em diversas situações é possível observar a utilização de recursos de ontologias para apoio às tarefas de recuperação de informações, melhorando o resultado obtido, uma vez que informações semânticas sobre os documentos estarão sendo utilizadas, como já descrito por Baeza-Yates (1999). Em geral são utilizadas anotações baseadas em ontologias para a identificação precisa dos documentos, favorecendo a criação de um mecanismo de indexação que permite a superação dos limites impostos pela sintaxe dos termos, com a utilização do seu significado (Belew, 2000).

Em outra abordagem, observa-se o uso da estrutura dos documentos como forma de melhorar o resultado dos mecanismos de recuperação de informação. Para isso são utilizados documentos descritos na linguagem XML, com o aproveitamento de características importantes da mesma, como a facilidade para separação entre conteúdo e estrutura dos documentos, bem como a facilidade para a descrição e para a validação da estrutura de documentos. No trabalho descrito por Hayashi et al. (2000) são utilizadas informações sobre a estrutura dos documentos XML para a criação de índices que irão apoiar a recuperação de informações com a ordenação por relevância de documentos, mapeando um conceito de campos de pesquisa à sub-estruturas de conjuntos de documentos. Em Grabs e Schek (2002) pode ser observada uma extensão desta proposta, na qual observa-se uma maior abrangência a partir da geração de informação de relevância independentemente da estrutura do documento. Em Cohen et al (2003) a estrutura do documento XML é considerada para fins de geração de índices de relevância a serem associados às respostas do sistema de recuperação de informações, a partir de perguntas formuladas pelo usuário e do uso de linguagem de pesquisa específica.

Entretanto, trabalhos neste sentido, apesar de ampliarem as possibilidades do mecanismo geral observado em sistemas de recuperação de informações existentes - como o mecanismo utilizado no sistema de recuperação de informações Google[34], descrito inicialmente por Brin e Page (1998), ou então mecanismos voltados para situações mais específicas, como o sistema Thumba, descrito por Silva (2003) - não utilizam o apoio de ontologias para sua tarefa. Um dos primeiros trabalhos com esta característica foi a linguagem SHOE (*Simple Html Ontology Extensions*), conforme descrito inicialmente por Luke (1996) e mais tarde empregado no contexto da Web Semântica, com uma plataforma mais robusta (Helfin, 2001). Esta abordagem consiste na utilização de diversas ontologias de domínios e em marcadores específicos, utilizados em documentos HTML, o que possibilita associar a estes documentos os termos desejados (conforme descritos na ontologia indicada) e valores para os mesmos. Assim um documento pode conter a informação sobre o seu autor ou assunto (entre outras), de forma a que esta possa ser utilizada automaticamente em um ambiente de recuperação de informações, a partir da ontologia de domínio especificada.

De forma geral, o uso de ontologias para o apoio à recuperação de informações pode ser visualizado na figura 2.10 abaixo, onde está resumida a forma de utilização (inicial) de ontologias empregadas em outro trabalho pioneiro, o Ontobroker[35]. Segundo a descrição feita por Erdman (1998), a utilização de ontologias neste sistema está associada aos seus diversos módulos componentes, ou seja: junto à base de conhecimento, junto ao módulo de acesso aos documentos, associado ao mecanismo de inferência e também ao mecanismo de resolução das perguntas.

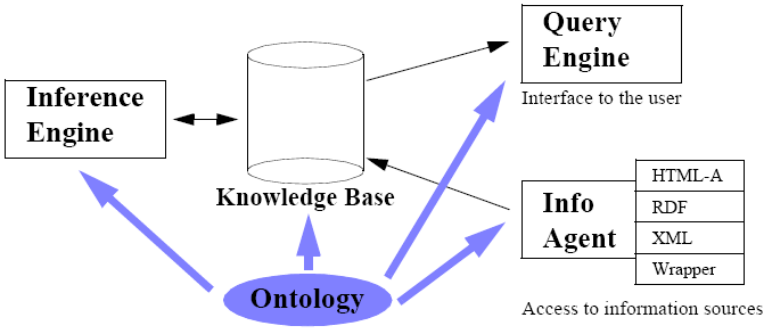


Figura 2.10: Esboço do uso de ontologias no sistema Ontobroker

Os trabalhos relacionados ao sistema Ontobroker atualmente são disponibilizados pelo projeto Ontoprise[36], sendo inclusive observadas ofertas de diversos produtos para empresas e entidades, como sistemas para edição e manutenção de ontologias (Ontoedit), sistema de inferência (Ontobroker), sistemas para gerência de conhecimento ou utilização destas tecnologias em ambientes de trabalho (SemanticMinner e OntoOffice). Este projeto também está relacionado com o desenvolvimento do ambiente KAON[37] (*The KArlsruhe ONtology and Semantic Web Tool Suíte*), um framework de código aberto para o gerenciamento de ontologias,

---

34 disponível em http://www.google.com

35 http://ontobroker.aifb.uni-karlsruhe.de

36 http://www.ontoprise.de/home

37 http://kaon.semanticweb.org/

voltado para o desenvolvimento de aplicações, com foco na escalabilidade e na eficiência de mecanismos de inferência. A partir da descrição feita por Oberle et al. (2003), observa-se a preocupação de implementação de um ambiente capaz de permitir flexibilidade ao gerenciamento de ontologias utilizadas e à criação de aplicações, com a intenção de posicionar o ambiente como um servidor de aplicações voltado para a Web Semântica.

A utilização da linguagem RDF como suporte para a descrição de ontologias e anotação de metadados é observada nestes projetos citados e também em diversos trabalhos, sendo a abordagem inicialmente adotada, juntamente com outras linguagens para descrição de ontologias como DAML (*Darpa Agent Markup Language*)[38], OIL (*Ontology Inference Layer*)[39] e DAML+OIL[40]. No desenvolvimento do RDFSuite, descrito por Alexaki et al. (2001) observa-se uma ferramenta para a manutenção e manipulação, em um SGBD objeto-relacional, de descrições em RDF. Outro modelo para a manipulação de dados descritos em RDF por ser observado no trabalho descrito por Reggiori (2003) onde uma indexação específica é adotada como forma de adaptação às suas características de descrição de recursos, propriedades e valores. No trabalho descrito por Harmelen (2003) uma arquitetura para armazenamento e consulta a dados em RDF e RDF Schema é apresentada, tendo como característica a abstração em relação ao repositório de dados e a possibilidade de uso do RDF Schema. O projeto Jena, desenvolvido pelo laboratório de Web Semântica da Hewlett-Packard também apresenta um conjunto de funcionalidades voltadas inicialmente para a manipulação de dados em RDF, segundo Mcbride (2001). Nestes projetos observa-se a continuidade do trabalho inicial e a utilização de linguagens como DAML+OIL e, em seguida, da OWL, como linguagem para representação de ontologias.

Sistemas de recuperação de informações baseados em RDF são conhecidos, como o QuizRDF (Davies 2004). Também o sistema PISTA emprega inicialmente o RDF, em um contexto de aplicação voltado para questão de informações para segurança, entretanto, como descrito por Bertram et al. (2005), realiza posteriormente uma migração para o uso da linguagem OWL. Alguns trabalhos neste sentido atuam em domínios específicos, o que facilita o tratamento de algumas questões relativas ao mapeamento de informações, como no caso do projeto ITTALKS, voltado para material a respeito de palestras na área de Tecnologia da Informação (Cost, 2002).

Observa-se também o esforço no sentido da integração de informações obtidas em diversas ontologias, o que pode ser bastante desejável. O ambiente ASCS (*Agent Semantic Comunication Service*) utiliza a Linguagem DAML+OIL para implementar estas possibilidades a partir de agentes que realizam a tradução e adequação de termos entre ontologias (LI, 2002). Em outros trabalhos pode ser vista a utilização de traduções para linguagens com disponibilidade de um maior suporte às operações de inferência, como a linguagem Prolog. Exemplos são descritos, como por Wielemaker (2003), no caso de uma implementação de interface para tratamento de RDF e OWL, ou por Quaresma (2003, 2001) onde observa-se a utilização da linguagem Prolog como forma de tratamento mais eficiente de interrogações a serem realizadas, após uma operação de

---

38 http://www.daml.org

39 http://www.ontoknowledge.org/oil

40 http://www.daml.org/committee

conversão entre formatos de linguagens de descrição de ontologias (DAML+OIL e OWL) para fatos Prolog.

Em contexto diverso, empresas como a TopQuadrant[41], disponibilizam atualmente ambientes a partir dos quais é possível implementar serviços de integração de dados de repositórios diversos ou outras aplicações nas quais são necessárias ontologias e sistemas de inferência. Em outro exemplo neste âmbito, temos o serviço disponibilizado pela Intellidimension[42], que consiste em uma plataforma de aplicação baseada em descrição de dados a partir da linguagem RDF e de componentes complementares diversos que possibilitam o acesso a bases de dados e a operações de inferência.

2.11.2 Aplicações em Anotação Semântica

A utilização de ontologias, em geral, pressupõe a aplicação de anotação semântica de informações junto aos documentos que compõe o domínio de pesquisa. A anotação semântica pode ser entendida, neste contexto, como o processo, referido anteriormente, de associação entre documentos e informações sobre os mesmos (metadados). Além da anotação de documentos deve ser considerada como de interesse também a anotação de serviços Web (Medjahed, 2003), ou então a anotação de processos, situações e papéis (Gangemi, 2003). Observam-se diversas formas de realização de anotações, que podem ser referenciadas, de modo geral, como anotação manual dos documentos, observada em alguns sistemas, ou então anotação automática, possível a partir de um contexto diferenciado, onde a origem dos documentos deve ser conhecida ou então algum processo de análise dos mesmos deve ser desenvolvido. Quando a informação associada aos metadados adequados para a descrição de documentos encontra-se disponível em um ambiente integrado, acessível ao usuário durante a criação dos documentos, a anotação manual pode ser uma opção proveitosa, pelo fato de possibilitar uma maior correção, visto que seria realizada pelo próprio autor do documento, contando com o conjunto de termos a partir da ferramenta de criação.

Esta abordagem manual pode ser observada, como descreve Bechhofer (2001), no projeto COSHE (*Conceptual Open Hipermedia Service*) onde anotações a respeito de documentos podem ser realizadas a partir de uma ferramenta utilizada como um *plug-in* associado a um navegador Web. Outro projeto similar e em desenvolvimento junto ao âmbito do W3C, bastante promissor em suas possibilidades é o projeto Annotea[43], que implementa uma infra-estrutura para a anotação de documentos de forma distribuída e de forma colaborativa entre diversos usuários. Em alguns trabalhos podem ser encontradas também preocupações relacionando a anotação ao próprio projeto dos documentos anotados, como proposto por Woukeu (2003), onde modelos mais gerais de projeto para documentos de hipermídia como *Web Modelling Language* (WebML[44]) ou *Object-Oriented Hypermedia Design Model* (OOHDM[45]), entre outros, são analisados em relação ao possível suporte em tarefas de anotação voltadas para o

---

41 http://www.topquadrant.com

42 http://www.intellidimension.com

43 http://www.w3.org/2001/Annotea/

44 http://www.webml.org/webml/page1.do

45 http://www.telemidia.puc-rio.br/oohdm/oohdm.html

seu uso na Web Semântica. Um portal para descrição e acesos a serviços de arquivamento digital, no qual a anotação é realizada pelos autores dos arquivos é descrito por Yeh (2003). No projeto SEAL (*Semantic portAL*), como referenciado por Staab et al. (2002), podem ser vistos mecanismos complementares, com a anotação manual e também com a geração de informações de modo automático. Já no projeto descrito em Berrios (2003) utilizam-se as anotações semânticas beneficiando o usuário do sistema através da facilitação da formulação de perguntas, a partir de uma interface que é construída com base nas anotações disponíveis. Tal abordagem também pode ser observada no projeto OntoWeb (Fensel, 2003), onde a anotação disponível para os documentos é utilizada como campos a serem indicados pelos usuários em um formulário de perguntas. Ou ainda, em domínios bem específicos, como documentos de imagens, é possível também observar a existência de sistemas como o descrito em Schereiber (2001), que permite a anotação de termos para a descrição de fotografias, facilitando a recuperação posterior das imagens desejadas.

Em situações onde os documentos são obtidos a partir de repositórios conhecidos é possível a utilização de formas automáticas para a realização da anotação, visto que um procedimento de geração dos documentos pode usar informações de uma ontologia para a representação de metadados adequados a cada documento sendo criado (Vieira e Rigo, 2002; Chismann et al. 2003). Uma linha de trabalho bastante promissora aponta para o uso de ferramentas automáticas de geração de anotações a partir da análise dos textos dos documentos, ou então a partir da classificação destes, de modo a possibilitar que todo um grande conjunto de documentos já existentes na Internet possa ser acessado com base nos conceitos da Web Semântica. Como exemplo, temos Engels (2001) descrevendo o ambiente "Corporum", no qual um módulo de análise semântica dos documentos possibilita a geração de informações sobre os conceitos descritos nestes documentos, implícita ou explicitamente, sendo esta informação utilizada posteriormente como anotação. Abordagens para a construção de ontologias a partir de documentos também são conhecidas, como em Saias (2003) onde um conjunto de documentos a ser utilizado em sistema de recuperação de informações é analisado levando-se em conta informações lingüísticas e a partir desta análise são gerados conceitos de forma automática, para uso na descrição dos documentos, de forma relacionada a uma ontologia de domínio. No trabalho descrito por Popov (2003) observa-se um sistema que permite a anotação semântica de termos e a utilização destas anotações para indexação automática de documentos, com vistas à recuperação de informação baseada na associação entre as perguntas formuladas pelo usuário e no conjunto destas informações semânticas.

No contexto da educação à distância, a personalização pode ser vista como uma importante aliada ao processo de aprendizado realizado pelo aluno, uma vez que os conteúdos podem ser apresentados ao mesmo de acordo com suas preferências e características. Podem ser levadas em conta as suas experiências anteriores, verificadas no ambiente ou relatadas, os seus objetivos de aprendizado, suas características cognitivas e mesmo detalhes do equipamento utilizado pelo aluno, como tipo e resolução do monitor ou velocidade de conexão disponível. Juntamente com estas informações são processadas as informações disponíveis sobre os recursos didáticos no ambiente, sendo possível a indicação de conteúdos adequados ao aluno, a partir de diversos fatores e dentro de um contexto conhecido. Observa-se, em aplicações voltadas ao ensino à distância, a utilização de vocabulários controlados como forma de possibilitar a resolução de problemas de acesso ao material didático disponibilizado.

Alguns padrões de descrição de metadados como o SCORM (*Sharable Content Object Reference Model*) (SCORM 2003) ou o LOM (*Learning Objects Metadata*) (IEEE 2002) se preocupam em fornecer um conjunto de elementos padronizados, representando os conceitos observados no conjunto de documentos, para que tarefas como reaproveitamento, localização e relacionamento de material possam ser realizadas de forma eficiente. Diversas organizações com foco em educação à distância participam de seu desenvolvimento, como no caso do LOM, onde se observa a participação da ARIADNE [46] (*Alliance of Remote Instrucional and Distributed Networks on Europe*), o IMS *Global Learning Consortium*[47] e o grupo associado ao *Dublin Core Metadata Initiative* (DCMI[48]).

Estes padrões de anotação podem ser associados aos materiais didáticos na sua criação, quando o autor pode indicar, para cada documento gerado, o conjunto de termos descritores adequado. Com este processo o conjunto dos documentos poderá ser posteriormente consultado com a utilização destas informações. Em diversos trabalhos observa-se a preocupação de utilização destes padrões no contexto da Web Semântica, a partir de suas linguagens. Em Aroyo et al. (2003) é descrita a utilização do SCORM com apoio de OWL ou descrições DAML-S[49] no ambiente OntoAims. A criação do material didático é acompanhada pela anotação semântica, segundo o padrão empregado, o que possibilita a recuperação posterior dos conteúdos, permitindo que sejam acompanhadas seqüências de material ou determinadas escolhas a partir de condições indicadas. Não é observada neste sistema a personalização automática do conteúdo apresentado, pois não é descrito um modelo consistente do usuário. Em outro trabalho, apresentado por Nilsson (2003) o padrão LOM e a linguagem RDF são explorados para a descrição de um modelo de associação e de resolução de questões formuladas segundo um vocabulário controlado. O ambiente foi testado no projeto SHAME[50].

2.11.3 Aplicações em Personalização e Adaptação

Alguns trabalhos suportam a personalização em educação à distância oferecendo mecanismos de adaptação de conteúdo que levam em conta informações diversas sobre o usuário, como seus objetivos, preferências e conhecimento adquirido, além de considerações mais técnicas acerca de suas possibilidades de conexão, sistema operacional ou software para acesso a Internet (navegador Web). Esta abordagem pode ser vista em trabalhos como o apresentado por Brusse (2003) ou Stojanovic et al. (2001), ou ainda comentada por Palmer (2002). Em Sancho (2002) o uso de padrões associados à anotação semântica é também empregado e permite que o usuário tenha o conteúdo adaptado a partir de diversos fatores, entre eles o nível de dificuldade descrito para o material.

Aggarwal e Yu (2002) descrevem um sistema completo para a personalização em um contexto de portal de notícias. Neste caso são empregadas técnicas de análise

---

46 http://www.adlnet.org

47 http://www.imsproject.org

48 http://dublincore.org/

49 http://www.daml.org/services/ISWC2002-DAMLS.pdf

50 http://kmr.nada.kth.se/shame

léxica dos documentos compondo o portal, de onde são obtidas cadeias de termos correlacionados que são utilizadas em conjunto com informações de uso por cada usuário. Com isso é possível a obtenção de um mapeamento de conjuntos de documentos que supostamente seriam de interesse para cada usuário, permitindo tarefas como a recomendação de notícias e a realização de operações de busca personalizadas.

Liu et al. (2002) apresenta uma abordagem para aplicação de recursos de personalização a um contexto de um sistema de recuperação de informações. Para isso é descrito um perfil de usuário que é acrescido de informações na medida em que este interage com o sistema. Ao mesmo tempo são utilizados mapeamentos das perguntas realizadas para categorias de intenções (possíveis) dos usuários, sendo que com este mecanismo o sistema de recuperação de informações reage de modo personalizado, levando em conta tanto o perfil aprendido do usuário como a possível intenção em uma pergunta, que pode ser algo não relacionado com o seu comportamento normal.

Kim et al. (2003) apresenta uma abordagem para a personalização de operações de recuperação de informações a partir da manutenção de um modelo de preferências do usuário e do uso de um esquema de representação denominado "*Weigthed Semantic Taxonomy Tree*", utilizado no sistema para a representação de termos e de conceitos relacionados com estes, de modo a auxiliar na localização do conteúdo desejado pelo usuário. Um mecanismo de aprendizado baseado em uma rede neural é utilizado, juntamente com informações (explícitas ou implícitas) sobre a satisfação do usuário, para que o modelo de intenções deste seja aprimorado.

Em função de dificuldades no tratamento de grandes quantidades de texto existente na Internet atualmente, observa-se a utilização de diversas metodologias para o apoio aos processos de recuperação de informações e outros relacionados, como personalização. Técnicas de processamento de linguagem natural (e também a mineração de dados na web) podem ser observadas, voltadas para o apoio aos sistemas de recuperação de informações. O uso de ontologias pode ser encontrado, de modo geral, em diferentes situações, tanto auxiliando nos processos de recuperação de informações como também em processos de personalização. O uso de informações lingüísticas apresenta-se como um caminho promissor no sentido de auxiliar no processo de geração de ontologias e anotações semânticas a partir de conjuntos de documentos existentes na Internet ou associados a domínios específicos (Mobasher e Dai 2004).

A aquisição de ontologias de domínio, tarefa importante neste contexto, pode ser observada em diversos trabalhos, com diferentes técnicas. Também existem diversas possibilidades de apoio a esta tarefa, associadas à extração de informações e descoberta de conhecimento. Em Loh et al. (2000), é utilizada uma abordagem semi-automática onde ferramentas de processamento de linguagem natural são empregadas para a obtenção de conceitos associados a domínios. Maedche e Staab (2000) apresentam uma abordagem para a obtenção de relações conceituais generalizadas a partir de regras de associação à mineração. A criação de uma ontologia base a partir da análise de documentos e da extração de seus conceitos mais importantes é descrita por Saias e Quaresma (2003). Após a análise sintática do texto e de uma análise semântica parcial são extraídas as entidades observadas no texto. Em uma segunda etapa, com intervenção manual de especialistas, estas entidades são associadas a uma ontologia de domínio com maiores relacionamentos.

A possibilidade de personalização de conteúdos e apresentação, em conjunto com o tratamento semântico dos documentos permite ainda que sejam vislumbradas possibilidades de tratamento das operações de navegação de uma forma conceitual, onde o usuário não teria a necessidade de interação a partir de palavras e *hyperlinks*, mas sim a partir dos conceitos desejados. Caberia ao sistema gerenciar esta interação de modo a localizar para os usuários os documentos adequados. Alguns exemplos neste sentido podem ser observados. Baldoni et al. (2003) utiliza um agente cognitivo que gera dinamicamente uma visão do *site* Web onde o usuário navega, a partir do conhecimento do usuário armazenado no sistema e do conhecimento dos conteúdos dispostos nas diversas páginas. A navegação é definida como conceitual e não baseada nas ligações entre documentos (*hyperlinks*). Nos trabalhos descritos por Crampes (2000), Martelli (2002) e Naeve (2003) esta abordagem pode ser também verificada, sendo que a principal vantagem apontada é a possibilidade de interação apenas através conceitos e não através de *hyperlinks*.

O estudo e a aplicação da lingüística possibilita que sejam obtidos resultados promissores em tarefas diversas, associadas à recuperação de informações, personalização, integração de informações ou descoberta de conhecimento. No trabalho de Ciorascu et al. (2003) observa-se um sistema de recuperação de informações projetado para a utilização da Wordnet[51], descrita em linguagem OWL, o que possibilita a ampliação da etapa de localização dos termos desejados, tendo em vista a possibilidade de relacionamento dos termos com informações lingüísticas associadas, como por exemplo sinonímia e hiponímia. O mapeamento de termos descritos em ontologias diversas também pode ser beneficiado pela abordagem lingüística, como descrito em Magnini (2002). Um sistema de extração de informações a partir de textos que utiliza ontologias onde são descritas algumas estruturas sintáticas parciais é apresentado por Todirascu et al. (2002), com o propósito de recuperar novos termos nos documentos analisados. Em outro trabalho (Andreasen et al., 2002) a recuperação de informações é realizada a partir do mapeamento e comparação dos conceitos contidos no texto. Variantes morfológicas e sinônimos lexicais são reconhecidos e utilizados em conjunto com as informações já descritas em ontologias.

Resumo do Capítulo:

Neste capítulo são apresentados resumidamente os conceitos da Web Semântica que são importantes para o trabalho desenvolvido. Foram apresentados os fundamentos das linguagens XML, XSL e RDF, bem como da linguagem OWL. As linguagens de consulta RDQL e SPARQL são descritas brevemente. São descritos alguns exemplos de utilização destes recursos em situações de recuperação de informação, personalização e anotação semântica.

---

[51] http://www.cogsci.princeton.edu/~wn/

3 MINERAÇÃO DE DADOS NA WEB

Neste capítulo são contextualizadas as áreas de Mineração de Dados e, mais especificamente, a Mineração de Dados na Web, em relação ao processo mais geral de Descoberta de Conhecimento em Bases de Dados. A seguir são detalhados aspectos importantes da Mineração do Uso da Web, com a apresentação de características relevantes para o trabalho aqui descrito.

3.1 Contextualização

A Mineração de Dados é uma das etapas do processo de Descoberta de Conhecimento em Banco de Dados e possibilita a exploração de repositórios de dados para a descoberta de novos conhecimentos (Fayad, 1996). Ela permite inferir informação útil e descobrir relacionamentos implícitos, tais como padrões de comportamento, conceitos, ou regras. Este processo constitui-se de várias etapas que são executadas tipicamente com a interação de responsáveis pela geração e análise dos dados. A necessidade de ajustes em parâmetros, conjuntos de dados e técnicas de mineração fazem com que sejam normalmente necessárias estas interações durante o processo (Zaiane, 2000).

As principais etapas deste processo podem ser agrupadas, segundo Addrians (1997), em ações sobre as bases de dados, ações de mineração e ações de representação e uso do conhecimento. As ações iniciais, relacionadas com as bases de dados, são as mais trabalhosas, pois envolvem a tarefa de conhecimento dos dados do domínio, seleção dos dados relevantes para o processo, tratamento dos dados selecionados com operações de limpeza para retirada de erros e redundâncias ou ambigüidades, sendo por fim aplicadas transformações necessárias para que estes dados possam ser utilizados na etapa posterior, de mineração. Estas transformações podem ser bastante diversas, pois dependem da natureza dos dados e do formato necessário para as técnicas empregadas na mineração. A partir da utilização de mecanismos para a mineração deste conjunto de dados transformado, serão descobertos padrões, que devem passar por um processo de análise e validação, para que seja verificada a possibilidade de sua utilização. Em freqüentes casos esta análise permite detectar a necessidade de retorno às etapas anteriores para que os dados originais sejam tratados de forma diferente ou então para que sejam organizados novos conjuntos de parâmetros para a etapa de mineração. Após este percurso, o conhecimento descoberto e validado será então representado ou utilizado pelos responsáveis, sendo que esta etapa é muito específica e associada às particularidades tanto do domínio de conhecimento como das técnicas disponíveis e dos objetivos finais.

Segundo diversos autores (Fayad, 1996; Zaiane, 2000; Addrians, 1997), estas etapas comentadas podem ser detalhadas em etapas mais específicas, relacionadas como

ilustra a figura 3.1 a seguir. Ressalta-se a natureza interativa do processo, sendo necessário o envolvimento dos responsáveis em diversas etapas.

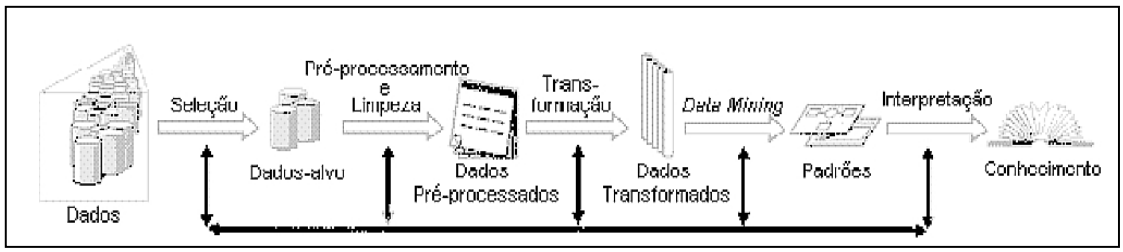


Figura 3.1: Visão geral do processo de descoberta de conhecimento em bases de dados

Na etapa de mineração de dados, podem ser empregados diversos tipos de algoritmos, sendo que cada possui maior ou menor adequação para com características específicas de problemas de descoberta de conhecimento. Alguns exemplos são os algoritmos de classificação e segmentação (*clustering*), algoritmos para mineração de regras de associação e de padrões. Estas classes de algoritmos determinam também tarefas nas etapas anteriores, de seleção e pré-processamento dos dados. Com o objetivo de proporcionar uma visão geral, são descritas a seguir algumas destas classes.

Os algoritmos de classificação partem de um conjunto conhecido de informações que delimitam classes de objetos. Durante sua aplicação, estes algoritmos analisam as características de cada objeto e o relacionam com as classes pré-definidas. O procedimento de construção das classes é efetuado com a utilização de um conjunto de exemplos previamente classificados, conhecido como conjunto de dados para treinamento. Esta abordagem é conhecida como aprendizagem indutiva, pois o conhecimento prévio do domínio e a identificação de similaridades são utilizados para a construção do modelo e a classificação. Uma abordagem diferenciada é encontrada nos algoritmos de segmentação, que partem do conjunto de dados disponível e buscam identificar subconjuntos com características similares (também conhecidos como agrupamentos ou *clusters*). Não existem classes definidas inicialmente, sendo este um diferencial em relação aos algoritmos de classificação.

As regras de associação constituem uma terceira abordagem, onde são identificadas relações entre registros de bases de dados, observando-se ocorrências significativas de valores. Elas permitem identificar, por exemplo, porcentagens significativas de ocorrência relacionadas destes valores. Uma típica regra de associação permite indicar que, dada a ocorrência de valores “A” e “B” em um registro, existe uma porcentagem “P” de ocorrência do valor “C” nestes mesmos registros. Esta porcentagem é tratada como fator de confiança da regra. Este pode estar associado com outro indicador, que seria o fator de cobertura da regra, o qual indica o percentual de ocorrência desta relação no conjunto dos dados tratados. O que pode ser considerada uma variação desta abordagem é a análise de seqüências, particularmente útil em situações ligadas ao contexto da Web. Seu objetivo é analisar o conjunto de dados e identificar seqüências específicas que ocorram de modo freqüente ou relacionadas pelo fator temporal. Estas seqüências podem representar itens adquiridos ou, no caso da Web, acessos a determinadas páginas de um *site*. Ainda podem ser encontrados algoritmos específicos para tarefas como sumarização ou regressão. Também observa-

se mais recentemente o uso de recursos de lógica nebulosa (*Fuzzy Logic*) e de algoritmos genéticos para tarefas específicas.

A mineração de padrões de comportamento na Web origina-se em trabalhos anteriores de mineração de dados e tem por objetivo a descoberta automática ou semi-automática de padrões de acesso gerados por usuários de *sites* Web. Algumas das aplicações que impulsionam este tipo de abordagem específica são os sistemas de recomendação ou sistemas voltados à personalização, como pode ser observado na seguinte definição, apresentada por Mobasher (2005):

> *"Para um processo de personalização mais efetivo, tanto as informações de uso como de conteúdo de um site devem ser integradas com mineração Web e usadas nos sistemas de geração de personalização".*

Como existem características específicas a serem observadas e tratadas no caso de mineração da Web, esta área de utilização dos recursos de mineração de dados é descrita a seguir em maiores detalhes.

3.2 Mineração da Web

Pode-se definir mineração da Web, de forma ampla, como sendo a descoberta e análise de informação útil da Web, onde, a partir das informações descobertas, seja possível demonstrar características, comportamentos, tendências e padrões de navegação do usuário e de conteúdo Web (Cook, 2000).

A Mineração na Web pode ser dividida em três áreas de interesse, segundo Zaiane (2000) e Mobasher (2005). Seriam elas a Mineração do conteúdo da Web (*Web Content Mining*), Mineração da estrutura da Web (*Web Structure Mining*), e Mineração do uso da Web (*Web Usage Mining*). A estas áreas podem ser associadas diferentes coleções de dados, que colaboram inclusive para o estabelecimento desta distinção. Desta forma podem ser utilizados dados originados nos servidores Web (registros de acesso) ou em servidores *Proxy*, dados originados no software cliente utilizado para a navegação, ou ainda dados disponíveis em bases de dados e relacionados com os conteúdos apresentados ou com operações realizadas.

Resumidamente, a Mineração do Conteúdo na Web é o processo de extração de informações úteis sobre o conteúdo, dados e documentos da Web. A Mineração da Estrutura na Web é o processo de inferência de conhecimento através da topologia, organização e estrutura de links da Web entre referências de páginas. Finalmente, a Mineração do Uso da Web é o processo de extração de padrões de navegação interessantes dos registros de acesso Web.

A tabela 3.1 abaixo apresenta uma sumarização de diversos aspectos destas três abordagens, relacionando os tipos de dados usuais, suas origens, a representação normalmente observada e categorias de aplicações possíveis. Entretanto estas divisões não são absolutas e existem abordagens possíveis integrando aspectos diversos, como forma de ampliar as possibilidades de aquisição de informações.

Tabela 3.1: Áreas de interesse na Mineração da Web

	Mineração de uso	Mineração de conteúdo	Mineração de estrutura
Tipo de dados	Resultados da interação, como visualizações de páginas	Textos estruturados ou semi-estruturados, conteúdo em bases de dados	Estrutura de hyperlinks
Fonte dos dados	Registros de uso do Servidor Web ou registros obtidos por aplicações de gerenciamento de conteúdo. Alguns dados são privativos.	Documentos de texto, hyperdocumentos, bases de dados de gerenciadores de conteúdo Web. Dados públicos.	Estrutura de hyperlinks publicados
Representa-ção usual	Tabelas relacionais, grafos	Lista de palavras, termos ou frases, conceitos de ontologias, tabelas em bases de dados relacionais	Grafos
Categorias de aplicação	Adaptação e personalização de *sites* Web, modelagem de usuários	Categorização e segmentação, extração de regras e padrões	Categorização e segmentação

A seguir são descritas em maiores detalhes as etapas necessárias para a tarefa de Mineração do Uso da Web, que foi a técnica adotada neste trabalho. Um dos motivos desta adoção é a possibilidade de aplicação dos resultados de Mineração de Uso da Web em contextos de adaptação de *sites*, objetivo do mesmo.

3.3 Mineração do Uso da Web

Mineração do Uso da Web é a atividade de mineração de dados que visa à descoberta automática de padrões de comportamento de usuários, através de seus dados de acesso a Web. Desta forma, utilizando-se os dados de sua interação atual, pode-se obter uma previsão de interações futuras. Esta técnica vem ampliando sua utilização, pois um número crescente de organizações utiliza cada vez mais a Internet para a divulgação e administração de seus negócios, logo fazendo com que as estratégias de comunicação e técnicas para análise de mercado, em decorrência deste contexto, passem a ser realizadas através deste meio de comunicação (Baeza-Yates, 2006; Mobasher, 2005).

A partir desta perspectiva são identificadas diversas áreas de aplicação para a Mineração do Uso da Web, dentre as quais podem ser citadas, como exemplos, aplicações de reconhecimento de perfis de usuários, personalização e recomendação, melhorias e reestruturação do projeto de *sites* Web, avaliação de *sites* em ambientes de educação a distância, inteligência de negócios e comércio eletrônico, bem como auxílio em operações de sistemas de Recuperação de Informações.

Diversas fontes de dados e os principais algoritmos e técnicas encontradas na Mineração de Dados (tais como agrupamento, geração de regras de associação e descoberta de padrões seqüenciais) podem ser utilizados para extração, descoberta e análise de conhecimento agregando assim valor informacional a estes dados de uso.

As fontes de dados empregadas neste caso, como já indicado, são bastante variadas. Podem ser usados dados gerados e manipulados pelos navegadores Web,

dados obtidos pelos servidores Web em seus registros de acompanhamento de acesso e também podem ser usados dados gerados pelas aplicações Web, a partir de inserções específicas de código em etapas de atendimento de requisições para páginas Web. Os dados mais comuns encontrados em padrões de uso de páginas Web, tais como o endereço IP, páginas referenciadas, data e tempo de acesso às páginas, são os mais utilizados. Tipicamente, os dados de uso dos registros de acesso dos servidores Web são mantidos em arquivos com formato CLF (*Common Log Format*[52]). Dados gerados por aplicações web utilizam diversos formatos, normalmente com uso de padrões XML, bem como bases de dados auxiliares. A análise feita nestes dados pode ajudar a detecção de uma série de padrões, tais como padrões demográficos, tempo médio de acesso, percursos freqüentes, resultados de campanhas de comunicação e estratégias de marketing direcionadas a determinados produtos.

Estes padrões podem ser detectados com diferentes técnicas. O tratamento de percursos freqüentes utiliza a identificação de sessões de usuários e realiza a organização destes dados resumindo os percursos mais utilizados, seja por um usuário ou por um grupo de usuários. Já as regras de associação, no contexto da mineração de uso da Web, servem, por exemplo, para relacionar páginas que foram freqüentemente referenciadas em conjunto, ou associar perfil de usuários aos acessos em determinadas páginas, identificando, por exemplo, probabilidades de que, ao visitar uma página "A", o usuário também visite uma página "B". A segmentação (ou *clustering*) permite a identificação de grupos de dados associados por características similares de acesso.

3.3.1 Fases da Mineração do uso na Web

O processo de mineração de uso da Web possui uma característica geral similar ao processo de mineração de dados. A descrição do processo apresentada em Fayad (1996), identifica etapas como seleção de dados e pré-processamento, transformação dos dados para o formato adequado ao processamento, mineração de dados com o uso das técnicas escolhidas e por fim a análise dos resultados. Podem ser destacadas, de modo geral, as etapas descritas abaixo, conforme Mobasher (2005, 2002). A figura 3.2 ilustra este processo. Inicialmente deve haver a identificação e aquisição dos dados de uso necessários e definição de sua representação. Estes dados podem ser complementados com informações adicionais, como metadados ou conhecimento do domínio. Juntamente com os dados de uso, as informações complementares podem ser utilizadas em tarefas como o pré-processamento e identificação de acessos e sessões de usuários. Com o uso de algoritmos definidos, é descoberto um conjunto inicial de padrões de navegação, que devem estar em acordo com os requisitos identificados para o processo. Este conjunto inicial deve ser avaliado, de forma a descartar padrões não desejados ou incorretos. Por fim, existe a necessidade de tratamento dos padrões de navegação considerados válidos de forma que possam ser utilizados para avaliação ou adaptação de *sites* Web, entre outras aplicações possíveis.

---

52 http://www.w3.org/pub/WWW/Daemon/User/Config/Logging.html#common_logfile_format

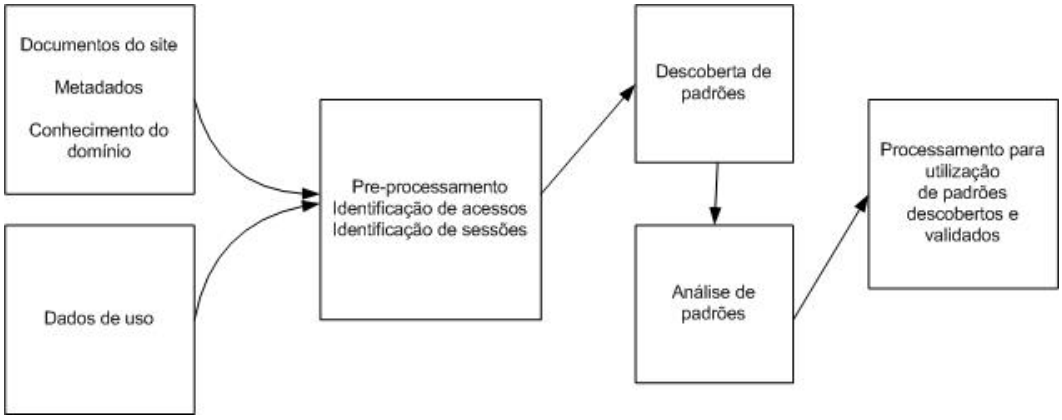


Figura 3.2: Resumo das fases do processo de Mineração de Uso da Web

Dadas as características específicas da Mineração de Uso da Web, algumas etapas são realizadas de forma diferenciada das demais formas de Mineração de Dados. A seguir são detalhadas algumas destas características.

3.3.1.1 Pré-processamento

O pré-processamento é a etapa onde, escolhidas as fontes de dados, são realizadas diversas operações que permitem a seleção dos dados, a limpeza dos mesmos quanto aos possíveis erros, a transformação destes para formatos adequados ao processamento posterior e sua organização em coleções diferenciadas por critérios como tempo de ocorrência, servidor de origem, ou outras possíveis.

Esta é uma tarefa difícil de ser executada devido à incompletude e a grande heterogeneidade dos dados disponíveis. Conforme citado, diversas fontes de dados podem ser utilizadas para esta tarefa, das quais a principal é o arquivo de registros de acesso mantido pelo servidor Web, que registra cada requisição a um recurso. É importante ressaltar que um registro é automaticamente adicionado a cada vez que um recurso é solicitado ao servidor Web. Como conseqüência, as entradas de todos os usuários estão agrupadas pela seqüência de ocorrência, sendo que uma única página requisitada por um usuário pode gerar múltiplas entradas no arquivo de registros do servidor. Isso ocorre porque os diversos elementos que compõe a página (imagens, arquivos HTML complementares, arquivos de estilo ou scripts auxiliares) geram, cada um, uma requisição. Além disso, alguns destes acessos são tratados com mecanismos de servidores *Proxy* ou de memória temporária de acessos (*cache*), modificando em alguns aspectos a situação existente. Assim, as tarefas principais desta etapa são a identificação de diferentes usuários, a eliminação de registros de acessos irrelevantes (figuras e estilos, por exemplo) e a reconstrução das sessões de usuários.

Em casos onde a fonte de dados é alimentada pela própria aplicação Web, existem algumas facilidades, como a identificação mais precisa do usuário e suas sessões. No caso de uso dos registros de servidor Web, existe sempre alguma margem para erros, pois a definição do início e final de uma sessão de usuário está sujeita a heurísticas e informações incompletas.

Os arquivos com o registro dos acessos, mantidos pelo Servidor Web, foram criados originalmente para a realização de operações estatísticas e para auxiliar em tarefas de depuração (Kohavi, 2001). Um dos formatos largamente utilizados para o armazenamento destes registros é o "Common Log Format" (Nielsen, 1995), mas existem alguns formatos proprietários e uma extensão deste, que adiciona algumas

informações. Estes formatos descrevem arquivos de texto nos quais são armazenadas as informações de cada requisição Web. Os campos principais são a identificação do endereço da requisição, data e horário do acesso, os parâmetros recebidos, o estado final da requisição, o número de dados transferidos e a identificação do navegador utilizado. Quando alguma informação não está disponível, utiliza-se um traço ("-") em seu lugar. Algumas linhas típicas destes arquivos de registro de acesso podem ser vistas abaixo, na figura 3.3, onde estão representados o acesso a um documento em formato HTML ("/cursos/intercambios/apresentacao/corpo.htm") e a um documento de imagem ("/_imagens/capa/banners/ban_extravest.jpg").

```
66.249.64.47 - - [13/Feb/2005:04:15:13 -0200] "GET
/cursos/intercambios/apresentacao/corpo.htm HTTP/1.0" 304 - "-" "Googlebot/2.1
(+http://www.google.com/bot.html)"
10.21.213.93 - - [13/Feb/2005:04:15:20 -0200] "GET
/_imagens/capa/banners/ban_extravest.jpg HTTP/1.1" 200 2968 "https://www1.unisinos.br/"
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.2; .NET CLR 1.1.4322)"
```

Figura 3.3: Exemplo de um trecho de arquivo no formato "Extended Common Log"

Alguns problemas para a Mineração de Uso da Web podem ser indicados. A anotação textual e de forma extensiva de todos os dados tende a gerar documentos bastante grandes e de difícil processamento. No caso de um acesso a um documento HTML as anotações incluem todos os documentos associados ao arquivo original solicitado. Neste caso tipicamente observa-se a referência a outros documentos de imagens, de descrição de estilo (como arquivos no formato CSS, por exemplo) ou contendo rotinas em alguma linguagem como Javascript. Para a tarefa de Mineração de Uso da Web, nem todos estes dados são relevantes. Além desta situação, as anotações são realizadas no momento de sua recepção pelo servidor Web, o que intercala requisições de diversos usuários e gera a necessidade de pré-processamento para a identificação de sessões individuais. Isso normalmente é feito levando-se em conta heurísticas específicas que se baseiam em dados da origem da requisição, data e horário. Porém estas abordagem estão sujeitas à ocorrência de erros, principalmente relacionados com aplicação de servidores *proxy* ou de memória temporária de acessos (*cache*). Nestes casos alguns acessos que já foram realizados anteriormente pelo navegador Web não geram novas requisições ao Servidor Web e são atendidos localmente (pelo servidor *proxy* ou pelo mecanismo de memória auxiliar). Portanto existirão acessos efetivos que não estarão registrados no arquivo de registros do Servidor Web. Também merecem menção as situações de acessos gerados por sistemas de recuperação de informação, que devem ser retiradas do conjunto a ser analisado.

Dependendo da finalidade da análise a ser realizada, nesta etapa os dados são organizados em diferentes níveis de abstração. O nível mais baixo de abstração é a visualização de uma página (ou "*pageview*"). Neste nível são agrupados diversos objetos que compõe uma página Web, tais como o documento HTML, imagens, rotinas e estilos. Dependendo da aplicação Web, cada visualização pode estar associada com uma tarefa ou evento específico, tal como a leitura de um texto, a escolha de um produto, o acesso à um formulário de dados ou a escolha de um vídeo. Entretanto neste nível de abstração não existe a compreensão de etapas anteriores e posteriores. Para isso utiliza-se um nível de abstração mais alevado, conhecido como a sessão de acesso do usuário. Uma sessão compreende todas as visualizações de páginas Web de um único

usuário em determinado período de tempo que constitui uma visita deste a um *site* Web. Normalmente uma sessão pode variar bastante em sua duração, o que pode ser associado ao tipo de tarefa sendo realizada. Além disso, uma sessão pode ser considerada, para fins de análise, como composta de diversos conjuntos de subsessões, que identificam tarefas diferenciadas realizadas pelo usuário em sua visita.

Um exemplo de identificação de sessões de usuários e da subdivisão destas sessões por um critério de tempo de acesso e navegador utilizado pode ser observado na figura 3.4 abaixo. O exemplo demonstra uma situação onde a identificação da origem do acesso permite claramente a descoberta das sessões de usuários. Também seria possível a utilização deste exemplo para ilustrar a separação de uma sessão de usuário em conjuntos menores, utilizando-se um limite de tempo entre os acessos. Pode ser observada na lista de acessos do terceiro usuário uma diferença significativa entre o tempo do segundo e do terceiro acesso, o que permitiria a conclusão de dois momentos distintos. Entretanto esta conclusão pode ser confrontada com situações em que a diferença de tempo deve-se realmente a um tempo maior gasto para leitura de material e não a uma modificação de objetivos.

Hora	IP	URL	Origem	Navegador		Hora	IP	URL	Origem
15:01	1.2.3.1	A	-	Mozilla/4.0	Usuário 1	15:01	1.2.3.1	A	-
15:09	1.2.3.1	B	A	Mozilla/4.0		15:09	1.2.3.1	B	A
15:10	2.2.2.2	C	-	Mozilla/4.0		15:19	1.2.3.1	F	B
15:12	2.2.2.2	A	C	Mozilla/4.0		15:25	1.2.3.1	G	F
15:15	2.2.2.2	B	A	Mozilla/4.0					
15:19	1.2.3.1	F	B	Mozilla/4.0	Usuário 2	15:10	2.2.2.2	C	-
15:22	3.3.2.3	A	-	IE5		15:12	2.2.2.2	A	C
15:22	3.3.2.3	F	A	IE5		15:15	2.2.2.2	B	A
15:25	1.2.3.1	G	F	Mozilla/4.0		15:33	2.2.2.2	H	B
15:33	2.2.2.2	H	B	Mozilla/4.0					
17:58	3.3.2.3	G	F	IE5	Usuário 3	15:22	3.3.2.3	A	-
17:59	3.3.2.3	H	G	IE5		15:22	3.3.2.3	F	A
						17:58	3.3.2.3	G	F
						17:59	3.3.2.3	H	G

Figura 3.4: Exemplo de identificação de sessões a partir de dados do servidor Web

As tarefas de pré-processamento mencionadas podem ser realizadas a partir de fontes diversas. Uma destas é o caso de utilização de *cookies*[53] e codificação auxiliar, junto a uma aplicação de gerenciamento de conteúdo Web ou de publicação de conteúdo Web. A cada acesso pode ser realizada uma operação de armazenamento dos dados pertinentes ao tipo de análise planejada.

---

53 Um *cookie* é um grupo de dados trocados entre o navegador e o servidor de páginas, colocado num arquivo de texto criado no computador do usuário. A sua função principal é a de manter a persistência de sessões HTTP. Alguns *sites* Web utilizam *cookies*, por exemplo, para guardar as preferências ou histórico de ações do usuário.

3.3.1.2 Descoberta de padrões

A descoberta de padrões sobre os acessos a páginas Web, sejam eles tratados a partir de dados resultantes da análise dos registros de acessos do servidor Web ou de aplicações específicas, corresponde à extração e reconhecimento de características, regularidades e regras. Para um melhor desempenho na aplicação dos métodos e algoritmos utilizados nas etapas de mineração de dados, se faz necessário adaptar alguns dos algoritmos que serão utilizados sobre os arquivos de dados da Web.

Dentre as técnicas disponíveis, observa-se a utilização de regras de associação e de padrões seqüenciais em um grande número de trabalhos (Woon, 2005). Apesar disso, também é possível a identificação de técnicas como agrupamento e técnicas de classificação em determinados contextos.

Para tal, são utilizados os dados obtidos da etapa de pré-processamento. Estes consistem em conjuntos de visualizações de páginas e conjuntos de sessões de usuários. Considera-se um conjunto $P=\{p_1, p_2, p_3, ..., p_n\}$ como sendo o conjunto de *n* diferentes páginas de um determinado *site* Web. O conjunto não vazio $L=\{l_1, l_2, l_3, ..., l_m\}$ representa uma sessão de acesso de um usuário, sendo cada l_i pertencente a P. Cada sessão de acesso pode ser também considerada como um conjunto de pares ordenados $(l_i, v(l_i))$, no qual l_i representa uma visualização específica e $v(l_i)$ representa o valor considerado para esta visualização. Para a definição do valor de cada visualização podem ser levados em consideração diversos fatores, normalmente associados ao tipo de análise desejada. Em alguns casos a definição pode realizada com valores binários, indicando a presença ou não da visualização. Em outros o tempo de duração da visualização pode ser aplicado, apesar desta opção acarretar algumas dificuldades de implementação relacionadas com a dificuldade de identificação precisa deste tempo de permanência em alguns casos. Por fim, outros exemplos possíveis para a identificação do valor de cada visualização são escalas de avaliação de cada página geradas pelo retorno dos usuários ou por algum mecanismo automático de avaliação de conteúdo.

Dado um conjunto de sessões de acesso e um determinado critério para a avaliação da importância de cada acesso, por exemplo, o tempo de permanência, é possível a montagem de uma matriz de acessos. Esta pode ser utilizada de maneiras diversas na Mineração de Uso. Na figura 3.5 abaixo é disponibilizada uma matriz com valores hipotéticos relacionando sessões de acesso com valores para a visualização de páginas Web. Neste caso cada visualização está relacionada com o tempo de permanência obtido. Porém estes valores podem ser associados com diversos outras características, como já comentado. No caso de integração de características textuais ou de informações semânticas a respeito de cada página Web, por exemplo, pode-se chegar a uma nova composição que permite análises mais abrangentes.

Em uma situação onde diversos fatores de análise estejam associados a cada página a matriz exemplificada na figura 3.5 passaria por uma transformação para incorporar múltiplas dimensões, cada uma associada a uma característica utilizada. Cada visualização passa a ser representada então por um vetor contendo elementos no formato $cv(c_i)$, que representa o valor da característica *i* para cada visualização. Dentre as características possíveis estão conceitos ou termos de interesse, tipo de tarefa representada pela visualização, escala de avaliação gerada pelo usuário, entre outras possíveis.

	Visualizações de páginas Web					
	A	B	C	D	E	F
Sessão 1	12	0	33	0	45	20
Sessão 2	0	132	44	2	6	89
Sessão 3	13	44	187	0	0	0
Sessão 4	44	12	76	4	0	129
Sessão 5	23	55	10	0	0	0
Sessão 6	0	123	0	56	0	142

Figura 3.5: Exemplos de matriz de sessões de acesso e valores de importância por acesso

As diversas características ligadas a conceitos ou termos contidos nas páginas Web podem também ser utilizadas de forma a gerar matrizes contendo a transformação de cada um dos acessos de uma visualização para uma lista de acessos na qual serão identificados todos os termos ou conceitos de cada página visitada. Uma utilização possível com este resultado é a análise de conteúdos associados com a navegação realizada pelos usuários. Em outra situação interessante, a larga utilização de dispositivos móveis permite observar a possibilidade da utilização de informações ligadas ao contexto, tais como informações relacionadas com as características específicas do dispositivo em uso (como informações sobre o local de acesso).

No contexto de Mineração de Uso da Web, mecanismos de segmentação (*clustering*) utilizam os dados pré-processados para a identificação de conjuntos de usuários ou então conjuntos de páginas Web contendo características comuns. De forma geral este processo pode ser descrito como a identificação de similaridades entre os vetores descritores das sessões de acesso. No caso de uma representação das sessões de acesso a partir de uma única dimensão, tal como o tempo de permanência, o cálculo de similaridades é relativamente simples. Quando o número de atributos analisados aumenta, tal como nos casos de utilização de informações semânticas, termos e conceitos, avaliação prévia por usuários, então a identificação de similaridade deve ser adequada e incorporar recursos para tratar a complexidade de cada caso. Algoritmos de segmentação como o *k-means* (Aldenderfer, 1984) partem de um número de segmentos desejado (k) e criam estes k segmentos com base na dimensão analisada. O algoritmo calcula a similaridade entre cada elemento e os segmentos iniciais. Então os elementos são alocados dentro do segmento mais próximo e o valor central de cada um dos segmentos é recalculado com a avaliação dos novos elementos. O processo do cálculo da similaridade de cada elemento e do valor central de cada segmento é repetido diversas vezes, sendo a condição de final da repetição a estabilização dos valores.

Algoritmos para a geração de regras de associação tratam os conjuntos gerais de visualizações de páginas Web que possuem uma grande quantidade de ocorrências em comum. O algoritmo *apriori* (Agraval e Srikant, 1994) é utilizado para minerar regras de associação. Ele utiliza grupos de itens associados em transações. No contexto da Mineração de Uso da Web estes conjuntos correspondem às sessões de acesso. A cada passo do algoritmo são identificados subconjuntos contendo um número mínimo

pré-definido de elementos em comum. Inicialmente são identificados os itens mais freqüentes. Em seguida são descartados os itens menos freqüentes deste conjunto e formados novos conjuntos com os itens restantes. O processo segue realizando estes passos até o momento em que todos os itens encontram-se acima do limite desejado. Neste ponto o processo finaliza e são geradas as regras de associação.

Padrões seqüenciais freqüentes permitem que dados originados em sessões diversas sejam analisados para a identificação de características comuns no comportamento dos usuários. Eles posibilitam a descoberta de seqüências de acesso repetidas por diversos usuários, em uma ordem definida. Servem assim para a localização dos caminhos percorridos pelos usuários em um *site* Web. Ao serem identificados possibilitam a avaliação de comportamentos futuros e análises cruzadas com outros fatores associados às páginas que consistem um padrão seqüencial.

3.3.1.3 Análise de padrões

A análise dos padrões extraídos pelas etapas anteriores deve estar integrada, no caso da mineração de uso da Web, com a finalidade do processo implementado. O tratamento a ser empregado para tarefas como personalização e adaptação de *sites* ou então para o acompanhamento de produtos e campanhas de comunicação é bastante diverso. No caso da personalização existe a necessidade de mecanismos que possam disponibilizar algum tipo de tratamento automático e de integração dos resultados com outros componentes de software, para uso imediato. Já no acompanhamento de produtos a informação obtida nos padrões minerados é utilizada para a confecção de relatórios gerenciais, tipicamente, não necessitando de tratamento integrado com alguma aplicação, de forma automática e online.

Assim, a representação dos padrões obtidos poderá ser empregada em informações que possam assessorar a tomada de decisões e que constituam crescimento para a modelagem de negócio. Para isto, o conjunto de padrões considerados interessantes deve ser analisado por um especialista no domínio da aplicação, que irá avaliar as regras e padrões encontrados a fim de detectar aquelas aplicáveis ao negócio.

Os resultados obtidos em processos de segmentação permitem a adaptação ou personalização de conteúdos. Em geral as informações obtidas são ilustrativas de interesses demonstrados pelos usuários durante suas visitas ao *site* Web. Processos de geração de regras de associação possibilitam a identificação de ligações importantes entre algumas páginas específicas e podem ser utilizadas tipicamente em situações de acompanhamento de promoções ou de campanhas de comunicação. Também servem para a identificação de comportamentos emergentes e importantes, dada a possibilidade de verificação das regras dentro de parâmetros de suporte e confiança.

Abordagens para personalização podem ser agrupadas em técnicas baseadas em conteúdo, filtagem colaborativa e formas híbridas contendo ambas as características. Suas diferenças situam-se nas estratégias e na informação usadas para a geração de opções de personalização.

Na abordagem baseada em conteúdo, o perfil pessoal do usuário é utilizado e identifica normalmente seus maiores interesses. Os conteúdos dos *sites* Web são classificados e representam alguns assuntos catalogados. Métricas para avaliar a similaridade de interesses descritos no perfil de usuários com os conteúdos apresentados são então realizadas e seus resultados aplicados em etapas de personalização. Esta abordagem pode ser observada em diversos trabalhos, com algumas variações, porém

mantendo o aspecto principal de adaptação do *site* Web de acordo com os interesses do usuário (Mikroyannidis, 2005). Esta abordagem permite a filtragem de conteúdos extensos com base no perfil do usuário e seu comportamento prévio. Porém existem situações em que este mecanismo pode não apresentar efetividade, como nos casos de um perfil recente e com pequeno histórico de navegação ou no caso de relações semânticas que não são capturadas pelo processo de análise de conteúdo.

As técnicas de filtragem colaborativa não realizam análise de conteúdo e atuam com base nas preferências ou atividades associadas com um usuário específico. Estas atividades são então comparadas com as atividades de todos os demais usuários, o que pode levar à identificação de um grupo de usuários que compartilha interesses e preferências. Para a identificação destas relações podem ser usadas diversas opções, tais como o acesso às mesmas páginas Web, a anotação de avaliação de itens de modo similar, entre diversas outras. Uma vez estabelecidos os grupos de interesses, a personalização pode ocorrer com base na sugestão dos itens ou páginas não visitados por um determinado usuário, mas constantes no conjunto de itens do seu grupo. Assumindo-se que o grupo possui interesses similares, considera-se provável o interesse de um usuário por estes itens constantes no grupo geral. Esta técnica apresenta problemas no caso de páginas recentes, para as quais não existe o tempo de acesso necessário para que sejam integradas a um determinado grupo (Konstan, 1997; Balabanovic, 1997).

Algumas abordagens utilizando ambas as técnicas são conhecidas (Middleton, 2002; Kleinberg, 2004) e permitem a diminuição das dificuldades apresentadas por cada uma individualmente. Utilizando as informações relacionadas com conteúdos e com interação, de forma integrada, o sistema de personalização pode apresentar maior eficiência.

3.3.2 Algoritmos de mineração de seqüências frequentes

A seguir são descritas características de algoritmos de mineração de uso da Web. São enfatizadas, por interesse para o trabalho presente, as peculiaridades que o contexto da Web traz para a implementação dos algoritmos. Em especial, são descritos algoritmos para o tratamento de seqüências freqüentes.

A mineração de seqüências freqüentes é descrita como de grande importância para uma gama importante de aplicações. A sua contextualização para o domínio de aplicações Web indica o processo de detecção de seqüências de acesso aos documentos Web, considerando-se as sessões dos diversos usuários. Para isso são descritas separadamente as técnicas associadas à obtenção de informações de uso e sua manipulação posterior.

Existem atualmente ferramentas de software que permitem a geração de informações de uso dos *sites* Web, porém com resultado insatisfatório para algumas demandas, como por exemplo, a identificação de percursos freqüentes. As alternativas observadas em trabalhos que buscam atender a esta necessidade geralmente envolvem, para a etapa de obtenção de informações, a utilização de mecanismos auxiliares para a complementação dos dados obtidos com os registros de acesso dos servidores Web. Uma das alternativas observadas é o uso de *cookies*, que contém informações específicas, armazenados no computador dos usuários. Outra é o uso de *scripts* junto às páginas Web, com o objetivo de gerar de informações adicionais. Este conjunto de informações pode ser utilizado para recuperar, por exemplo, percursos realizados por usuários no acesso a determinado *site* Web. A recuperação de percursos pode ser feita

sem a necessidade de identificação específica do usuário, o que torna possível a manutenção da sua privacidade.

O tratamento dos padrões de acesso observados permite a recuperação de informações bastante úteis ao contexto de sistemas de recomendação ou de sistemas de personalização. Estas informações podem ser tratadas por algoritmos específicos, tanto para a geração de regras de associação, como para geração de percursos (seqüências freqüentes). Além do tratamento específico destes dados de acesso, pode-se observar a identificação de novas possibilidades de associação das informações quando o tempo de ocorrência é levado em consideração. A identificação destes padrões é referida por autores como padrões de acesso seqüencial periódico, ou como mineração temporal de acesso Web. Nesse caso, um determinado padrão de acesso pode ser comparado com diversos outros para verificação do momento de ocorrência. A repetição de padrões pode também ser tratada a partir de um determinado intervalo de tempo.

Esta forma de mineração permite, por exemplo, que sejam detectadas seções de *sites* Web que são muito freqüentemente acessadas em determinados períodos do dia. Esta associação pode ser estendida aos hábitos dos usuários deste *site* Web. Um exemplo seria a descoberta de padrões freqüentes de acesso a seções de notícias durante o início da manhã e seções de informações financeiras no final da tarde. O intervalo de tempo considerado pode ser tratado de forma a variar entre horas, dias, semana, meses ou anos, se desejado. Também nestas situações em que o intervalo é maior, como no caso de semanas ou meses, podem ser descobertos padrões importantes como o acesso às informações de programação e transportes de um *site* Web de um evento nas primeiras semanas de seu lançamento e o acesso às informações de inscrição e lista de artigos aceitos no período próximo das datas de realização do evento.

Para o processo de análise dos dados obtidos, com a mineração das seqüências freqüentes, são conhecidos diversos algoritmos. Algumas abordagens observadas podem ser citadas, como Agrawal, (1996), com o algoritmo GSP (*Generalised Sequential Patterns*), que trabalha com seqüências candidatas e com múltiplas passagens pela base de dados, sendo que um dos de seus problemas pode ser identificado no caso de seqüências longas, o que é freqüentemente observado no caso de acessos a *sites* Web. O algoritmo SPADE, descrito por Zaki (2001), evita esta situação, tratando os dados previamente e gerando uma base de dados vertical contendo as seqüências candidatas e sua identificação, permitindo sua enumeração de forma mais eficiente. Algumas extensões deste algoritmo tratam de aspectos específicos, melhorando sua performance, como o caso de Leleu (2003), que introduz o conceito de ocorrências generalizadas no algoritmo GO-SPADE, ou de Demiriz (2002), que propõe uma versão paralela do algoritmo SPADE.

Abordagens mais recentes procuram evitar problemas associados ao tratamento de seqüências candidatas, com uso de recursos como a estrutura WAP-Tree (Pei, 2000), permitindo a descrição e tratamento das seqüências com maior performance e flexibilidade do que versões anteriores. Outras abordagens como CCSM (Orlando, 2003), FS-Miner (El-Sayed, 2004), CLOSET+ (Wang, 2003) ou FAS-Tree (Xiaoqiu et al., 2006) tratam de melhorias na performance e flexibilidade da etapa de mineração de seqüências e detecção de padrões, com uso de estruturas de dados mais eficientes.

Resumo do Capítulo:

Neste capítulo é apresentado brevemente o processo de descoberta de conhecimento em base de dados. A área de Mineração da Web é descrita resumidamente. São descritas características importantes da Mineração de Uso da Web e relatados aspectos gerais de alguns algoritmos de Mineração de Uso da Web para obtenção de sequências freqüentes.

4 HIPERMÍDIA ADAPTATIVA

Neste capítulo são apresentados os conceitos gerais da área de Hipermídia Adaptativa, as principais técnicas conhecidas e exemplos de sistemas existentes.

4.1 Visão Geral

A pesquisa na área de Hipermídia Adaptativa possui como objetivo a melhoria da experiência dos usuários durante o acesso aos conteúdos disponibilizados por sistemas de hipermídia. Considera-se que esta melhoria de usabilidade pode ser alcançada a partir da construção de modelos capazes da representação dos conhecimentos, habilidades, objetivos e preferências dos usuários. Além da modelagem do usuário são observadas técnicas de construção das interfaces de forma flexível e técnicas para acompanhamento do uso destas. A utilização das informações de tais modelos em conjunto com outras informações complementares como o contexto da aplicação, dados de interação do usuário ou de grupos de usuários, regras de adaptação, entre outros, permitem a identificação de possíveis tópicos de interesse, restrições de acesso e adaptações de conteúdo e formato dos *sites* Web (Brusilovsky, 1996; Brusilovsky, 2004; De Bra, 1999; De Bra 2004).

Seguem algumas definições conhecidas, obtidas de diferentes autores e em diferentes momentos, que colaboram para identificar uma tendência observada no sentido de ampliação das áreas de utilização, integração de técnicas das áreas de Processamento de Linguagem Natural, Inteligência Artificial, Banco de Dados e também a utilização em contextos de dispositivos móveis. Também podem ser observadas nas definições as menções a componentes como modelo de usuário, modelo do domínio e modelo de adaptação.

Segundo Brusilovsky (1996):

> *"A Hipermídia Adaptativa envolve a criação e manutenção de um modelo do usuário e acompanhamento de sua interação para a adaptação segundo suas necessidades."*

Segundo Markellou et al. (2005):

> *"Personalização pode ser uma solução para o problema de sobrecarga de informações, a partir de integração de recursos de áreas como Recuperação de Informação, Modelagem de Usuários, Inteligência Artificial, Banco de Dados e PLN, entre outras."*

Segundo Zimmerman et al. (2005):

> *"Um sistema de hipermídia adaptativa segue uma estratégia de adaptação e um objetivo de adaptação, considerando-se informação relevante a respeito do usuário do contexto de interação."*

Observa-se pelas definições apresentadas, a tendência de integração de novas técnicas e de uso de recursos existentes em equipamentos móveis e sensores, entre outros. Também fica evidenciada a ênfase na utilização de informações de contexto, para a geração das adaptações.

Diversos sistemas de Hipermídia Adaptativa vem sendo desenvolvidos, por diferentes grupos de pesquisa, em diversas áreas. Algumas das mais evidentes são a educação, interfaces para sistemas de recuperação de informação, sistemas de apoio à localização, turismo ou visitação de museus e bibliotecas. Muitos destes sistemas utilizam interfaces de identificação e caracterização do perfil de cada usuário, enquanto que outros empregam técnicas não-invasivas e buscam automaticamente a obtenção de informações para a geração e manutenção de informações a constarem no modelo representativo do usuário (Dolog, 2004).

É importante ressaltar uma diferenciação dos sistemas de Hipermídia Adaptativa para com os sistemas conhecidos como Adaptáveis, descritos a seguir. Observam-se, de forma bastante freqüente, sistemas que possibilitam ao usuário alternativas para a personalização, através da seleção explícita de conteúdos e formatos de apresentação desejados. Exemplos podem ser encontrados em alguns dos principais portais Web, onde as ferramentas de gerenciamento de conteúdo possibilitam a identificação de perfis gerais aos quais estão associadas diversas opções, a serem escolhidas pelos usuários sempre que desejado. Este tipo de sistema não possui um comportamento flexível e ativo observado nos demais sistemas de Hipermídia Adaptativa onde o formato e conteúdo podem ser obtidos de forma automática e independente da indicação de opções pelo usuário. Mesmo assim, possibilitam ao usuário a melhoria de sua experiência no uso das interfaces (Tsandilas, 2003).

Na figura 4.1 a seguir está descrito um exemplo de sistema adaptável. Alguns dos aspectos da interface são customizados pelo usuário, de forma estática, a partir de formulário específico e com recursos limitados. No caso pode-se ilustrar a personalização de conteúdo e formatos de apresentação em portais Web. Por exemplo, na porção esquerda da figura podem ser vistas opções de escolha de conteúdos e de escolha de uma mensagem de saudação. O usuário pode indicar os conteúdos desejados selecionando o item (*checkbox*) posicionado na esquerda do título de cada conteúdo. Na porção direita da figura são exemplificadas as opções de personalização do formato de apresentação destes conteúdos selecionados. O usuário pode escolher a apresentação em duas ou três colunas e depois disso pode selecionar a coluna onde deve ser exibido cada conteúdo. Ressalta-se que esta operação é definida pelo usuário e seus resultados são válidos para todas as futuras sessões de uso do sistema, até nova troca manual das opções.

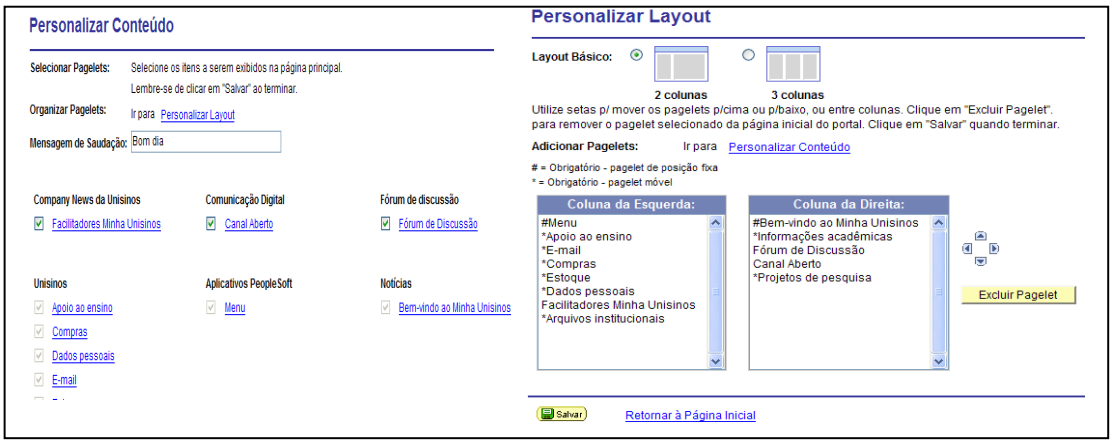


Figura 4.1: Exemplos de Hipermídia Adaptável

Para que seja possível a adaptação, relacionada com o modelo dos usuários, existem algumas premissas a serem observadas nos sistemas de Hipermídia Adaptativa. Os documentos disponibilizados necessitam de uma descrição em um modelo de domínio, onde podem ser observadas as relações entre os conceitos representados, nos seus diversos níveis. Estes podem ser mais abstratos (conceitos mais gerais), podem representar agrupamentos (tópicos de um assunto geral) ou podem descrever informações específicas e detalhadas de um determinado tópico. Esta representação deve ser empregada em conjunto com a indicação de relacionamentos diversos entre os elementos. Assim as diversas associações possíveis estarão disponíveis para uso pelo sistema, sendo que os conteúdos do sistema de hipermídia podem ser observados neste modelo de domínio (Wu, 2002).

De forma geral, este modelo do domínio é utilizado pelo sistema de Hipermídia Adaptativa em diferentes tarefas, observadas em um comportamento genérico deste. A primeira tarefa seria o acompanhamento de uso do sistema, onde as ações do usuário são anotadas e relacionadas com os elementos do modelo de domínio. No caso de sistemas voltados para uso na Internet este acompanhamento equivale, em geral, à seqüência de *hyperlinks* percorridos. Uma segunda tarefa seria o uso do modelo do usuário para a classificação das informações do conteúdo disponibilizado pelo sistema, de acordo com as preferências, objetivos e capacidade do usuário. A terceira tarefa seria a utilização de informações obtidas nas duas tarefas anteriores para a geração da interface a ser exibida para o usuário, disponibilizando assim os conteúdos de acordo com sua necessidade e interesse.

Uma parcela importante dos trabalhos de Hipermídia Adaptativa gera e atualiza o modelo do usuário com abordagens baseadas no acompanhamento de uso, variando em diversos aspectos. Além desta abordagem são conhecidos trabalhos envolvendo outros fatores para a descrição deste modelo. Uma destas abordagens trata a geração e manutenção destes modelos observando o perfil cognitivo dos usuários (Souto et al., 2005). Desta forma são mapeados, normalmente a partir de interações diretas ou entrevistas, aspectos cognitivos que serão posteriormente representados no modelo do usuário e tratados pelo sistema na geração de adaptações (Stash et al., 2006, Souto et al., 2002).

Abordagens para avaliação do uso de sistemas de Hipermídia Adaptativa podem ser encontradas na forma de acompanhamento do tempo necessário para a obtenção dos

resultados desejados pelos usuários em determinadas tarefas ou no acompanhamento da adequação das informações apresentadas (Brusilovsky, 2004). A justificativa destes sistemas de Hipermídia Adaptativa, entretanto, encontra-se relacionada de forma bastante evidente ao crescimento da Internet e ao aumento do número de informações disponibilizado a partir de tecnologias de digitalização de documentos, integração de sistemas de bases de dados e convergência de mídias. Observa-se um número cada vez mais elevado de informações disponíveis, o aumento da diversidade de formatos para armazenamento e integração destas informações e o aumento da velocidade de geração e atualização. Mesmo com o uso de ferramentas de apoio como sistemas de gerenciamento de conteúdo Web ou sistemas de gerenciamento de documentos, o grande volume e diversidade das informações não permite que estas sejam tratadas (em alguns contextos) de forma adequada pelos diferentes usuários interessados no acesso. Outra justificativa (De Bra, 2004) para o uso de sistemas de Hipermídia Adaptativa está relacionada com o aumento do número de usuários e com a sua diversidade de conhecimentos e interesses. Esta situação torna extremamente difícil o projeto de uma interface adequada, de forma estática, dadas as diferenças dos usuários.

De forma resumida, a adaptação gerada pelos sistemas de hipermídia adaptativa está relacionada com os conteúdos e seu formato de apresentação. Assim, podem ser disponibilizadas informações com maior ou menor profundidade ou detalhes sobre o assunto de interesse do usuário. Estas informações podem ser apresentadas em uma interface diferenciada, com maior presença de texto, imagens, determinada configuração de cores ou uso de recursos auxiliares como som, vídeos ou animações (Christopher, 2002; Paramythis e Stephanidis, 2005).

Com a melhoria observada em diversos dispositivos esta adaptação torna-se necessária em um novo contexto, mais abrangente do que a interface padrão de programas de navegação de hipertexto utilizados em computadores de mesa. A integração de equipamentos móveis e de telefonia vem gerando a necessidade de envio de informações para dispositivos que possuem diferentes condições de armazenamento, processamento e exibição dos dados (Petrelli, 2005).

A integração de informações geradas por dispositivos diversos, tais como sensores, também vem a contribuir para alterar o panorama na área de Hipermídia Adaptativa, por tratar-se de informações que podem ser utilizadas de forma muito efetiva na determinação de interesses e atitudes demonstradas pelos usuários, possibilitando um ciclo de interação mais curto e melhorias no processo de inferência e adaptação (Zimmerman et al., 2005).

Iniciativas para promover o reuso em sistemas de Hipermídia Adaptativa também são encontradas recentemente, dado que a implementação e manutenção destes sistemas é uma tarefa que envolve um esforço considerável. Assim alguns trabalhos tratam desta questão propondo a utilização de linguagens específicas, tais como em Stash, Cristea e De Bra (2007). Neste trabalho é proposta e demonstrada a viabilidade de uso de uma linguagem específica para a descrição de ações de adaptação.

4.2 Técnicas utilizadas

Nesta seção serão descritas técnicas envolvidas na geração de sistemas de Hipermídia Adaptativa. Para facilitar a identificação de técnicas empregadas nestes

sistemas, será traçado um breve resumo de suas características sob uma perspectiva histórica.

Os sistemas de Hipermídia Adaptativa são conhecidos desde a década de noventa, com trabalhos e iniciativas para a descrição de métodos para adaptação. A área com maior quantidade de aplicações foi sem dúvida a educação. Já entre 1996 e 2001, o aumento de utilização e a popularização da Web e consolidação de recursos multimídia permitiram a extensão de técnicas para diversas áreas de aplicação. Também são observadas iniciativas no sentido de melhorias na modelagem de usuários e domínios. A partir deste período, observa-se a implantação em escala crescente de dispositivos de computação móvel, fato que fez com que o contexto de uso das aplicações recebesse uma atenção importante e fosse tratado de forma mais específica em sistemas para adaptação. Também neste período observa-se o crescimento do uso de recursos da Web Semântica e de tecnologias como sistemas de recomendação, mineração de dados e processamento de linguagem natural.

Este breve relato histórico aponta para a ampliação de recursos de modelagem, a utilização de novas técnicas de Inteligência Artificial e da disseminação de sistemas adaptativos em áreas diversas, como um componente complementar.

Também é muito importante a ampliação do papel de dispositivos móveis na geração de informações de contexto. A integração de informações geradas por dispositivos diversos, tais como sensores, também vem a contribuir para alterações no panorama da área de Hipermídia Adaptativa, por tratar-se de informações que podem ser utilizadas de forma muito efetiva na determinação de interesses e atitudes demonstradas pelos usuários, possibilitando um ciclo mais dinâmico de adaptações.

A partir deste contexto mais geral, são conhecidos mecanismos de adaptação nos diversos sistemas desenvolvidos. As possibilidades de adaptação observadas estão relacionadas com diversas características e podem ser sumarizados na figura 4.2 a seguir. Esta figura é descrita por Brusilovsky (2001), como um resumo abrangente das possibilidades de Hipermídia Adaptativa. Nela podem ser observada a divisão da adaptação em um processo de adaptação de apresentação e outro de navegação. As possibilidades relacionadas com a primeira opção se organizam em adaptações para apresentação de multimídia, apresentação de texto e ainda adaptação de modalidade. Em relação às possibilidades de apresentação de textos, existem duas opções mais gerais, onde a primeira está voltada para a adaptação de linguagem natural, usada em sistemas que empregam abordagens mais ricas de interação, como simulações de diálogo. Já a segunda utiliza diversas possibilidades para a adaptação do texto disposto em páginas Web, tais como a inserção ou remoção de fragmentos, a alteração destes e sua ordenação. As opções de adaptação relacionadas com o suporte à navegação possuem seu foco na estrutura das páginas apresentadas. Podem ser observadas alternativas para promover navegações guiadas, ordenações diversas dos hyperlinks, apresentação ou não destes, associação de informações adicionais aos hyperlinks e ainda a geração adaptativa destes, que possibilita que a própria estrutura do *site* Web seja modificada de acordo com o contexto da adaptação.

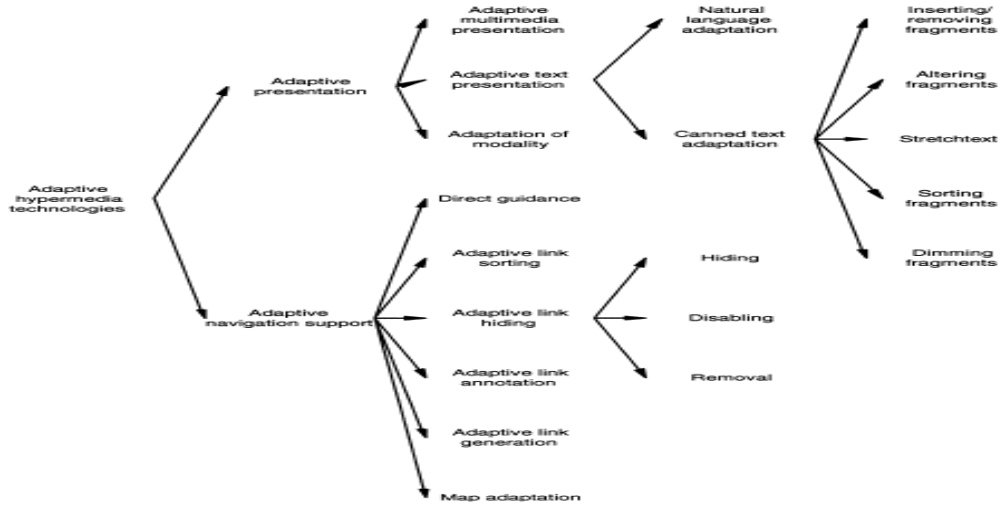


Figura 4.2: Resumo de possibilidades de adaptação

A figura 4.2 possibilita um retrato parcial, entretanto, das opções encontradas em sistemas atualmente, conforme será visto nos exemplos descritos. Isto se deve ao fato de haver diversas combinações entre estas formas mais gerais. Assim, apesar de importante como referência, devem ser analisados também outros fatores associados às possibilidades de adaptação, bem como os recursos necessários para a sua implementação. Por exemplo, Uma importante divisão encontrada em trabalhos mais recentes separa a adaptação de conteúdo e de apresentação, por serem relacionados com recursos diferentes para sua modelagem e implementação.

A tabela 4.1 a seguir apresenta uma organização das possibilidades de adaptação indicadas na figura 4.2, porém com uma coluna complementar indicando recursos utilizados. Ela serve, desta forma, para auxiliar a compor uma visão geral da utilização de diferentes recursos tecnológicos, como a integração de tecnologias da Web Semântica, técnicas como mineração de dados e processamento de linguagem natural, no contexto de adaptação.

Tabela 4.1: Associação de possibilidades e técnicas de adaptação

O que adaptar	**Descrição**	**Tecnologias associadas**
Conteúdo	Ajuste do conteúdo apresentado (por exemplo, uso de linguagem coloquial ou técnica, conteúdo simplificado ou detalhado).	Anotação semântica, mineração de conteúdo, PLN, mecanismos de inferência.
Navegação	Organização do conjunto de nodos e links utilizados (por exemplo, inclusão ou retirada de elementos de acordo com o nível de conhecimento do usuário, retirada de links em determinadas operações, tais como testes).	Modelagem semântica de aplicações Web, mineração de estrutura e de uso, mecanismos de inferência.
Apresentação	Escolha de elementos da interface apresentada ao usuário (por exemplo, uso de imagens ou texto dependendo do suporte, uso de texto ou ícones para links).	Modelagem semântica de aplicações, uso de regras.

A utilização de modelos descritores do usuário, domínio e formas de adaptação é observada nos sistemas de Hipermídia Adaptativa mais recentes. Algumas iniciativas de descrição de aplicações Web com modelos contendo uma semântica mais rica podem ser encontrados e se justificam, apresentando maiores possibilidades de tratamento

automático para as questões de adaptação. Neste sentido, alguns trabalhos recentes trazem a proposta de construção de um meta-modelo para a integração dos requisitos de descrição de um sistema de Hipermídia Adaptativa. Existem diversas vantagens relacionadas, como a possibilidade de compartilhamento de dados do usuário entre sistemas diferentes e também a possibilidade de reuso e melhor configuração de componentes.

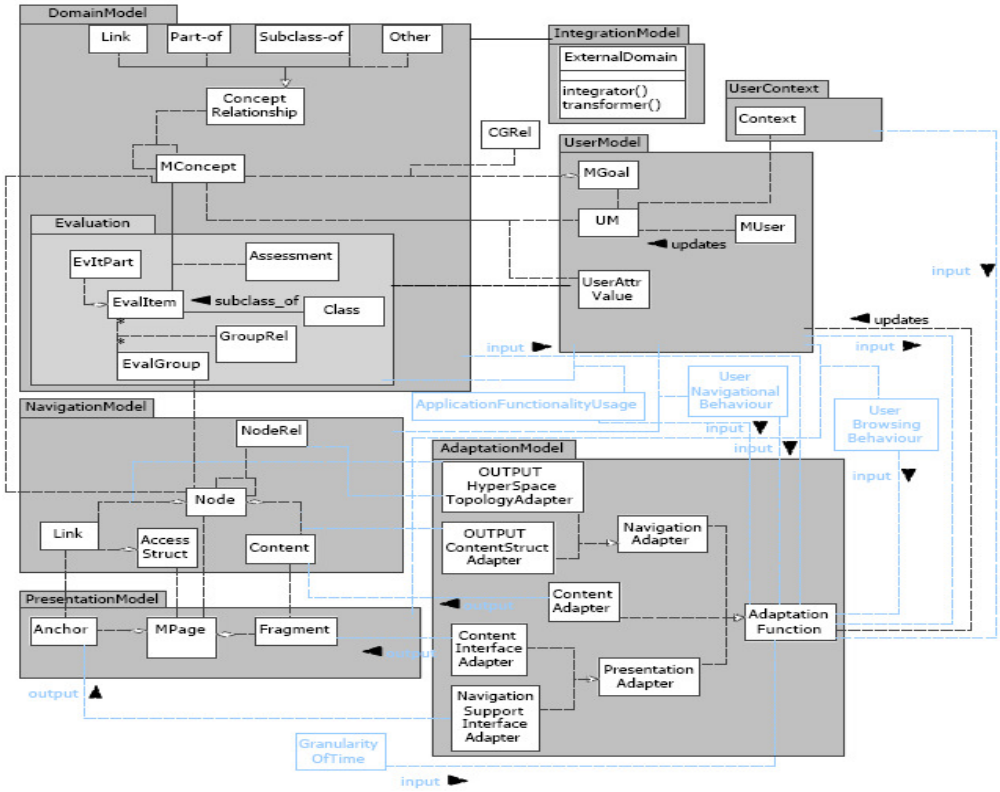


Figura 4.3: Exemplo de modelo para meta-adaptação

Uma destas iniciativas é ilustrada na figura 4.3, onde é descrita uma proposta de um metamodelo envolvendo os componentes principais deste tipo de sistema, como o modelo do usuário, o modelo do domínio, o modelo de adaptação e também um modelo de navegação. Neste caso o modelo é descrito com uso de notação UML (Assis, 2004).

A personalização, no contexto da Web, pode ser considerada como o processo de adequação das páginas visitadas por um determinado usuário, a partir de suas preferências e características (Mobascher et al., 2000). Neste sentido, está associada em sua origem a processos de marketing e tem como objetivo a busca do atendimento mais eficiente e adequado das necessidades de usuários. Exemplos deste tipo de abordagem de personalização podem ser encontrados em muitos trabalhos e diversos portais de comércio eletrônico, ou em *sites* Web de instituições bancárias. Pode-se, entretanto, observar o uso de técnicas de personalização em múltiplos contextos, como no caso de situações de ensino à distância, onde a adequação do conteúdo e das formas de apresentação a um determinado aluno, em seus interesses e a partir de seu histórico no sistema seriam o objetivo principal. Também pode ser observada a utilização de conceitos de personalização em sistemas de recuperação de informações, onde o perfil do usuário e seu histórico no uso do sistema pode ser utilizado para tarefas de classificação de documentos retornados em pesquisas, por exemplo.

Diferentes técnicas podem ser utilizadas para esta finalidade, especificamente na obtenção e atualização de um perfil ou modelo de usuário e na filtragem de informações para a identificação de comportamentos específicos e que possam ser generalizados para um grupo de usuários. As mesmas técnicas podem ser utilizadas em um contexto diferenciado, denominado de recomendação, que será brevemente comentado a seguir.

A recomendação busca identificar características de comportamento de grupos de usuários que possam ser utilizadas para realizar sugestões a um determinado usuário, com uma boa probabilidade de que este as aceite (Herlocker, 2000). Também são observadas características dos documentos ou produtos a serem recomendados e busca-se a integração destas informações. Neste sentido a recomendação pode se valer de algoritmos mais complexos e utilizar grandes conjuntos de dados sem que isso prejudique o seu resultado, pois normalmente não existe uma ação imediata a ser realizada, em função de comportamentos do usuário. De modo diverso da personalização, que possui características importantes nesta associação com uma situação que demanda resposta imediata, a recomendação pode ser implementada a partir de processos mais extensos e com resultados já previamente definidos, onde um conjunto de itens a ser recomendado para um dado perfil de usuário é obtido e a partir deste momento fica disponível para ser indicado, quando necessário. Existem enfoques diferenciados possíveis para a implementação de recomendação, como a filtragem colaborativa (onde são levadas em conta as informações obtidas com avaliações de usuários considerados em grupos similares ao grupo ao qual pertence o usuário em questão) ou a filtragem baseada em conteúdo (onde as preferências do usuário em questão e o conhecimento acerca do conteúdo dos itens são determinantes). Segundo Balabanovich (1997), a união destas duas abordagens, referenciada como abordagem híbrida pode levar a melhores resultados, por utilizar vantagens de ambas para o processo de recomendação. Em Middleton (2002) é justificado o uso de ontologias em conjunto com sistemas de recomendação, para evitar problemas decorrentes da pouca informação existente sobre o usuário quando de seu período inicial de interação com o sistema. Este fato normalmente pode ocasionar resultados insatisfatórios, sendo que uma ontologia pode ser utilizada, como neste sistema, para que informações gerais sobre usuários e domínios de interesse possam ser empregadas como forma de melhorar os resultados nestas situações.

As tarefas de personalização e recomendação diferenciam-se, portanto, de modo a evidenciar situações com necessidades imediatas e situações onde não existe esta necessidade imediata. Entretanto as duas podem ser observadas em um mesmo processo geral, onde existe inicialmente a necessidade de conhecimento do usuário, normalmente a partir de um perfil deste. Este perfil será utilizado em um processo de associação onde serão obtidas informações relacionando assuntos ou itens que provavelmente serão de interesse de usuários que estejam associados a este perfil. Neste processo normalmente podem ser observadas técnicas diversas como mineração de dados, mineração de textos, mineração de uso, combinadas com procedimentos diversos de classificação ou agrupamento. Neste sentido estes processos podem enfrentar os mesmos problemas que sistemas de recuperação de informações, pois durante o processo de mineração de dados na Web, por exemplo, os textos normalmente estarão sendo utilizados sem que sejam levadas em conta informações semânticas associadas, o que pode gerar erros nos resultados.

Resumo do Capítulo:

Neste capítulo são relatadas características gerais da área de Hipermídia Adaptativa. Foram descritas as principais abordagens para adaptação e também relacionadas tecnologias necessárias para sua implementação. A tendência ao uso de recursos semânticos nas diversas etapas foi destacada, a partir da possibilidade de uso de meta-modelos e de outras situações relacionadas, como recomendação.

5 METODOLOGIAS PARA APLICAÇÕES WEB

Este capítulo relaciona algumas propostas para a descrição de aplicações Web. Estas abordagens são de interesse para o presente trabalho, pois contribuem com melhorias nas possibilidades de projeto e também de utilização de recursos Web. Em alguns dos casos podem também ser observadas ocorrências de recursos semânticos no escopo das abordagens.

5.1 Apresentação

Acompanhando o cenário do desenvolvimento da Web, observa-se que os primeiros sistemas não possuíam as características evidenciadas atualmente, tais como grande integração com bases de dados, grande diversidade de áreas de aplicação e possibilidades de compartilhamento de serviços, entre outros. Ao relatarmos alguns dos tipos de sistemas atualmente disponíveis, encontram-se exemplos diversos tais como portais corporativos, portais de notícias, sistemas colaborativos e voltados para comunidades mediadas pela Internet, sistemas de comércio eletrônico ou bibliotecas digitais. Todos estes sistemas compartilham a mesma necessidade de tratamento de um volume expressivo de dados com características diversas e a geração de resultados que sejam adequados aos diferentes usuários e situações de consulta. Além disso, este tipo de sistema tem se tornado o sistema dominante em diversos contextos, existindo atualmente uma grande quantidade e necessidade de sistemas deste tipo. O estabelecimento de abordagens sistemáticas que levem em conta suas características específicas é desenvolvido a partir de uma disciplina emergente, denominada Engenharia de Web (Murugesan e Deshpande, 2001; Pressman 2005).

Diversas metodologias têm sido propostas para o tratamento desta classe de aplicações, sendo que neste capítulo são destacadas e descritas brevemente algumas destas que possuem em comum características orientadas por modelos ou a integração de recursos semânticos. Além de vantagens como facilidade de reuso e de manutenção, deve-se destacar outras como a facilidade de descrição e compreensão do sistema e seus componentes, possibilidade de verificação de validade, bem como a possibilidade de geração automática ou semi-automática de resultados a partir de modelos. As metodologias relacionadas utilizam recursos da Web Semântica para a descrição da aplicação, possibilitando maior interoperabilidade entre sistemas e maior flexibilidade na utilização destes modelos. A possibilidade de personalização dos resultados, um fator de interesse para o atual trabalho, é bastante facilitada a partir de uma abordagem de modelagem com maior descrição semântica.

5.2 Exemplos de metodologias

A seguir são descritas brevemente algumas metodologias para descrição de aplicações voltadas para a Web, empregando abordagens voltadas a modelos e recursos da Web Semântica.

5.2.1 RMM (Relationship Management Methodology)

A abordagem seguida pela RMM está ligada ao gerenciamento de relacionamentos entre os objetos de informação (Isakowitz et al., 1995; Diaz et al.,1997). Esta proposta é uma das primeiras conhecidas neste sentido e está relacionada neste trabalho por motivos históricos. Utiliza o modelo de entidade – relacionamento (ER) para a descrição e integração de quatro atividades básicas: projeto ER, projeto da aplicação, projeto da interface de usuário, construção e teste.

A metodologia propõe o uso de algumas ferramentas para complemento do trabalho e busca, com a utilização dos modelos ER, para a descrição em cada domínio atendido por uma aplicação, de suas principais entidades, seus relacionamentos e seus atributos. A etapa de projeto da aplicação relaciona estes elementos com outros como visões, índices, hyperlinks e roteiros. Na etapa seguinte são descritas composições destes elementos com padrões de páginas HTML. Assim existe uma formalização e descrição do contexto da aplicação que pode ser utilizada nas etapas de projeto seguintes e também na etapa de construção da aplicação.

5.2.2 OOHDM (Object-Oriented Hypermedia Design Methodology)

A metodologia OOHDM, descrita por Schwabe (1996) e Schwabe e Rossi (1998), utiliza o modelo de Orientação a Objetos e define cinco etapas para tratamento de aplicações Web: especificação de requisitos, projeto conceitual, projeto navegacional, projeto de interface abstrata e aplicação.

A partir da especificação dos usuários do sistema e de suas atividades, definidas na etapa inicial de especificação de requisitos, são descritos os principais componentes do sistema, de forma independente de usuários e atividades, a partir de uma notação de classes. Cada classe descrita pode ser complementada com atributos específicos e pode ser referenciada na etapa seguinte, em visões específicas de navegação, que compõe o modelo navegacional. Na etapa de aplicação, existe o mapeamento das descrições de classes navegacionais para componentes de páginas HTML.

Extensões deste método, descritas adiante neste capítulo, acrescentam recursos da Web semântica às etapas definidas.

5.2.3 WSDM (Web Site Design Method)

Esta metodologia parte do pressuposto da modelagem para diferentes públicos, a partir do qual pretende facilitar tarefas como personalização. Assim o WSDM, segundo De Troyer (1998) consiste em cinco etapas*: mission statement, audience modeling, conceptual design, implementation design, implementation*.

As duas etapas iniciais tratam da definição de tarefas e do agrupamento destas, o que descreve audiências específicas. A etapa seguinte descreve dados e tarefas, sendo que uma descrição destas é providenciada por uma linguagem voltada para notação de ontologias (no caso, a linguagem OWL). O projeto de navegação permite que sejam agrupados os componentes descritos anteriormente em um grafo, onde cada audiência diferenciada nas etapas anteriores de modelagem é indicada como uma área independente deste grafo. A etapa de projeto de implementação descreve os elementos de páginas, apresentações e dados, permitindo que atributos sejam descritos para cada um destes, de acordo com as necessidades das audiências. A última etapa permite que o resultado descrito anteriormente seja adequado a uma linguagem específica de apresentação.

5.2.4 OOWS (Object Oriented Web Solution)

A metodologia orientada a objetos pode também ser verificada na proposta do método OOWS (Pastor, 2005; Fons, 2002). Nesta abordagem são descritas duas etapas: modelagem conceitual e desenvolvimento da solução. A primeira etapa incorpora a descrição de requisitos, a modelagem conceitual dos elementos do domínio e a descrição do modelo de navegação e apresentação. Todos estes componentes são descritos com diagramas da UML.

Para a implementação da aplicação são empregados agentes que realizam a tarefa de associação de modelos, dados e geração de resultados de apresentação.

5.2.5 WebML (Web Modeling Language)

No caso da WebML, descrita por Ceri (2000, 2003), o sistema Web é descrito por uma linguagem de alto nível para modelagem, a partir de uma interface visual e do uso de XML como forma de serialização. São utilizadas apenas três etapas na metodologia: projeto de dados, projeto de hypertexto e implementação.

O projeto de dados possui uma característica similar ao método RMM, descrevendo os elementos de dados em um esquema similar a um diagrama ER. São descritas as principais entidades da aplicação, seus atributos, relacionamentos e também características que podem ser utilizadas para personalização. O projeto de hypertexto baseia-se em unidades pré-definidas, que podem ser associadas com páginas. Uma página pode assim realizar a descrição da estrutura de navegação da aplicação e organizar visões específicas para determinados usuários. Existem diversos tipos de hyperlinks pré-definidos na metodologia para explicitar as ligações entre unidades de páginas e também entre páginas. A etapa de implementação realiza a associação entre as unidades de dados definidas com fontes de dados e a descrição das páginas é mapeada, através de *templates* JSP (*JavaServer Pages*). Existe uma ferramenta visual para a descrição dos modelos e um ambiente de execução das aplicações.

5.2.6 XWMF (eXtensible Web Modeling Framework)

O XWMF (Klapsing e Neumann, 2000; Klapsing et al., 2001) é um framework para modelagem de aplicações Web a partir do uso de RDF e da linguagem WOCM (*Web Object Composition Model*), que permite a definição da estrutura e do conteúdo da aplicação Web. A representação da WOCM é feita em RDF e ela permite definir grafos diretos acíclicos, onde elementos chamados "*complexons*" são descritores de nodos e "*simplexons*" são descritores de folhas. Características de adaptação para

dispositivos específicos ou para validade de dados, por exemplo, podem ser definidas com atributos específicos, verificados em tempo de geração da codificação de apresentação.

A descrição de aplicações é feita em um ambiente específico, desenvolvido com o uso da linguagem Tcl e de Prolog. Neste ambiente é possível a descrição da aplicação e também a sua instanciação, para finalidades de validação e de testes.

5.2.7 OntoWebber

O OntoWebber (Jin et al., 2001) consiste em uma metodologia voltada para a descrição de aplicações Web ricas em semântica, sendo adotada a perspectiva do uso de ontologias. São previstas três camadas na arquitetura: integração, composição e geração. Na camada de integração são previstas tarefas de integração de dados e de resolução de diferenças de sintaxe ou de semântica, sendo usadas para isso descrições em RDF e uma ontologia de domínio como referência. A camada seguinte também se vale de descrições em ontologias para o manuseio das dimensões envolvidas na aplicação. São definidas ontologias para a descrição de: navegação, conteúdo, apresentação, personalização e, por fim, manutenção. A camada de geração se vale destes dados, em formato RDFS e realiza tarefas de verificação de restrições e de instanciação, gerando os resultados para a apresentação final.

Para as tarefas de composição, as ontologias descrevem a navegação a partir dos seguintes elementos: "*cards*", "*links*" e "*pages*". Um elemento "*page*" integra os demais elementos, que podem ser dinâmicos, de acordo com a visão desejada do *site* a ser gerado. A descrição de conteúdo é feita de forma diferenciada para conteúdos estáticos (texto, imagens) e dinâmicos. Estes últimos são definidos como entidades na ontologia de domínio. Para sua consulta é usado o mecanismo disponibilizado pelo sistema de inferência TRIPLE (Sintek e Decker, 2002). A descrição de apresentação associa elementos de estilo aos elementos de navegação definidos anteriormente. A modelagem de personalização permite a identificação de recursos de personalização para grupos ou para usuários. Estes recursos são associados a modelos de usuários ou grupos, onde são descritas características destes (nome, idade, gênero) e também propriedades indicando interesses e condições de navegação desejadas.

A implementação de *sites* Web com esta metodologia é suportada em um ambiente que permite a criação e manutenção das diversas ontologias indicadas e sua integração para a geração dos resultados finais.

5.2.8 SEAL (SEmantic PortAL)

A proposta do SEAL (Maedche et al., 2002; Maedche et al., 2003) é similar à proposta do OntoWebber e trata-se de uma metodologia baseada em ontologias para a descrição de portais Web. Trata das possibilidades de organização de uma grande quantidade de dados com recursos que permitam maior integração de semântica. As etapas para a construção destes portais Web seriam: projeto de ontologia, integração de dados, projeto de *site* e implementação. Na primeira etapa é utilizada a linguagem RDFS e regras para a descrição de conceitos e restrições. A segunda etapa trata da descrição e conversão de todos os dados manipulados para o formato RDF. São descritos diversos mecanismos de integração de dados existentes em formatos diferenciados. Para o projeto de *site* Web existem tarefas integradas que seriam a construção do modelo de navegação, modelo de entrada e modelo de personalização. A principal informação descrita nestes modelos é a associação das informações descritas

na ontologia de forma a comporem uma visão integrada do *site* Web, para um usuário específico.

Para sua implementação é utilizado o framework KAON, a partir do qual existem ferramentas integradas para a descrição de ontologias, integração de dados e geração do portal Web, a partir da identificação de usuários e seu contexto.

5.2.9 SHDM (Semantic Hypermedia Design Method)

A proposta do SHDM (Lima, 2003) trata da expansão da metodologia descrita pelo OOHDM (referido no item 6.2.2) com o uso de uma abordagem baseada em ontologias, expressas em linguagem OWL. Para sua descrição e implementação são integradas ferramentas de edição de ontologias, conversão de formatos e geração da navegação, a partir das ontologias. O método parte de cinco etapas, que são: levantamento de requisitos, modelagem conceitual, modelagem navegacional, projeto de interface abstrata e implementação. Em cada uma das etapas é descrito um modelo contendo as informações relevantes para a tarefa tratada. Observa-se que existe uma diferenciação entre a descrição conceitual do domínio do sistema e a descrição da navegação no sistema. Esta separação, ainda segundo Lima (2003), permite maior flexibilidade e o tratamento de características específicas a cada etapa.

5.2.10 ASHDM (Adaptive Semantic Hypermedia Design Method)

O método ASHDM propõe a extensão do método SHDM, com o acréscimo de uma camada de adaptação que permite também o tratamento de questões como meta-adaptação no contexto da descrição das aplicações (Assis, 2005; Assis e Schwabe, 2006). Desta forma, não apenas os modelos da aplicação, mas também as regras e o processo de adaptação podem ser delineados e modificados em função de informações de contexto.

A arquitetura possui como objetivo auxiliar na etapa de projeto de aplicações de Hipermídia Adaptativa, permitindo que sejam tratadas questões de identificação e tratamento de adaptações. O trabalho acrescenta características adaptativas aos modelos de navegação e de interface propostos pelo Método SHDM. Além disso, acrescenta o modelo de usuário e o modelo de adaptação, permitindo o tratamento automático das situações de adaptação necessárias em projetos adaptativos. No trabalho também são descritas diretrizes para a arquitetura de implementação do mesmo.

5.2.11 HERA

A metodologia proposta pelo sistema Hera (Vdovjak, 2003) utiliza uma distinção de modelos para o conteúdo conceitual da aplicação e para os aspectos ligados a descrição no formato de hipermídia. O ambiente designado para a geração dos resultados possibilita que sejam gerados automaticamente dados em formatos diversos, como HTML, WML, SMIL. Para tanto é utilizada uma combinação entre a representação dos dados em RDFS e sua transformação com uso de XSLT.

A metodologia prevê as seguintes etapas: projeto conceitual, projeto de aplicação e projeto de interface. São utilizados meta-modelos descritos em RDFS, sendo que estes permitem a descrição de propriedades e restrições de apresentação, utilizadas para a geração automática do resultado final.

5.2.12 AHAM

O AHAM (DeBra, 1999) se propõe a ser um modelo de referência para sistemas de hipermídia adaptativa e, neste sentido, descreve separadamente o modelo de

domínio, modelo de usuário e modelo de adaptação. No modelo de usuário são relacionadas informações estáveis de um determinado usuário, aspectos de seu ambiente de trabalho e descritas possíveis relações entre este usuário e o seu conhecimento e interesse para com elementos do modelo de domínio. Neste modelo são descritos os conceitos apresentados pela aplicação, junto com atributos que possam ser utilizados em adaptações. O modelo de adaptação trata de relacionar os conceitos descritos no modelo de domínio com opções de tratamento durante o seu acesso. Para esta tarefa são empregadas regras, de forma que possam ser descritas as opções desejadas.

5.3 Análise geral

Após esta descrição sucinta de diversos exemplos de metodologias para a descrição de aplicações Web, são relacionadas a seguir algumas de suas características para que possa ser observada, de forma geral, a maneira pela qual estas se complementam. A tabela 5.1 abaixo indica para cada metodologia uma breve descrição e indicação dos principais recursos empregados, juntamente com o destaque, na última coluna, para a previsão de recursos de adaptação.

A análise destes exemplos indica claramente uma preocupação com o tratamento flexível de diversos aspectos de uma aplicação Web. Assim, são elencados modelos para a descrição da apresentação, navegação, visões conceituais, adequação a formatos de dispositivos e modelos de adaptação. As tecnologias para a descrição destes modelos são bastante diversificadas em um momento inicial. Entretanto, observa-se a existência de diversos trabalhos que foram originalmente implementados sem recursos da Web Semântica e que apresentam versões posteriores incorporando estas características. Um dos motivos desta ampliação ou adequação dos trabalhos está ligado à flexibilidade obtida, tanto para a descrição dos modelos, como para a sua manipulação. Esta pode ser feita de forma mais flexível e com uso de linguagens de consulta específicas ou com mecanismos de inferência.

Tabela 5.1: Quadro comparativo de metodologias para descrição de aplicações Web

Nome	Descrição	Recursos	Ad.
RMM	Objetivo principal: projeto da aplicação e descrição do domínio	Diagrama ER, templates HTML, sem previsão para adaptações	N
OOHDM	Orientado a Objetos. Define requisitos, projeto conceitual, navegacional, interface abstrata e modelo de aplicação	Componentes descritos em notação de classes. Aplicação mapeia elementos HTML	N
SHDM	Expansão do método OOHDM com inclusão de recursos de Web Semântica	Ontologias (OWL)	N
OOWS	Segue metodologia Orientada a Objetos. Define modelo conceitual e modelo de desenvolvimento	Diagramas UML	N
WebML	Descreve projeto de dados, de hipertexto e implementação. Incorpora recursos de Web Semântica	Linguagem de alto nível para modelagem, serializada em XML. Templates JSP	N
WSDM	Define requisitos, públicos, projeto conceitual, design, implementação	Ontologias (OWL). Elementos de interface podem ser associados a públicos diferentes	S
XWMF	Adaptações a dispositivos específicos e validação de dados	RDF e linguagem específica (WOCM). Geração da inerface com Tcl e Prolog	S
SEAL	Diversos modelos (navegação, apresentação, usuários, adaptação)	Utilização de ontologias (RDFS, OWL) e regras semânticas. Framework KAon, Inferência (Triple)	S
ASHDM	Expansão do método SHDM, com previsão explícita de modelo de adaptação	Ontologias (OWL)	S
HERA	Geração de formatos finais diferenciados (HTML, WML, SMIL)	Ontologias (RDFS) e XSLT	S

Resumo do Capítulo:

Neste capítulo são descritas brevemente algumas iniciativas para a descrição de modelos voltados para aplicações Web. Em função de sua crescente importância e de suas peculiaridades, torna-se cada vez mais necessário o uso destas metodologias. Em boa parte das propostas mais atuais observa-se o uso de recursos da Web Semântica.

6 TRABALHOS RELACIONADOS

Neste item são descritos brevemente alguns trabalhos que utilizam recursos de adaptação ou de mineração de uso na Web, sendo ressaltadas as vantagens, desvantagens e as características de interesse para o trabalho aqui apresentado.

Os exemplos são organizados por características similares, permitindo que se possa distinguir os primeiros sistemas, onde não são utilizados expressivamente recursos mais complexos, tais como modelos ou recursos da Web Semântica. Este grupo é exemplificado pelos itens 6.1, 6.2 e 6.3. Além deste grupo é destacado um conjunto de trabalhos no qual existe a utilização mais extensiva de modelos, ilustrado pelos itens 6.4, 6.5 e 6.6. O item 6.7 ilustra um exemplo de trabalho onde modelos específicos são utilizados para o tratamento de adaptações de textos. Os itens 6.8 e 6.9 ilustram sistemas que utilizam extensivamente modelos de domínio, de usuário e de navegação, para a geração dos resultados. Os três últimos exemplos, itens 6.10, 6.11 e 6.12 ilustram sistemas onde existe a integração de resultados de mineração de uso como forma de colaborar na adaptação promovida pelo sistema. O sistema exemplificado no item 6.13 trata da construção de um portal semântico onde o modelo do usuário é alimentado com preferências e interesses indicados pelo próprio usuário, ao ser cadastrado no sistema.

6.1 ELMART

Os sistemas ELM-ART e ELM-ART II (Weber, 1997), são citados como os primeiros sistemas de Hipermídia Adaptativa disponibilizados na Web. Neles o usuário é modelado como uma sequência de eventos de interação com o sistema. Os conceitos disponibilizados para acessos relacionam-se entre si por conjuntos de requisitos. Com o objetivo de estruturar um livro eletrônico, os conceitos são organizados em lições, seções, sub-seções e páginas terminais. Na figura 6.1 são exibidos componentes de um curso, exercícios relacionados e uma área para anotações particulares.

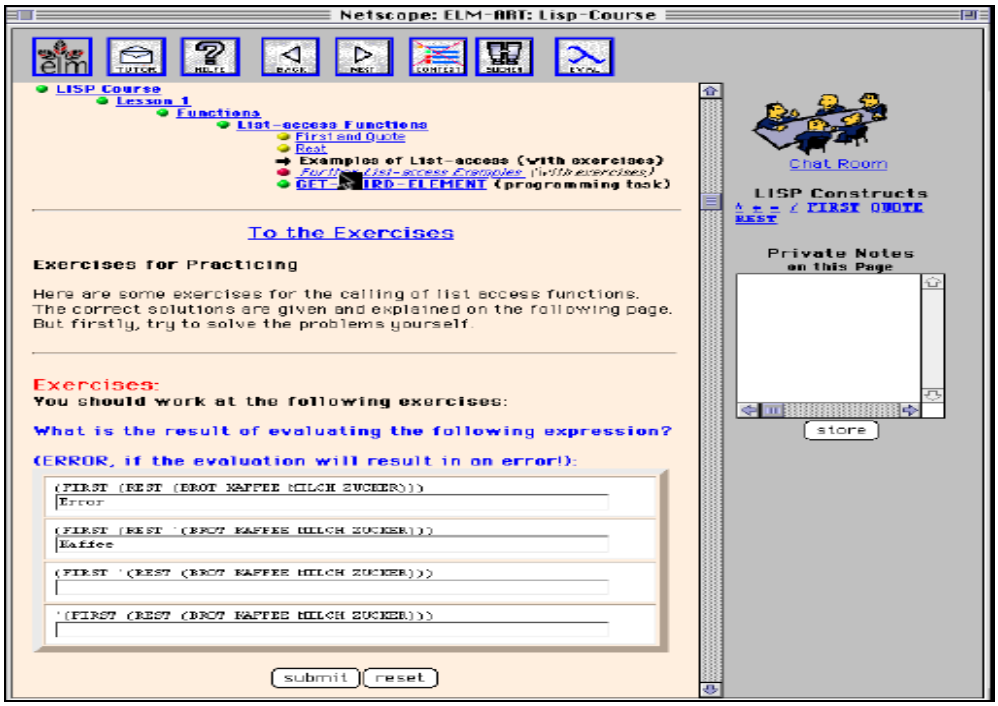


Figura 6.1: Exemplo de conteúdo do sistema ELMART

6.2 PUSH (Plan and User Sensitive Help)

O sistema PUSH (Espinoza, 1996) apresenta opções de apoio em tarefas de busca de informações. Permite ao usuário diferentes níveis de interatividade e faz uso de regras simples para associação de perguntas com possíveis tarefas associadas. A arquitetura do sistema, descrita na figura 6.2, prevê a geração de imagens e de conteúdo textual. São descritos também componentes para o tratamento do modelo do usuário e uma base de conhecimentos, contendo as regras de adaptação.

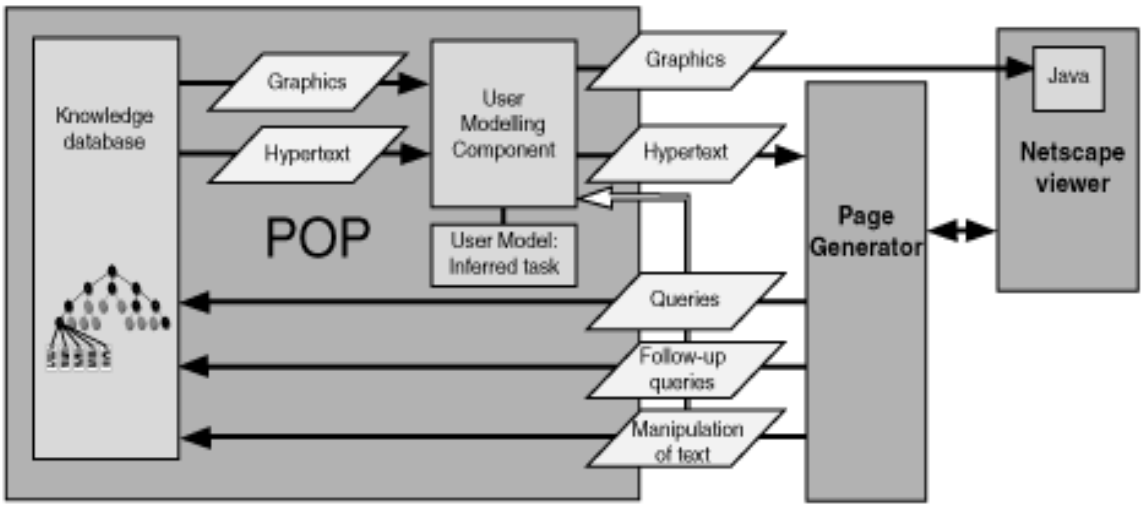


Figura 6.2: Arquitetura do sistema PUSH

6.3 AVANTI

O sistema AVANTI (Nill, 1995) tem como foco a adaptação de informações sobre área metropolitana para usuários com diferentes necessidades. Utiliza modelos de usuário e de conteúdos, com apoio de um componente de adaptação. Propõe o atendimento de usuários com deficiências, a partir de diferentes dispositivos de interação.

Na figura 6.3, que descreve a arquitetura do sistema, podem ser observadas algumas de suas características principais, como a adaptação voltada para diferentes dispositivos e também o uso de regras para adaptação de conteúdos apresentados.

O sistema prevê a utilização de características de sistemas adaptáveis e também de sistemas adaptativos, justamente para facilitar o atendimento a usuários com necessidades específicas. Este tipo de interação pode ser realizado a partir de um componente específico.

O mecanismo de adaptação utiliza bases de conhecimentos para o tratamento das características dos usuários e estilos de interação, para adaptação sintática e léxica. Também está prevista na arquitetura o monitoramento das ações e sua utilização junto aos demais componentes.

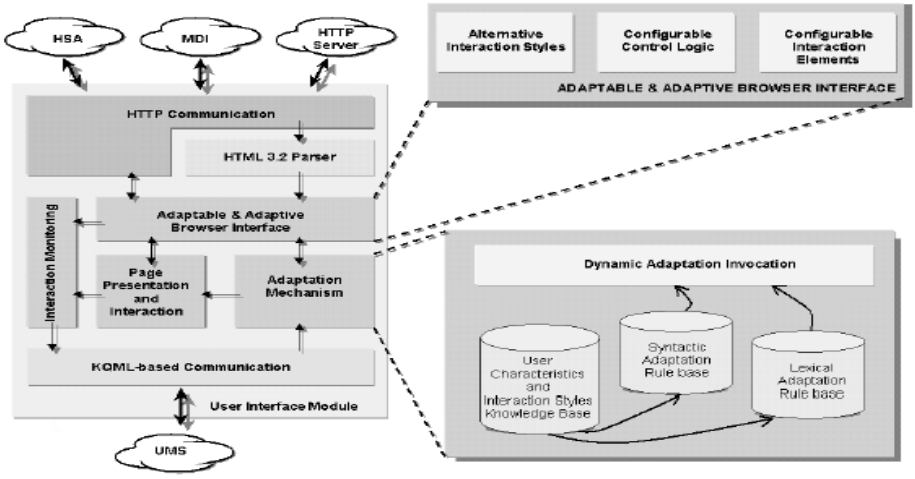


Figura 6.3: Arquitetura do sistema Avanti

6.4 Elena PLA (Personal Learning Assistant Service)

O sistema Elena PLA (Dolog, 2004) possui como foco principal o suporte para personalização em um contexto de ambiente distribuído, usando uma arquitetura baseada em serviços (com uso de *Web-Services*). Alguns dos componentes importantes são o modelo do usuário, repositórios de conteúdos, serviços de anotação e de recomendação. Estes componentes e a proposta geral do sistema estão ilustrados na figura 6.4 abaixo.

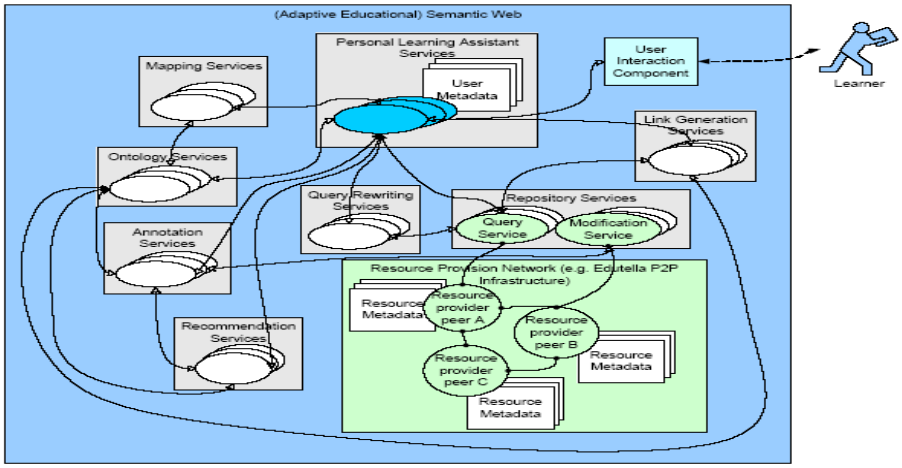


Figura 6.4: Arquitetura baseada em serviços do sistema PLA

Os resultados da adaptação podem ser vistos na figura 6.5 a seguir. Parte deles consiste em recomendações sobre conteúdos, o que pode ser notado na figura em duas colunas situadas à esquerda da tabela de resultados. As cores verde, vermelho ou amarelo indicam os itens recomendados ou não.

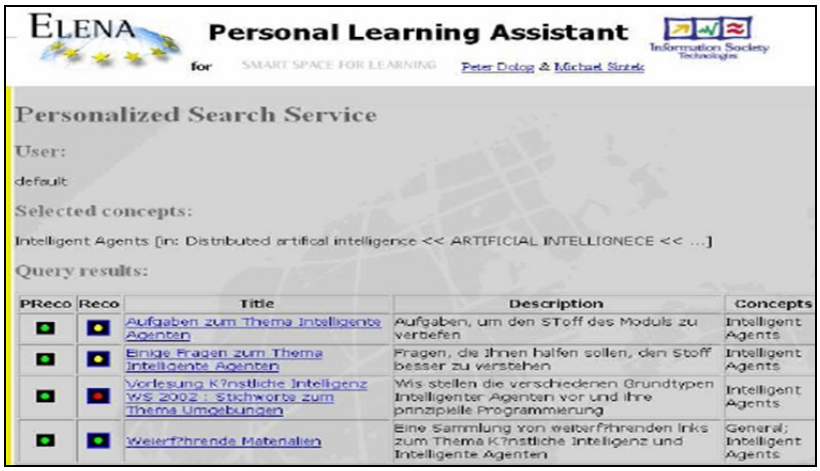


Figura 6.5: Exemplos de recomendação

6.5 AdaptWeb

O ambiente AdaptWeb (Freitas et al., 2002; Oliveira e Muñoz, 2004) disponibiliza um ambiente adaptativo para aprendizagem, onde são levadas em conta o programa de aprendizado, o conhecimento do aluno e suas preferências de interação (navegação). Trata-se de um exemplo no qual pode-se observar a utilização de recursos como a anotação semântica, a descrição de estrutura de conteúdos e a adaptação de características dos conteúdos, de acordo com o contexto do usuário. Um ambiente de autoria permite a descrição de metadados associados ao material sendo publicado. Uma descrição do modelo do usuário é utilizada, a partir de uma ontologia, prevendo operações de personalização. Estas são realizadas levando-se em conta a estrutura dos conteúdos e o modelo de usuários. Os conteúdos educativos são descritos

com o uso de padrões que possibilitam seu compartilhamento. A visão geral da arquitetura do sistema e de seus componentes está ilustrada na figura 6.6 abaixo.

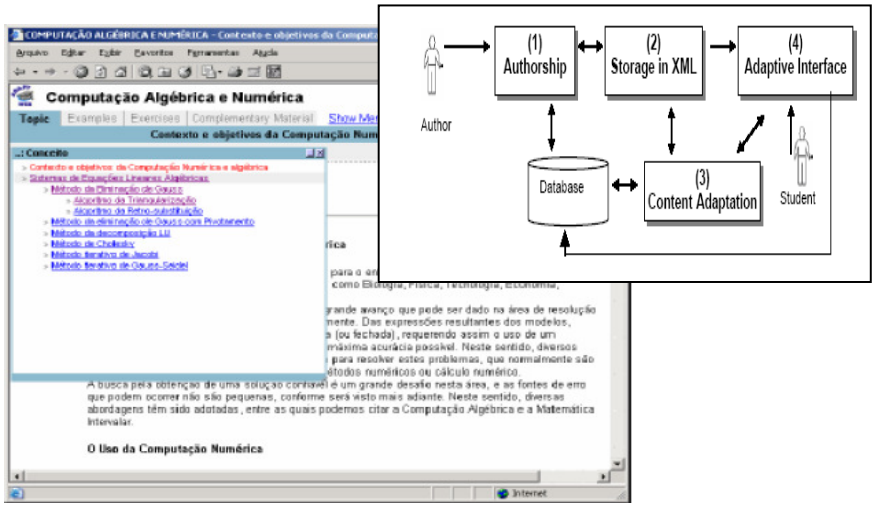


Figura 6.6: Adaptações e metadados no ambiente Adaptweb

Este Ambiente permite a autoria e apresentação adaptativa de disciplinas integrantes de cursos na modalidade EAD na Web. Na figura 6.7 está apresentada a interface de autoria, para uma situação típica. O objetivo do AdaptWeb é permitir a adequação de táticas e formas de apresentação de conteúdos para alunos de diferentes cursos de graduação e com diferentes estilos de aprendizagem, possibilitando diferentes formas de apresentação de cada conteúdo, de forma adequada a cada curso e às preferências individuais dos alunos participantes. O projeto é baseado na linguagem PHP[54] e no banco de dados MySQL[55], usando as tecnologias de Web Semântica e ontologias para prover adaptabilidade e interoperabilidade.

---

54 http:// www.php.net

55 http:// www.mysql.com/

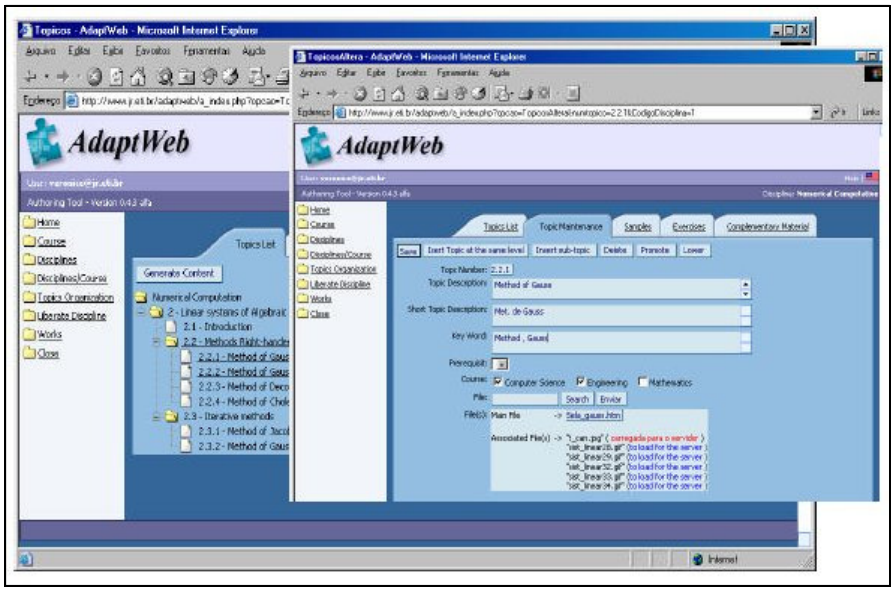


Figura 6.7: Ferramenta de autoria do ambiente AdaptWeb

6.6 AHA!

O sistema AHA: Adaptive Hypermedia Architecture (De Bra, 2003) permite adaptação de texto e da estruturas de hyperlinks. Utiliza modelo de usuários, de domínio e de adaptação. A aplicação é composta por conceitos e relações, apresentados aos usuários de acordo com suas características. A figura 6.8 ilustra sua arquitetura, onde podem ser identificados os modelos utilizados e também a existência de opções de autoria e de gerenciamento do ambiente.

Este sistema vem sendo melhorado constantemente, servindo como base para um conjunto interessante de experiências com diferentes aspectos da Hipermídia Adaptativa. Assim, a partir deste trabalho inicial, são conhecidos outros trabalhos derivados, que exploram a utilização de modelos para a descrição da aplicação, a aquisição de dados de usuários para a geração do perfil de usuários e opções de descrição de adaptação, como com a utilização de linguagens específicas (De Bra et al.; 2007, Romero et al., 2007; Stash et al., 2007).

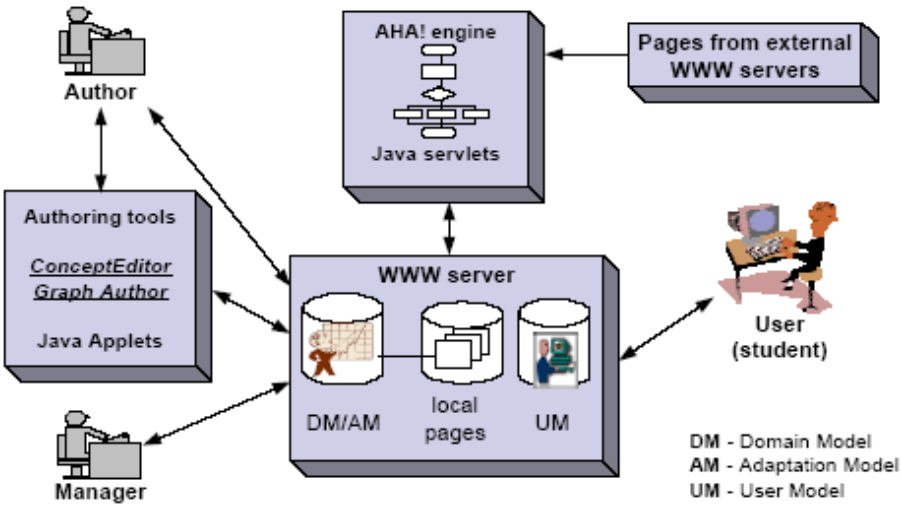


Figura 6.8: Arquitetura do sistema AHA!

6.7 Hylite

O sistema Hylite+ (Wilks, 2005) apresenta características importantes de adaptação de conteúdo. No sistema são manuseadas informações técnicas sobre equipamentos de informática para que o usuário possa receber um resultado com maior ou menor número de detalhes. As informações são manipuladas com base em uma representação prévia com RST (*Rethorical Structure Tool*). Na figura 6.9 é exemplificada uma consulta para um termo ("*tape driver*") e a exibição dos resultados. A informação de subtipos apresentada (DLT, DAT, Ultrium) pode ser suprimida automaticamente, por exemplo, a partir da identificação do interesse do usuário.

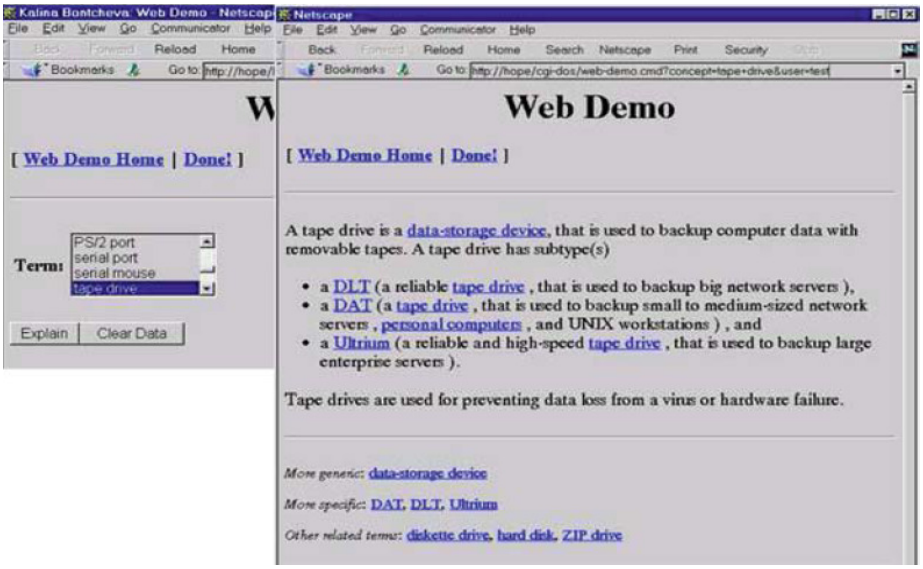


Figura 6.9: Resultado de adaptação de texto no systema Hylite+

6.8 OntoWeaver

Descrito por Lei (2003, 2004, 2005), o OntoWeaver é uma proposta de portal de conhecimento. Possui tratamento de modelo de usuários, modelo de domínio, modelo do site e modelo de apresentação, conforme indicado na figura 6.10, que descreve sua arquitetura. Sugestão de uso de *Web-Services* para colaboração e integração de serviços. A customização da visualização é feita a partir de regras de adaptação, de acordo com o usuário.

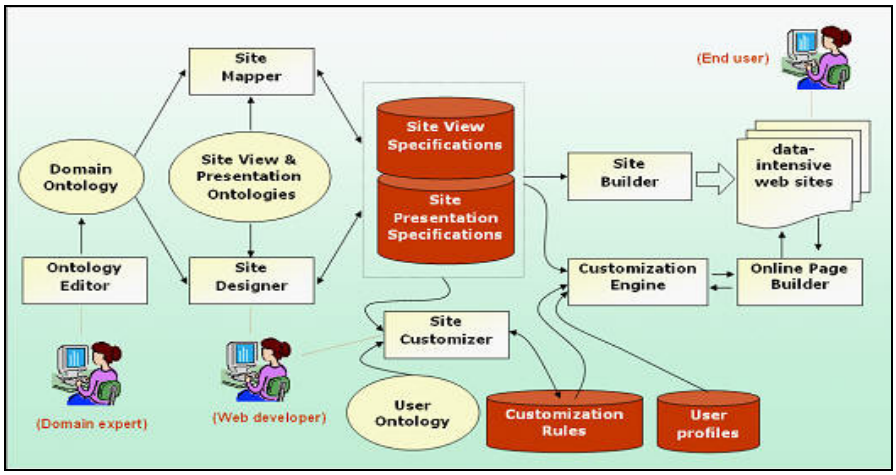


Figura 6.10: Arquitetura do sistema OntoWeaver

Neste sistema destaca-se a utilização extensiva de ontologias para a descrição do modelo do *site* Web e para o auxílio nas tarefas de adaptação. Na figura 6.11 estão descritos alguns destes componentes utilizados no projeto para a descrição de informações sobre o domínio, sobre a navegação e apresentação. Seguem alguns comentários sobre os elementos desta figura. Nos dois itens na parte superior esquerda (marcados como *(a)* e *(b)*) pode ser observada a estrutura de classes considerada para os elementos de conteúdo publicados no portal, assim como o seu relacionamento. Na parte superior direita está um resumo dos elementos considerados na ontologia que descreve a estrutura de publicação de um portal desenvolvido com esta abordagem. Na parte inferior esquerda pode ser visto o resultado da publicação do portal, tal como observado por um usuário que acessa o mesmo a partir de um navegador Web. Por fim, na parte inferior direita são exibidos os relacionamentos entre as seções que compõe o portal. Estas relações podem ser relações descritas na construção do portal, ou relações contextuais, detectadas a partir da estutura de relações definida.

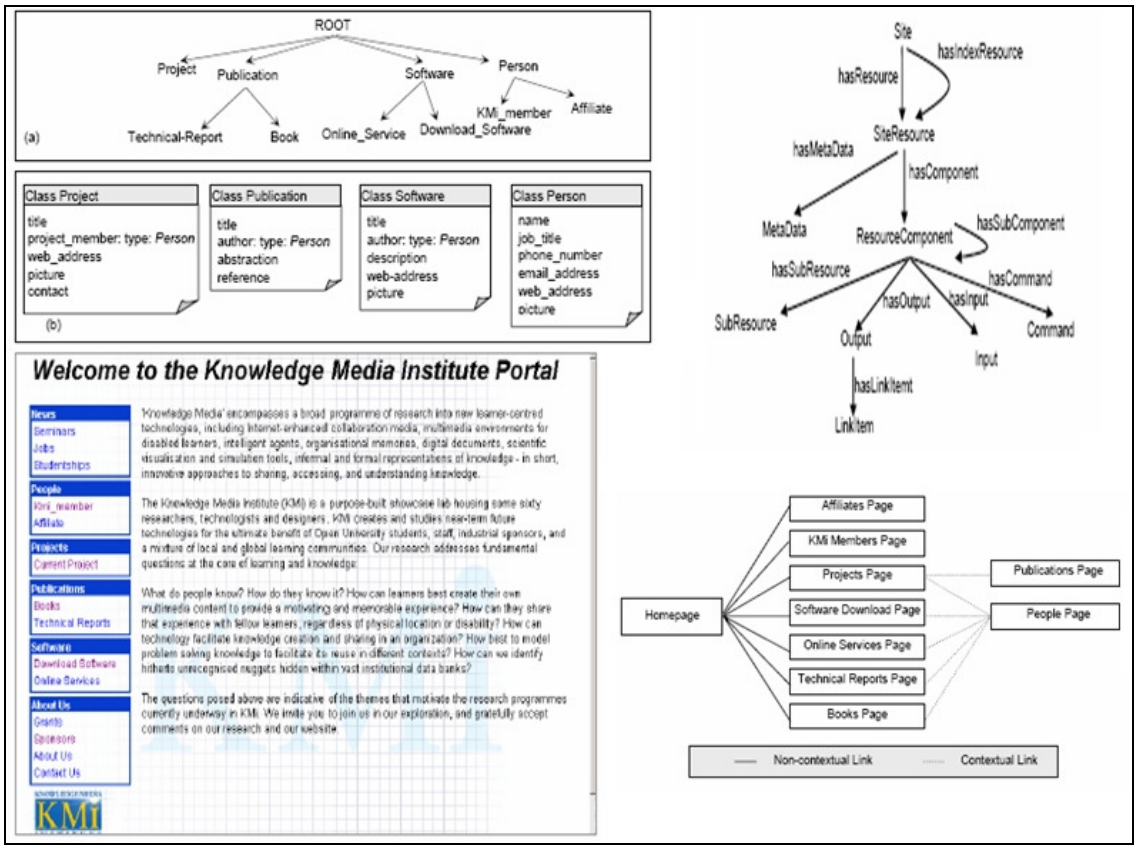


Figura 6.11: Ontologia e resultados no sistema Ontoweaver

6.9 CXMS: Context Management Framework

O sistema *CXMS* (Zimmerman, 2005) se propõe a realizar o tratamento de informação de contexto e interação com dispositivos móveis. Existe a integração de modelos de usuário e de conteúdo com informações adicionais originadas em dispositivos externos, tais como sensores. Assim observa-se a interação do sistema, durante etapas de adaptação, com informações mais diretas. Também são previstas no ambiente anotações de conteúdo, como forma de facilitar ao autor dos conteúdos a identificação de situações de adaptação.

Na figura 6.12 abaixo é descrita a arquitetura deste sistema, na qual pode ser observado que o mesmo prevê a geração de resultados para diferentes dispositivos, bem como a entrada de dados originados de sensores. A figura evidencia também o cuidado com o tratamento de contextos, sendo que neste caso a descrição e gerenciamento dos contextos é considerada de forma relacionada com a própria aplicação e não apenas com detalhes de componentes de software.

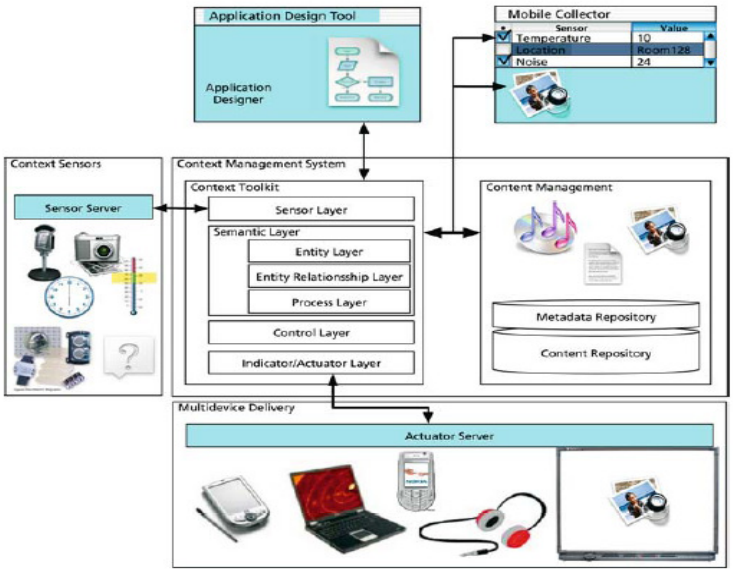


Figura 6.12: Arquitetura do sistema CXMS

A anotação de conteúdo e seu relacionamento com contextos diversos é uma tarefa que exige recursos dedicados. Sendo assim, existe um componente neste sistema que foi dedicado à tarefa de associação de conteúdos e contextos. Pode ser observado na figura 6.13 uma tela exibindo parcialmente este componente do sistema. Na parte esquerda da figura está disposta a informação contextual, que pode ser indicada como relevante para a exibição dos elementos de conteúdo que estão dispostos na parte direita da figura.

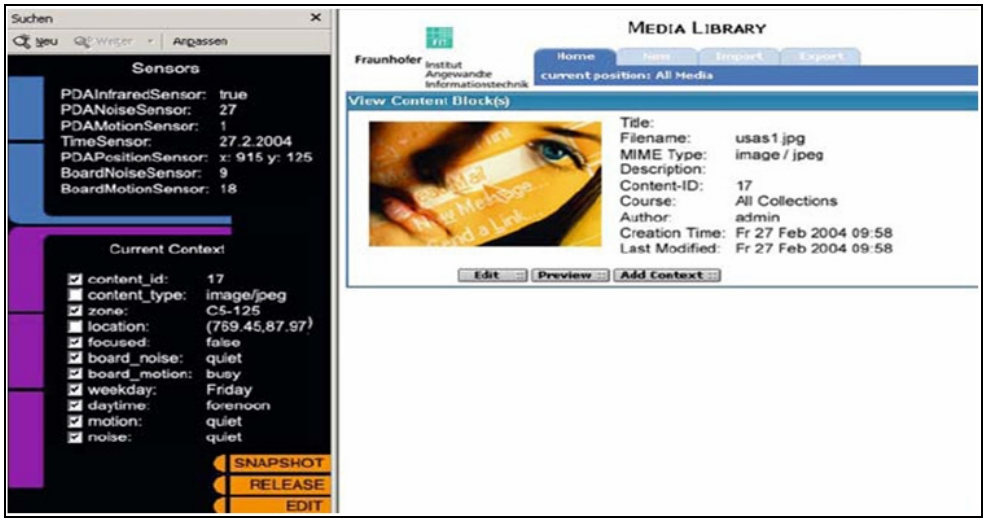


Figura 6.13: Exemplo de anotação de contextos no sistema CXMS

6.10 Framework para mineração de uso da Web

Um Framework para adaptação de site Web com aplicação de técnicas como mineração de uso é apresentado por Mikroyannidis (2005). A aplicação de Mineração de Uso Web para geração de sugestões de modificação da estrutura de navegação é complementada com o uso de ontologia com estrutura do *site* Web. Na figura 6.14 está

ilustrado o processo geral adotado neste sistema e uma ontologia de domínio que foi utilizada para a descrição de um *site* Web a ser adaptado.

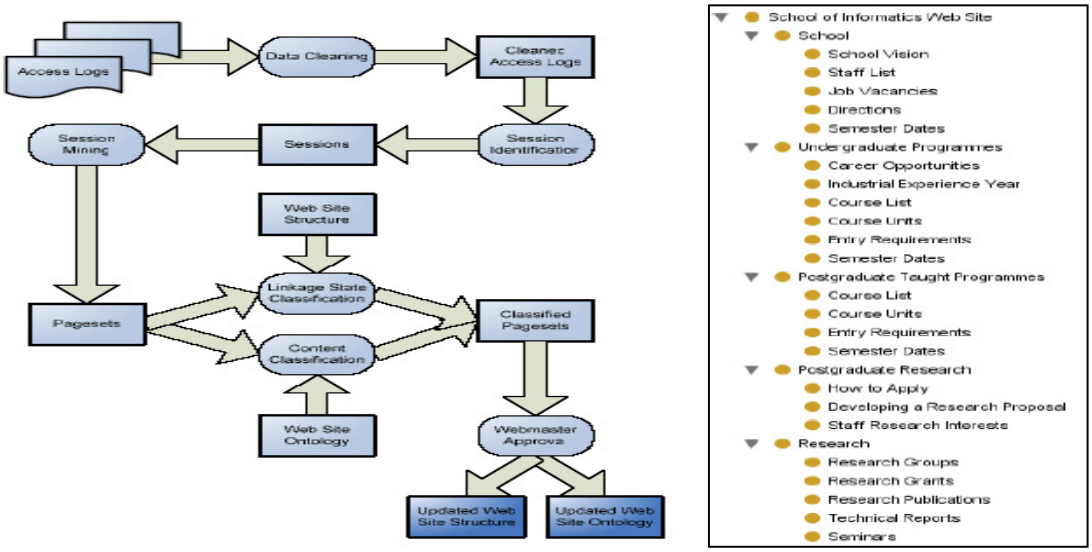


Figura 6.14: Arquitetura e ontologia

Neste sistema os dados de uso são empregados para a geração de percursos freqüentes. Estes, por sua vez são integrados como informações descritas na ontologia da aplicação, em duas formas complementares. A primeira leva em conta a estrutura de relacionamentos das páginas de um percurso freqüente. A segunda leva em conta os conteúdos associados a cada uma das páginas. Após estes mapeamentos as informações obtidas são disponibilizadas como adaptações no *site* Web. O trabalho prevê também a possibilidade de alteração permanente da estrutura do *site*, a partir das informações de mineração obtidas. Esta alteração permanente é feita em um processo manual.

6.11 SEWEP

O sistema SEWEP (Eirinaki, 2003) destaca-se pelo uso de mineração de textos como apoio na construção do modelo de conteúdo, utilização de mineração de uso em conjunto com informações do modelo de conteúdo para a geração de recomendações. Na figura 6.15 é apresentada a arquitetura do SEWEP, um sistema de personalização Web que integra processos de análise de uso com semântica como forma de enriquecer o conjunto de recomendações que são providas para o usuário.

Uma característica da arquitetura desse sistema é o uso de "C-*logs*", uma extensão dos logs de uso da Web, que contém a semântica do conteúdo. A anotação semântica do conteúdo é executada usando uma hierarquia, permitindo assim uma série de recomendações ao usuário. Este trabalho envolve o uso de dados do perfil do usuário, levando em consideração suas preferências, para filtrar as recomendações. Por fim, é previsto no sistema o uso de um algoritmo para geração de regras de associação para aumentar o poder de recomendação do SEWEP.

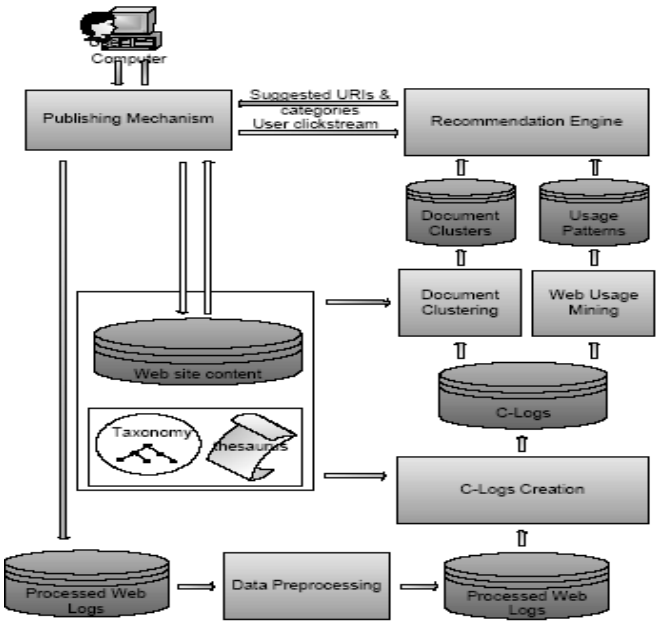


Figura 6.15: Arquitetura do sistema Sewep

6.12 Suggest

O sistema Suggest é um exemplo de sistema que permite a personalização de *sites* Web tomando como base a mineração de uso (Baraglia e Silvestri, 2004; Baraglia e Palmerini, 2002; Baraglia e Silvestri, 2007). O formato de personalização possível é a geração de sugestões de hiperlinks de navegação, com base na identificação de grupos de páginas altamente relacionadas, que são avaliadas quanto à similaridade para com a sessão de um usuário. A figura 6.16 mostra um exemplo do resultado das sugestões de navegação geradas. Elas estão identificadas na janela com o título "*suggestion*", composta por uma lista de hiperlinks, posicionada na parte superior esquerda da figura.

Figura 6.16: Resultado das adaptações do sistema Suggest

O sistema de mineração e de geração das recomendações pode ser tratado de forma independente em relação aos *sites* Web que serão atendidos, sendo integrado ao servidor Web Apache[56]. Ele possui como características principais o fato de empregar uma arquitetura interna que permite a geração das sugestões sem a necessidade de identificação pessoal dos usuários e também o fato de realizar a coleta e o tratamento dos dados de forma automática e ininterruptamente. Grande parte dos sistemas de personalização tratam a coleta dos dados em etapas separadas, realizadas periodicamente.

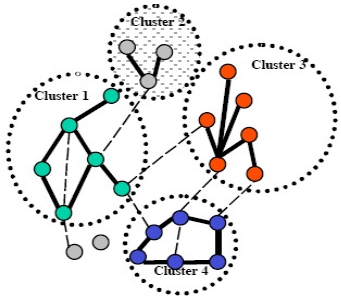


Figura 6.17: Geração de agrupamentos pelo Suggest

A figura 6.17 acima exemplifica o resultado processo de geração dos agrupamentos neste sistema. O sistema mantém uma representação das páginas acessadas em forma de grafo, onde os nodos indicam cada página e os arcos indicam o relacionamento entre elas, sendo que este relacionamento é evidenciado durante a navegação entre as duas páginas conectadas por um arco. De acordo com a maior quantidade de acessos, os arcos são considerados mais significativos, formando assim os agrupamentos que serão usados na geração de sugestões.

6.13 SEMPort

O sistema SEMPort implementa um portal semântico, com opções de personalização. Um dos principais objetivos do sistema é proporcionar ao usuário uma navegação baseada em conceitos e não apenas em hiperlinks (Sah e Hall, 2007, Sah et al., 2008). O usuário indica seus interesses e preferências explicitamente, durante o cadastro ou utilização. O sistema é implementado com uma arquitetura baseada em ontologias para a descrição do domínio de conhecimento, do modelo do usuário e seu perfil.

---

[56] http://www.apache.org

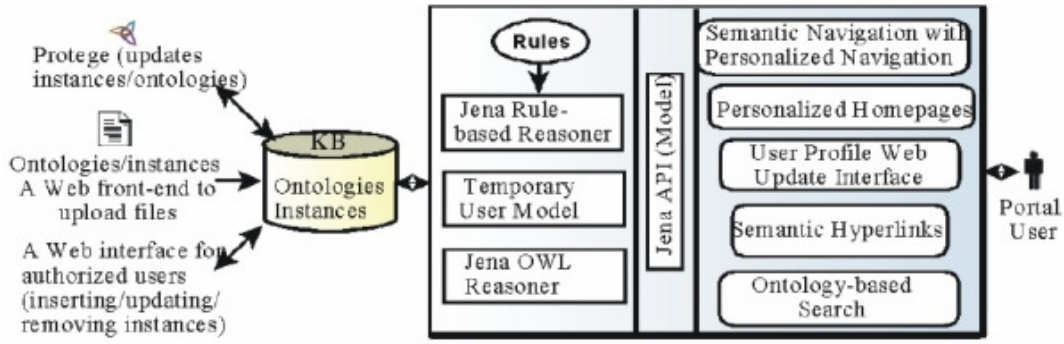


Figura 6.18: Arquitetura do SEMPort

Os conteúdos são inseridos no sistema em uma interface simplificada, gerando instâncias na ontologia. A figura 6.18 acima ilustra a arquitetura do SEMPort, onde podem ser destacados os seguintes componentes: base de conhecimentos, contendo as instâncias das ontologias de domínio; modelo do usuário, contendo suas preferências; módulo de navegação personalizada; módulo de busca na ontologia.

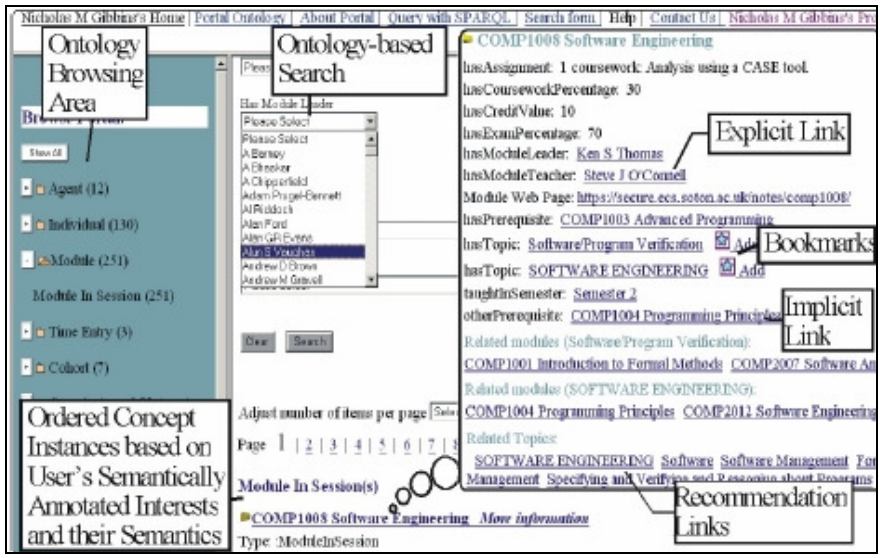


Figura 6.19: Exemplo de interface gerada pelo SEMPort

Na figura 6.19 está exemplificada uma sessão típica de navegação. Na parte esquerda da imagem é possível observar a navegação pela ontologia. Isso permite que o usuário tenha acesso aos conteúdos desejados a partir dos conceitos de seu interesse e não a partir de hiperlinks. Já na parte direita da imagem está exibida a janela contendo os resultados detalhados para o item selecionado. Nesta janela são exibidos alguns itens personalizados de acordo com as informações do perfil do usuário mantido pelo sistema.

6.14 Avaliação de características gerais

Esta breve descrição de exemplos de trabalhos relacionados com a tese apresentada possibilita a identificação de alguns pontos de atenção. A partir de uma avaliação dos trabalhos pioneiros na área de Hipermídia Adaptativa já é possível a identificação de um contexto geral onde a utilização de modelos é evidente e contribui para a geração das caracerísticas necessárias aos sistemas, tais como a flexibilidade na

geração dos resultados e o acompanhamento da utilização. Também pode ser traçado um paralelo entre as formas adotadas para a adaptação do sistema e as necessidades de tecnologias. Uma definição de adaptação de conteúdo passa necessariamente pela adoção de tecnologias para a representação mais rica dos documentos, seus conteúdos e formatos. O atendimento a dispositivos diversificados passa pela geração de resultados em múltiplos formatos.

As arquiteturas identificadas nos exemplos representam diversas escolhas na forma de modelagem do perfil dos usuários e na utilização destes resultados. Estas opções apresentam uma clara tendência para a adoção cada vez maior de técnicas automáticas para a geração dos modelos de usuários. Alguns fatores que contribuem para esta adoção podem ser indicados. São eles o grande volume de usuários, a variação freqüente de seus interesses, a utilização de sistemas diversos pelos usuários, com interfaces específicas. Outro fator associado é a possibilidade de geração e atualização de modelos de usuários de forma automática e sem a necessidade de interação com os mesmos.

Muitos trabalhos utilizam de modo bastante amplo estas possibilidades de aquisição automática de informações, ampliando a opção de utilização das informações de uso para a incorporação de informações relacionadas com o conteúdo e formatos de apresentação. Em casos de dispositivos com informações diferenciadas, como os dispositivos móveis, estas informações podem ser observadas como componentes importantes em alguns trabalhos.

Por fim, trabalhos recentes apresentam de modo bastante evidente a utilização de recursos de semântica, em diversas formas, para apoio de atividades de aquisição de modelos de usuários ou para atividades de adaptação. Com esta abordagem podem ser evitadas algumas limitações impostas pelo conjunto de informações de uso, quando não qualificadas e relacionadas.

Resumo do Capítulo:

Neste capítulo são apresentados exemplos de aplicações Web com características de adaptação diversas. Esta diversidade permite a identificação de recursos importantes como modelos descritores do usuário, domínio ou adaptação. São identificados exemplos de sistemas que integram de alguma forma os modelos citados com informações de uso. Também são descritos sistemas sem uso destes recursos, como forma de contextualização da evolução observada.

7 ARQUITETURA GERAL E EXPERIMENTOS PARA O SISTEMA DESENVOLVIDO

Neste capítulo é apresentada a arquitetura geral para o sistema desenvolvido neste trabalho. Também são descritos experimentos realizados com o objetivo de validação da proposta.

7.1 Considerações iniciais

Considera-se que a validação da proposta para aquisição de informações do perfil de classes de usuários com a integração de informações de uso e semânticas está associada com três aspectos já mencionados anteriormente no texto. Seriam eles: a experimentação da aquisição e tratamento de dados de uso da Web, a descrição semântica de aplicações Web e a adaptação da estrutura de *sites* Web. Estes são os componentes do contexto geral no qual esta proposta encontra-se inserida. Assim, neste capítulo serão detalhados os trabalhos realizados para o tratamento destes tópicos.

Inicialmente foram encaminhados estudos sobre sistemas de Hipermídia Adaptativa, Mineração do Uso da Web e sobre tecnologias da Web Semântica. Além disso, foram analisados trabalhos relacionados, tanto voltados para a aplicação de recursos da Web Semântica na descrição de aplicações Web como direcionados para a adaptação e personalização em diversos contextos, especialmente aqueles com utilização de recursos de mineração e semântica. Com esta etapa identificou-se um cenário no qual a utilização de mineração de dados desponta como uma técnica bastante adequada para a geração de informações para sistemas de personalização e sistemas adaptativos. A geração e manutenção de modelos de usuários é uma etapa de sistemas de Hipermídia Adaptativa que pode obter grandes benefícios com mineração de dados. Os principais fatores associados são a grande quantidade de dados disponíveis em documentos e em registros de interações na Web. Estes podem ser tratados, de forma automática, com Mineração de Conteúdos e com Mineração de Uso da Web, respectivamente.

Em especial, a Mineração de Uso da Web pode apresentar resultados bastante interessantes quando associada a tarefas de geração e manutenção de perfis de usuários. Diversos trabalhos registram esta abordagem, com objetivos bastante diversos. Alguns deles são: a identificação de interesses profissionais em sistemas de Recuperação de informações (Mrabet, 2007), a obtenção de subsídios para ações de marketing (Spiliopoulou, 1999; Cooley, 1999), a predição de padrões de acesso futuros (Nasraoui, 2000; Shahabi, 1997), a geração de buscas contextuais (Bamshad, 2004), o atendimento a pessoas com necessidades visuais especiais ou com uso de dispositivos específicos (Zimmerman, 2005).

As aplicações de Mineração de Uso da Web para a obtenção de perfis de usuários podem utilizar diversas opções de algoritmos, tais como agrupamentos (Wang, 1999), classificação (Deshpande, 2003), regras de associação (Agrawal e Srikant., 1994) e percursos freqüentes (Agrawal e Srikant, 1996; Han et al., 2004, Yang et al. 2005; Xin et al. 2005). Cada um possui maior adequação a determinadas situações, tal como o uso de agrupamentos na descoberta de perfis levando em conta conteúdos de documentos acessados ou a mineração de percursos freqüentes no caso da geração de perfis com base na interação. Entretanto, grande parte dos trabalhos nesta área busca o desenvolvimento de algoritmos eficientes para o tratamento do grande volume de dados disponível em situações características de *sites* Web. As abordagens levando em consideração informações semânticas são mais recentes e em menor número (Mei et al., 2007).

Com base nestas constatações, foi definido o processo descrito a seguir.

Para tratar a etapa inicial, de aquisição de dados de uso, foi definido um processo geral de acompanhamento de uso da Web. Neste processo, a aquisição dos dados de uso é realizada com a utilização de codificação específica associada a aplicações Web. Esta codificação possui como objetivo a geração e manipulação de informações armazenadas em *cookies*. Desta forma não existe a necessidade de acesso aos dados de registros de uso (*logs*) de servidores Web, o que facilita algumas etapas de pré-processamento, como a identificação de sessões de usuários. O armazenamento dos dados é realiziado em formato XML, facilitando a sua posterior manipulação. A partir desta forma de aquisição são realizadas as etapas necessárias para o tratamento e pré-processamento dos dados de uso.

O mecanismo de aquisição de dados de uso foi utilizado em um contexto controlado e também em outros contextos mais amplos, relacionados com um *site* pessoal e com diversos *sites* com maior volume de acessos. O objetivo desta abordagem foi testar a aquisição de dados de uso em situações diferenciadas e também testar a abordagem para verificar características específicas de contextos diversos. Outro objetivo desta abordagem está relacionado com a independência do mecanismo implementado em relação às aplicações Web. O mesmo consiste em componente que pode ser facilmente adaptado à diferentes contextos. Como exemplo, foi empregado em situações diferentes, como uma aplicação Web simples, um gerenciador de conteúdo Web (Joomla) e uma aplicação Web como arquitetura baseada na Web semântica.

O armazenamento dos dados em formato XML possibilita que estes sejam facilmente manipulados e com isso aproveitados em diferentes situações. Uma possibilidade é a manipulação dos dados de uso para sua utilização com diferentes algoritmos ou sistemas. No trabalho desenvolvido foram testadas possibilidades de pré-processamento destes dados para a integração com sistemas existentes de mineração de dados. Desta forma é viável utilizar mesmo conjunto de dados com diferentes algoritmos, tais como algoritmos para geração de agrupamentos, regras de associação e percursos freqüentes. O sistema de mineração de dados utilizado para testes de aplicação foi o Weka[57], sendo que detalhes destas aplicações são descritos adiante. Entretanto com a utilização de sistemas de mineração já existentes não é possível a utilização de novas características. Assim, foi implementado um algoritmo para tratamento de seqüências freqüentes, para que fosse possível maior controle na

57 http://www.cs.waikato.ac.nz/~ml/weka/

experimentação da integração de recursos semânticos com os padrões obtidos com mineração do Uso da Web.

As metodologias para descrição de aplicações Web foram estudadas com o objetivo de avaliação das possibilidades existentes. Como a descrição da aplicação Web a partir de um modelo permite que sejam efetuadas operações de adaptação de forma mais adequada e flexível, estes modelos são de interesse para o trabalho desenvolvido. Algumas destas metodologias permitem a descrição de modelos para o domínio da aplicação e para aspectos como modelos de navegação e apresentação de resultados, junto com modelos de usuários. Neste trabalho foram adotadas duas abordagens relacionadas com este aspecto. Na primeira foi realizada a descrição de um modelo de domínio da aplicação a partir de uma ontologia de domínio com foco na área de desenvolvimento escolhida para testes (área educacional). Neste modelo do domínio são descritos os principais conceitos e relações envolvidas. Como este modelo foi descrito em uma ontologia independente da aplicação Web, esta ontologia foi utilizada para a anotação semântica dos conteúdos publicados. Esta abordagem se justifica pelo fato do foco do trabalho estar inicialmente ligado ao mecanismo de aquisição dos dados para descrição de padrões de navegação. Outro fator importante nesta opção está relacionado com a flexibilidade e independência de aplicações, de modo a permitir que o processo de anotação semântica de conteúdos e descrição semântica do domínio da aplicação possa ser livremente portado entre plataformas distintas. Já na segunda abordagem, foi descrita e implementada uma aplicação de gerenciamento de conteúdo Web baseada em recursos da Web Semântica. São empregadas duas ontologias, uma contendo a descrição da aplicação e outra contendo as informações de apresentação necessárias para a geração dos resultados. Nesta abordagem considera-se que a descrição da aplicação e apresentação com recursos da Web Semântica possibilitam mais flexibilidade nas operações de adaptação.

A seguir são descritas em detalhes as etapas envolvidas nestas duas abordagens e também ilustradas experimentações realizadas.

7.2 Integração de informações de uso e informações semânticas

Para evitar a geração de adaptações baseadas apenas no uso do *site* Web, definiu-se um processo de integração com informações semânticas. Para isso parte-se da coleta e tratamento de informações de uso da Web e da anotação semântica dos documentos com base em uma ontologia de domínio. Os dados de uso empregados são os padrões seqüenciais de acesso, que indicam os percursos mais freqüentes realizados pelos usuários do *site* Web. Com o objetivo de identificar possíveis conjuntos de tarefas sendo realizadas no *site* Web foram escolhidas relações que descrevem sua estrutura e outras relações complementares. Estas relações são dependentes do contexto geral do *site*, sendo consideradas no caso da aplicação na área da educação, por exemplo, indicações de precedência entre elementos, requisitos para compreensão e aproveitamento, ou ainda o tipo de conteúdo.

A abordagem baseada em semântica permite que as adaptações levem em conta informações que não estão disponíveis nos padrões de acesso, como as relações indicadas acima (precedência, tipo de conteúdo). Desta forma o contexto tratado é mais significativo, possibilitando melhores resultados. Um exemplo desta possibilidade é descrito a seguir, com a análise de um serviço de adaptação para o caso de *sites* Web

voltados para a descrição e comércio de produtos. A figura 7.1 a seguir ilustra, na parte esquerda, uma seqüência de páginas Web obtidas por um determinado usuário em uma típica sessão, onde o mesmo buscaria por um produto e seguiria acessando o *site* até encontrar detalhes do mesmo. Na primeira página é exibido o conjunto de todas as ofertas de um determinado tipo de produto. Na segunda um produto escolhido foi exibido com um conjunto simplificado de informações. Na terceira página são exibidos todos os detalhes disponíveis sobre o produto. Esta seqüência de acessos pode ser acompanhada por um mecanismo de Mineração do Uso da Web, mas o resultado, conforme já comentado antes, será mais proveitoso para o processo de adaptação caso seja relacionado com a informação semântica associada às páginas.

No exemplo da figura 7.1 está ilustrada a associação possível entre a primeira página acessada e um conceito descrito em uma ontologia de domínio ("classe_produto"). Também a segunda página está associada a outro conceito da ontologia ("LCD"). Este conceito possui uma relação ("is_a") com o conceito anterior, indicando hiponímia. Desta forma, Pode ser detectada uma seqüência de navegação de um conceito mais geral para uma ocorrência deste conceito. O acesso à última página do exemplo pode estar associado com a operação de consulta aos atributos desta ocorrência do conceito mais específico. A informação de uso pode ser utilizada desta forma para que sejam obtidos relacionamentos mais abrangentes do que aqueles possíveis apenas com os resultados de mineração de percursos freqüentes. As relações entre as páginas selecionadas pelos usuários podem ser obtidas consultando-se a ontologia de domínio. Este processo permite constatar, por exemplo, a navegação em profundidade para um determinado assunto ou a navegação em tópicos mais gerais do *site* Web.

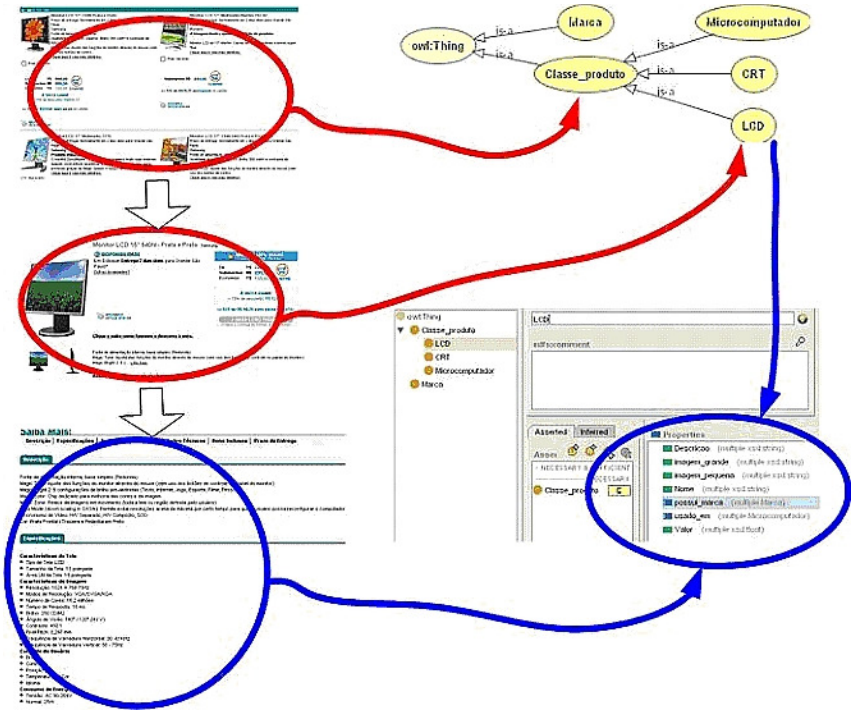


Figura 7.1: Associação entre dados de acesso e conceitos de ontologia de domínio

O objetivo deste exemplo é ilustrar a possibilidade de utilização deste processo em diversas situações, identificando relações específicas em seqüências de acesso que podem ser obtidas pelos dados de sessões de usuários. Com isso podem ser geradas

possibilidades de adaptação mais interessantes. Também pode ser visualizada a flexibilidade da forma de prototipação proposta em relação a outras aplicações Web existentes, pois a coleta de dados e a geração de adaptações podem ser integradas com facilidade ao contexto das aplicações.

A seguir, na figura 7.2 está ilustrado de forma geral o processo de integração implementado. Nela podem ser observados os elementos gerais considerados neste trabalho, tais como a aquisição dos dados de acesso (*log* de acessos), a descrição de informações semânticas (ontologia de domínio), o pré-processamento dos dados de acesso e a descoberta de padrões, além do serviço de adaptação e sua integração com a aplicação Web. A etapa de pré-processamento recebe também informações da ontologia de domínio, além das informações do registro de acessos (*log*). Com estas duas fontes de informações é possível a identificação não apenas da seqüência de páginas acessadas, mas dos conceitos em cada uma e sua relação.

A partir da interação com o usuário, a aplicação Web coleta os dados de uso, armazenados no componente "log de acessos" da figura 7.2. Considera-se a existência de uma ontologia de domínio para a aplicação em questão. Esta ontologia foi descrita manualmente nos experimentos realizados, com uso do software Protégé[58], sendo utilizado o formato OWL para sua representação. O pré-processamento é realizado de forma a integrar estas informações e possibilitar a etapa seguinte, de descoberta de padrões, sendo que os padrões são tratados de forma automática. Após sua geração, os padrões ficam disponíveis para uso pelo componente de adaptação, que interage com a aplicação Web, relacionando as informações do usuário, com o objetivo de geração de adaptações de estrutura das páginas Web disponibilizadas.

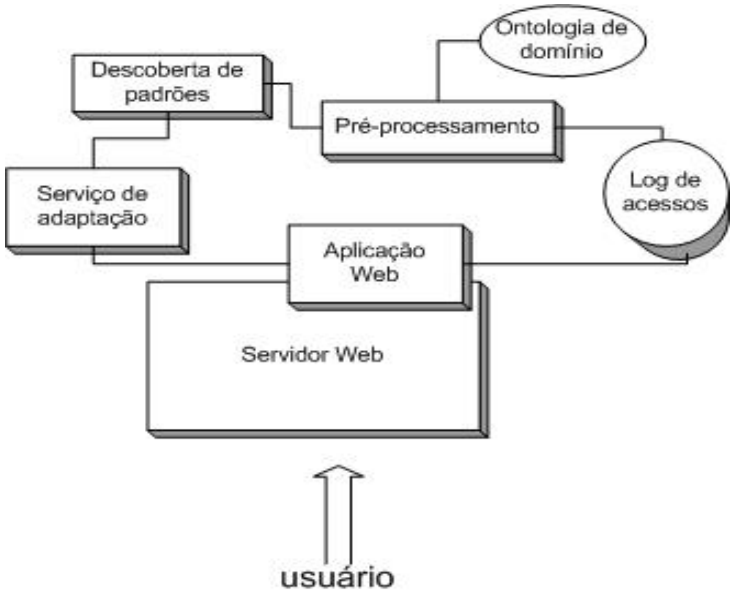


Figura 7.2: Integração de informações semânticas com Mineração do Uso da Web

São descritos no item 7.4 experimentos com a adaptação de um sistema de gerenciamento de conteúdo Web. O experimento implementa a coleta dos dados de uso e também a geração das adaptações de estrutura. No item 7.5 está descrita uma

[58] http://protege.stanford.edu

abordagem complementar, na qual foi implementada uma aplicação Web baseada em informações semânticas para a estruturação do conteúdo de um *site* Web.

De modo geral, a manipulação de perfis de usuários para aplicações de adaptação de um *site* Web pode ser realizada a partir de etapas gerais tais como: identificação de um conjunto padrão de perfis que represente interesses dos usuários de um *site* Web; detecção do padrão de acesso de um determinado usuário; realização de medidas de similaridade entre o perfil do usuário e os padrões previamente armazenados; utilização das informações do perfil com maior similaridade para dirigir ações de adaptação; acompanhamento do uso do *site* Web e atualização dos perfis.

A identificação de perfis padrão que representem interesses dos usuários de um site Web é realizada a partir de uma diversidade de informações. Trabalhos recentes indicam um aumento no interesse pela utilização de informações semânticas. Observa-se a obtenção de perfis com agrupamentos de anotações semânticas de páginas Web (Bamshad et al., 2000; Mobasher e Dai, 2004). Em outro trabalho são geradas taxonomias com mineração de conteúdo de documentos Web e os resultados são associados com dados de mineração de uso para a obtenção do perfil de usuários (Eirinaki 2003). Ontologias de domínio descrevendo a estrutura do site Web também podem ser utilizadas em conjunto com informações de uso (Mikroyannidis 2005). Em outras abordagens, como Aroyo et al. (2006), são usadas ontologias para relacionar fatores temporais, espaciais e léxicos, relacionados com a coleção de conteúdos e utilizados na geração de adaptações. Li e Zhong (2006) apresentam um método para mineração de ontologias descrevendo conjuntos de dados como forma de geração automática de modelos de interesses de usuários. Ou então podem ser observadas abordagens baseadas na aquisição de perfis com informações de uso e ontologias descrevendo conteúdos apresentados nas páginas do *site* (Mrabet, 2007).

O trabalho desenvolvido nesta tese parte de alguns pressupostos para delinear a forma de tratamento dos dados de uso e sua integração com informações semânticas. Busca-se a descrição de um mecanismo geral, que possa ser utilizado em diferentes aplicações Web. A identificação ou participação ativa do usuário não deve ser necessária para o processo de aquisição dos perfis e nem para a etapa de adaptação. A detecção e a aplicação dos perfis devem estar definidas a partir de informações semânticas, integradas com as informações de uso. Os perfis são detectados principalmente com as relações entre os documentos Web acessados, objetivando a delimitação de tarefas realizadas. Informações descrevendo conteúdos disponibilizados em documentos Web acessados podem ser empregadas nas adaptações, como relações específicas. São tratadas apenas as informações de curto prazo, geradas em sessões de usuários não identificados, sem que seja necessário o conhecimento das sessões anteriores de cada usuário. Por fim, os perfis detectados são considerados para conjuntos de usuários e não para usuários individualmente.

As informações de uso empregadas, como citadas anteriormente, são padrões de percursos freqüentes. Os padrões obtidos por mineração de uso são submetidos a um limiar arbitrário indicando o suporte do padrão e com isso são selecionados como parte do conjunto de padrões a serem integrados com informações semânticas. A primeira parte desta integração relaciona cada elemento do percurso freqüente com os demais elementos de um percurso, tomando como base a ontologia de domínio para a aplicação. Assim, o percurso que contém apenas a seqüência de indicadores de páginas acessadas é transformado em um perfil candidato, contendo as diversas relações descritas na ontologia para aquela seqüência de páginas. Um exemplo do resultado desta

manipulação é ilustrado na figura 7.3 abaixo, na qual podem ser vistas as informações de um padrão freqüente de acesso no item “a”, que representam uma seqüência de visualização de três páginas Web, com quinhentas e trinta e três ocorrências e composta pelos seguinte identificadores do sistema de gerenciamento de conteúdo Web: 49, 50, 51. Cada um destes identifica uma página publicada no sistema.

a) ID, Nro pageviews, Nro ocorrencias, lista de paginas
2, 3, 533, 49, 50, 51

b) Percurso: 2 IDs: 49, 50 (composto_por_topico) [type - topico]
Percurso: 2 IDs: 50, 51 [type - topico] [parteDe - ID_49] [tipo_de_material - TM_AULA]
[tipo_de_conteudo - linguagem_de_programacao]

Figura 7.3: Comparação entre padrão de acesso e padrão semântico

A partir das informações semânticas obtidas com a consulta a ontologia de domínio é possível obter um resultado como indicado no item “b” da figura 7.3. Nele podem ser observadas as relações detectadas entre as páginas do percurso freqüente. Por exemplo, pode ser observado que as páginas identificadas com os códigos “49” e “50” possuem entre si uma relação “composto_por_topico”, o que indica que o item “49” é composto pelo tópico “50”. Também observa-se que as duas páginas possuem a mesma relação “type” com valor “tópico”, o que indica que estas duas páginas são do mesmo tipo. Já as páginas identificadas com os códigos “50” e “51” possuem diversas relações em comum. As duas compartilham a relação “type” com valor “tópico”, a relação “parteDe” com valor “ID_49”, a relação “tipo_de_material” com valor ”TM_AULA” e a relação “tipo_de_conteudo” com o valor ”linguagem_de_programacao”.

Estas relações possibilitam diferenciar contextos entre os percursos obtidos com a mineração de uso. Por exemplo, na figura 7.3 observa-se que as duas últimas páginas do percurso possuem em comum a relação “parteDe” com o mesmo valor, sendo que este corresponde ao código da primeira página. Como exemplo, a figura 7.4 ilustra um grafo gerado a partir destas informações obtidas com a consulta das relações descritas na ontologia de domínio para os componentes deste percurso freqüente. Nela pode ser identificado, por exemplo, que os itens 50 e 51 são parte do item 49, evidenciando uma situação de composição. Tambémpode ser verificado que estes dois itens (50 e 51) possuem os mesmos valores para as relações descrevendo o tipo de material e o tipo de conteúdo.

Desta forma o contexto semântico pode ser empregado para diferenciar os percursos entre si e também para a geração de subsídios para adaptação. Além disso, é possível a identificação de similaridades entre os percursos, de modo a gerar um conjunto mais resumido de perfis semânticos de acesso. No caso de um percurso que tenha início em uma página geral e continue percorrendo todas as páginas componentes desta, o tipo de acesso é similar ao acesso exemplificado nas figuras 7.3 e 7.4 e pode ser assim agrupado.

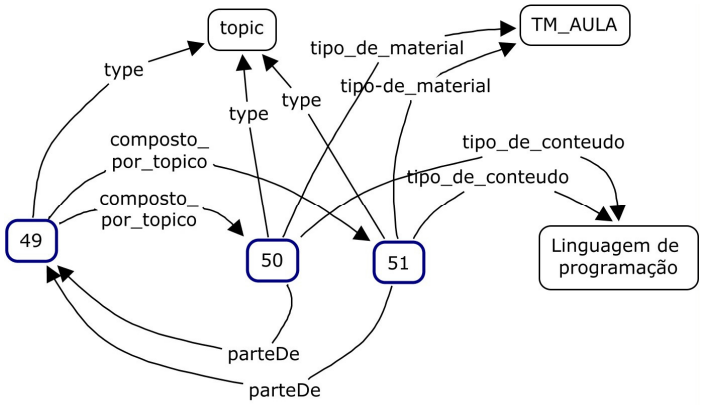


Figura 7.4: Grafo com ilustração de um contexto semântico

As relações empregadas na geração do contexto semântico estão descritas na ontologia de domínio e podem ser adequadas a cada tipo de aplicação. O processo definido para a geração do contexto semântico é independente de relações específicas. Para cada percurso representado por um conjunto de páginas L = {l_1, l_2, l_3, ..., l_m} são verificadas as relações entre os elementos de cada par de páginas (l_i, l_{i+1}) e também as relações destes dois elementos com algum outro conceito qualquer na ontologia. O processo de adaptação pode utilizar tanto as informações da seqüência de acessos como as informações obtidas com as relações do contexto semântico, que podem ser generalizadas.

A seguir são descritos os elementos envolvidos neste processo.

7.2.1 Anotação semântica e Ontologia

Para a obtenção de melhorias nas possibilidades de adaptação foram adotados recursos de anotação semântica para o conteúdo do *site* Web. Também foi realizada a descrição da estrutura do *site* Web e de algumas relações significativas em uma ontologia do domínio da aplicação. Uma ontologia permite a definição de conceitos e de relações entre estes, sendo possível descrever conceitos e relações mais gerais ou mais específicos. No presente trabalho foi adotada a abordagem de descrição mais específica, sendo que uma vantagem neste caso é a indicação das relações e dos conceitos de interesse para cada aplicação. Entretanto, esta decisão implica em revisão da ontologia a cada domínio de aplicação diferente, para adaptação e descrição de relações significativas.

Para a implementação de estudos de caso e validação da abordagem proposta foram definidas aplicações na área educacional. Assim a ontologia de domínio possui como objetivo descrever conceitos da área educacional. Entretanto a metodologia pode ser aplicada a diversas outras áreas. Esta ontologia foi descrita manualmente por especialistas no domínio da aplicação, com uso do editor de ontologias Protégé[59], tendo sido utilizada a linguagem OWL[60] para sua representação e manipulação. Esta forma de representação proporciona vantagens para etapas posteriores por facilitar a sua manipulação e integração com os dados tratados nos processos implementados.

---

59 http://protege.stanford.edu

60 http:// www.w3.org/2004/OWL

A figura 7.5 ilustra um trecho da ontologia de domínio descrita nesta aplicação. Nela podem ser observadas relações entre os elementos “topico” e “curso”. As relações “parteDe” e “composto_por_topico” indicam composição de elementos. Já as relações como “possuiRequisito” e “ehRequisitoDe” permitem indicar dependências entre os tópicos de um curso. A relação “do_tipo” possibilita qualificar aspectos de cada componente do *site* Web. A relação “tipo_de_conteudo” serve para descrecer uma anotação semântica sobre conceitos tratados no documento Web. Estas definições compõem uma parte da estrutura geral da ontologia. O relacionamento desta com o conteúdo disponibilizado no *site* Web é feito pela anotação semântica do mesmo. Para tanto cada item do *site* Web é descrito como uma instância na ontologia, com todas as relações necessárias. Estas instâncias serão utilizadas para o relacionamento com as informações de uso, conforme será detalhado adiante.

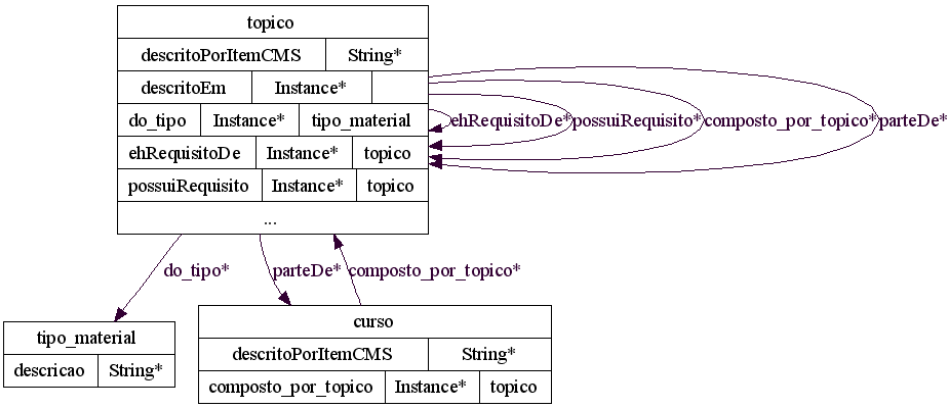


Figura 7.5: Trecho da ontologia de domínio utilizada

A utilização do editor de ontologias permite que sejam descritos e atualizados todos os elementos do *site* Web em um ambiente integrado com opções diversas. Cada página Web publicada no ambiente de gerenciamento de conteúdo Web que foi utilizado nos experimentos iniciais foi indicada na ontologia como uma instância específica. Assim esta informação pode ser utilizada ainda dentro do ambiente do editor de ontologias para fins de validação ou de consulta. Após verificada a sua correção, os dados podem ser exportados em formatos adequados para a sua manipulação, como no caso da utilização da linguagem OWL nos experimentos.

A figura 7.6 abaixo ilustra uma tela do editor de ontologias utilizado onde pode ser observada a edição de uma instância da ontologia para a anotação semântica de um item publicado. Na tela estão dispostas todas as propriedades associadas com a instância em questão. Podem ser ressaltadas, neste caso, a informação designada para o nome, no campo “*Name*”, onde está indicado o valor “ID_12” que será utilizado nas tarefas de integração semântica. Também a relação “*composto_por_topico*” pode ser observada, sendo que estão descritos dois tópicos (“ID_43” e “ID_33”) que são relacionados por esta propriedade com o tópico exemplificado.

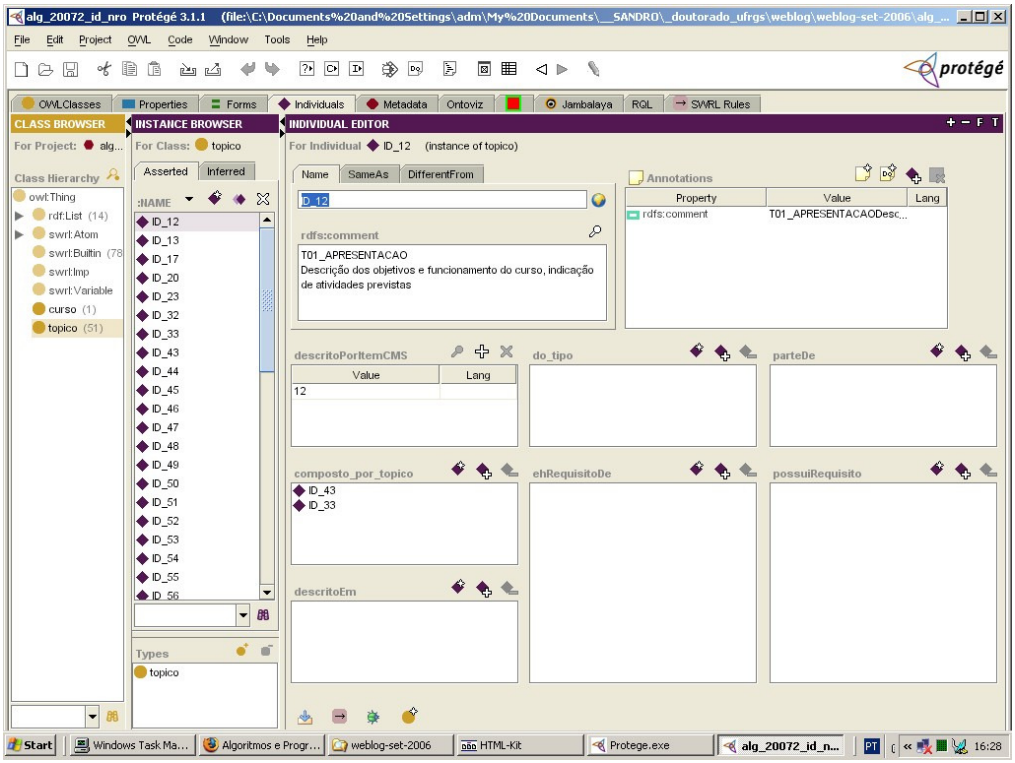


Figura 7.6: Exemplo de edição de anotações semânticas na ontologia

7.2.2 Aquisição e tratamento de informações de uso da Web

A Mineração de Uso da Web pode ser realizada a partir de etapas bem definidas envolvendo aquisição e pré-processamento de dados de uso, geração e análise de padrões. Os dados utilizados normalmente são extraídos dos registros de acesso do próprio servidor Web, ou então gerados a partir de inclusões específicas de código nas páginas do *site* Web. Neste trabalho é utilizado um mecanismo de captura a partir de acréscimos à codificação das páginas Web, permitindo a coleta de informações armazenadas com uso de um *cookie*. As páginas publicadas recebem a inclusão de um módulo de codificação específica, permitindo a geração de um *cookie* e a gravação de informações de acesso. O objetivo do uso deste *cookie* é auxiliar na identificação de sessões de usuários, pois mecanismos baseados em heurísticas usando o horário de acesso e números IP (Internet Protocol[61]) sofrem problemas em situações onde existe o uso de servidores *proxy* e tradução de endereços. O maior problema da abordagem adotada ocorre caso a opção de geração de *cookies* seja desabilitada pelo usuário, pois assim o processo de identificação não pode operar corretamente.

As informações armazenadas a cada acesso de um usuário são descritas em um formato XML[62]. Este formato é utilizado para proporcionar facilidades ao tratamento dos dados de uso. A flexibilidade proporcionada pelo formato XML facilita o pré-processamento resultando em subsídios para a utilização de mecanismos de geração de padrões freqüentes de acesso e também para a utilização de outros sistemas de mineração de dados.

---

61 http://www.ietf.org/rfc/rfc0791.txt

62 http://www.w3c.org/xml

A figura 7.7 abaixo ilustra o formato e elementos utilizados para a gravação de dados de acesso. Nesta etapa em particular, segue-se a metodologia descrita em Oliveira (2006). Os elementos permitem armazenar dados como a data e horário de acesso, número IP (*Internet Protocol*[63]) do local de origem do acesso, URL (*Universal Resource Locator*) acessada, parâmetros de acesso, navegador utilizado, data e horário de acesso e identificador (*cookie*). A identificação de origem do acesso é armazenada no elemento "userid", conforme o cookie gerado durante o acesso inicial de um usuário. O elemento "adapta" permite diferenciar os acessos feitos em páginas da estrutura normal do *site* e os acessos feitos nas demais páginas, geradas com o processo de adaptação.

```
<acesso>
        <ip>201.37.126.43</ip>
        <page>/cms01/index.php</page>
        <parametro>34</parametro>
        <agent>Mozilla/5.0 (...) Gecko/20050717 Firefox/1.0.6</agent>
        <data>12/11/2006</data><horario>13:04:23</horario>
        <userid>f4b3173f4a4efb248f6f200c5ce678e3</userid>
        <adapta>0</adapta>
</acesso>
```

Figura 7.7: Exemplo de elemento obtido com a coleta de dados

O módulo adicional necessário para a captura dos dados de uso é bastante simples e um trecho da codificação do mesmo pode ser visualizado na figura 7.8, onde pode ser visualizada a captura de informações de acesso como número IP e parâmetros (linhas 1 até 5); a captura de informações do *cookie* e a geração do mesmo (linhas 6 até 13); e a montagem do elemento de acesso (linhas 14 até 21), seguida da gravação em um arquivo XML (não incluída neste exemplo).

Com este mecanismo implementado, a cada acesso de um usuário ao *site* Web, será gerada a informação descrita na figura 7.7 acima. Apesar de uma coleta de informações similar já estar sendo feita automaticamente pelo servidor Web, os *logs* gerados não incluem a informação de identificação do usuário, sendo que para a identificação de uma sessão do mesmo serão necessárias algumas heurísticas que normalmente usam como base os dados de localização (IP) e de horário do acesso. Entretanto estas informações não são completamente confiáveis ou suficientes e podem induzir a erros de interpretação em casos bastante conhecidos. Assim identifica-se no mecanismo proposto neste trabalho a vantagem de facilitar a identificação dos acessos de uma sessão de usuários. Outra vantagem é a possibilidade de tratamento dos dados relevantes e não de todos os dados mantidos em um registro de uso de um servidor Web, que possuem diversas informações que não são empregadas no método de acompanhamento de uso e podem gerar volumes bastante grandes de dados, dependendo da freqüência de acessos observada no *site* em questão.

---

63 http://www.ietf.org/rfc/rfc0791.txt

```
1       $ip = $_SERVER['REMOTE_ADDR'];
2       $page = $_SERVER['PHP_SELF'];
3       $agent = $_SERVER['HTTP_USER_AGENT'];
4       $data = date("d/m/Y");
5       $horario = date("H:i:s");
6       if(empty($_COOKIE["CookieSessionUserID"])) {
7               $IDuser = md5($ip);
8               $tempoVida = time() + 86400*365;
9               setcookie("CookieSessionUserID", $IDuser, $tempoVida);
10      }
11      else {
12              $IDuser = $_COOKIE["CookieSessionUserID"];
13      }
14      $string  ="<acesso><ip>".$ip."</ip>";
```

Figura 7.8: Trecho de codificação (em PHP) para captura de dados de acesso

Os dados capturados são processados com determinadas heurísticas, resultando em seqüências que descrevem o percurso (sessão) de cada usuário. Estas heurísticas estão relacionadas basicamente com o tratamento de situações ligadas ao uso da Web. Como existe, por exemplo, a possibilidade de um determinado usuário acessar uma determinada página e deixar o navegador com esta página aberta enquanto outras tarefas são realizadas, este fato pode gerar discrepâncias quanto ao tempo de acesso. Para minimizar estas ocorrências são admitidos tempos máximos de permanência a partir dos quais se supõe que a atitude do usuário não é mais a leitura da página.

Deve-se ressaltar também problemas conhecidos desta abordagem (já referenciados anteriormente), como a possibilidade do uso de *cookies* ter sido bloqueado a partir de uma decisão do usuário. Como esta opção é disponibilizada pelos navegadores Web, é possível que alguns usuários não sejam identificados pelo mecanismo, em função da escolha desta configuração.

Os dados obtidos podem ser utilizados para a simples geração de informações já bastante conhecidas e oferecidas em sistemas convencionais de acompanhamento de uso de sites Web, como estatísticas de páginas mais acessadas em um determinado período. Entretanto, neste trabalho, o objetivo maior é seu uso na obtenção de percursos para usuários e também na obtenção de padrões de acessos seqüenciais freqüentes. No tratamento das informações capturadas é possível a obtenção de percursos, identificados como uma seqüência de acessos a um mesmo *site* e relacionada a um mesmo identificador (mantido pelo *cookie* descrito anteriormente). Com isso é viável, em análise posterior, reunir os percursos idênticos observados, para seções diferentes.

Conforme descrito no item 3.3.1.2 anteriormente, o contexto de análise de padrões seqüenciais para uso de *sites* Web pode ser formalizado considerando-se um conjunto $P=\{p_1, p_2, p_3, ..., p_n\}$ como o conjunto de *n* diferentes páginas de um determinado *site* Web. Uma seção de acesso de um usuário pode ser considerada como o acesso a um conjunto L de *m* páginas pertencentes a P, sendo que estes acessos são referenciados freqüentemente na literatura como *page-views*. Assim considera-se para um determinado usuário, em visita ao *site* Web, a geração de um conjunto não vazio $L=\{l_1, l_2, l_3, ..., l_m\}$, sendo cada página l_i pertencente a P. O acompanhamento dos acessos realizados por diferentes usuários, ou pelo mesmo usuário em diferentes

momentos, permite a geração de diversos conjuntos do tipo L. Ao final de um período de tempo em que os acessos a um determinado *site* estejam sendo acompanhados, o resultado obtido será o conjunto de todos as seções de acesso dos usuários deste *site*.

Um padrão seqüencial é considerado, neste contexto, como uma seqüência de acessos a páginas do *site* Web (ou *page-views*) observado de forma repetitiva com diversos usuários. Um padrão seqüencial freqüente seria considerado como um padrão seqüencial que se encontra dentro de limites desejados quanto ao número de ocorrência em relação aos padrões observados de forma geral e um limiar de repetição arbitrário

A seguir, na figura 7.9, estão relacionados dois trechos exemplificando estes resultados, descrevendo um conjunto de percursos gerais e um conjunto de percursos resumidos para este mesmo conjunto geral. Cada percurso é indicado com uma nova linha onde está identificado em numeração seqüencial seguida pela informação do nome da seção do *site* e do tempo de permanência em segundos nesta seção, descrito entre parênteses. Assim, para a relação de percursos gerais, a linha "Percurso [1] apresentacao (3s) capa (172s) transporte (9s)" estaria indicando o número seqüencial do percurso (1) e o acesso às seções do *site* identificadas como "apresentacao", "capa" e "transporte", sendo que o tempo de permanência nestas é de 3, 172 e 9 segundos, respectivamente. Para a relação de percursos resumidos são indicados o número de ocorrências deste percurso, o tempo médio geral do percurso e a seqüência de seções acessadas. Assim, a linha "Percurso Resumido [2] contagem = 91 Tempo total médio=286s [0] bsb [1] programacao [2] capa" deve ser entendida como o percurso resumido de número 2, para o qual existem 91 ocorrências detectadas, sendo que as seções "bsb", "programacao" e "capa" compõe o percurso, nesta ordem indicada.

Percursos Gerais Obtidos:

Percurso [1] apresentacao (3s) capa (172s) transporte (9s) eventos (113s) til (4s) apresentacao (2s)

Percurso [2] apresentacao (42s) turismo (6s) wie (316s) faleconsco (14s) ctic (8s)

...

Percursos Resumidos:

Percurso Resumido [1] contagem = 23 Tempo total médio=196s [0] eventos [1] til [2] semish [3] capa

Percurso Resumido [2] contagem = 91 Tempo total médio=286s [0] bsb [1] programacao [2] capa

...

Figura 7.9: Geração de percursos gerais e percursos resumidos

Observa-se que para todos os sites analisados este conjunto de percursos resumidos é sempre pequeno em relação ao número de percursos totais, o que se atribui às características particulares de cada usuário e de seus interesses durante o acesso aos sites. Este fato apontou à necessidade de localizar padrões repetidos, como forma de generalização de comportamentos. A obtenção de padrões freqüentes é implementada neste trabalho como uma etapa adicional para a obtenção de informações complementares que possam ser relacionadas às demais informações utilizadas. A geração de seqüências de acessos freqüentes foi desenvolvida com a linguagem C++[64],

64 http://www.research.att.com/~bs/C++.html

com uso do parser Expat[65], como forma de tratar de modo eficiente os dados obtidos. A implementação está baseada no algoritmo Spade e em melhorias descritas na literatura, conforme Zaki (2001) e Leleu (2003). A solução descrita permite o tratamento de seqüências de dados com volume médio, em torno de 30 Mbytes, com pequeno tempo de processamento e sem uso excessivo de memória.

Dados de 19/06/2005 até 26/06/2005

Número total de padrões obtidos = 2330

Nro Sequência

31 secomu semish til

33 enia jai secomu

34 jai secomu semish

34 Capa hospedagem Capa

38 Capa inscricao programacao

39 apresentacao eventos programacao

39 Capa programacao eventos

40 Capa inscricao inscricao

43 Capa programacao Capa

65 inscricao Capa inscricao

86 Capa inscricao Capa

130 Capa Capa Capa

Figura 7.10: Resultados para padrões freqüentes

A figura 7.10 acima ilustra os resultados obtidos em um experimento. Ela apresenta uma lista de percursos resumidos e seu índice de freqüência. Podem ser observados períodos diferentes, de forma a possibilitar a obtenção de comportamentos típicos de cada período. O caso analisado neste exemplo é o *site* Web do Congresso da Sociedade Brasileira de Computação (SBC) da sua edição do ano de 2005.

O acompanhamento de uso da Web, conforme citado, é realizado a partir da coleta dos dados com uso de codificação específica e de *cookies*, sendo o pré-processamento dos dados modificado para que possam ser integradas as informações da ontologia de domínio da aplicação. Deste modo, o tratamento de informações de percursos freqüentes obtidas pela análise das sessões de usuários pode ser ampliado, observando-se a informação semântica de cada página presente no percurso. Esta integração está descrita no item a seguir.

7.2.3 Utilização da Ontologia em operações de consulta

As informações das instâncias disponibilizadas na ontologia podem ser aproveitadas por mecanismos de inferência e por linguagens de consulta, como SPARQL[66]. Nestas linguagens é possível verificar a ocorrência de diversas formas de relacionamento das instâncias da ontologia. Como exemplo, pode-se identificar, a partir de uma instância conhecida, todas as relações com ela associadas. Ou então recuperar, a

---

65 http://expat.sourceforge.net/

66 www.w3.org/TR/rdf-sparql-query/

partir de uma propriedade, todas as instâncias relacionadas pela mesma. Ainda é possível descobrir todas as relações existentes entre duas instâncias conhecidas, entre diversas outras possibilidades. Um exemplo da descrição de elementos do domínio da aplicação com a ontologia e de utilização de comandos de consulta podem ser observados nas figuras 7.11 e 7.12 abaixo.

```
<topico rdf:ID="ID_24">
  <parteDe rdf:resource="#Banco_de_dados"/>
  <composto_por_topico>
    <topico rdf:ID="ID_25">
      <parteDe rdf:resource="#ID_24"/>
      <descritoPorItemCMS>24</ descritoPorItemCMS >
      <rdfs:comment >T05_01_menu_cms</rdfs:comment>
    </topico>
  </composto_por_topico>
  <descritoPorItemCMS>24</descritoPorItemCMS>
  <rdfs:comment >T05_ATIVIDADES</rdfs:comment>
</topico>
```

Figura 7.11: Exemplo de trecho da descrição das instâncias na ontologia

Na figura 7.11 observa-se um trecho, na linguagem OWL, contendo algumas instâncias da ontologia de domínio descrevendo informações semântica relacionadas com a aplicação usada para testes. São exemplificadas as descrições dos itens "T05_ATIVIDADES" e "T05_01_menu_cms", identificados respectivamente como "ID_24" e "ID_25". A relação "composto_por_topico" define a hierarquia entre os dois. A propriedade "descritoPorItemCMS" permite a anotação semântica dos conteúdos, tais como armazenados no sistema de gerenciamento de conteúdo Web utilizado. A relação "parteDe" relaciona o tópico identificado como "ID_24" como sendo parte do curso "Banco de dados". Todas estas descrições são serializadas em conjuntos compostos por três elementos onde o primeiro é uma instância da ontologia, o segundo uma propriedade e o terceiro elemento outra instância ou um literal. Estes elementos podem ser consultados de forma eficiente por linguagens de consulta.

a)
```
PREFIX v:<http://www....br/.../...owl>
    SELECT ?x, ?y WHERE (v:ID_24, ?x, ?y),
                        (v:ID_25, ?x, ?y)
```

b)
```
PREFIX v:<http://www....br/.../...owl>
    SELECT ?x WHERE (v:ID_24, ?x, v:ID_25)
```

Figura 7.12: Exemplos de trechos de consultas em SPARQL

Na figura 7.12 são ilustradas algumas possibilidades de identificação de relações com a linguagem de consulta SPARQL. Os identificadores dos tópicos na ontologia coincidem com os parâmetros gerados pela ferramenta de gerenciamento de conteúdo empregada no *site* Web. Isso permite a utilização das informações de acesso para integração com instâncias descritas na ontologia. Estas consultas são realizadas com os elementos já identificados por padrões de uso freqüente. No primeiro exemplo (a) o resultado será a identificação de todas as relações e instâncias associadas com os tópicos

indicados ("ID_24" e "ID_25"). No segundo exemplo (b) serão recuperadas todas as relações entre estes dois tópicos. Além de linguagens de consulta podem ser utilizados mecanismos de inferência e também a integração com informações contidas em outras ontologias complementares. Por exemplo, podem ser usadas ontologias que descrevam informações sobre localização, sobre tempo ou sobre os conteúdos de forma mais detalhada.

O resultado destas consultas permite a identificação do contexto que não se encontrava explícito nas informações do percurso freqüente. Este contexto apóia a definição da tarefa sendo executada, considerada neste trabalho para a identificação das classes de comportamento que serão empregadas nas adaptações. A cada classe de comportamento são associadas regras específicas de adaptação, adequadas ao contexto identificado.

7.2.4 Aplicação de regras de adaptação

O tratamento das regras de adaptação foi estudado a partir de duas abordagens. Na primeira foram utilizadas regras simples, nas quais a seqüência de acessos de um percurso delimita as adaptações de estrutura que são sugeridas. Desta forma, ao ser detectada uma coincidência parcial entre a sessão do usuário e algum dos padrões de percusos freqüentes minerados utiliza-se a lista de páginas ainda não acessadas pelo usuário como alvo da adaptação. Este formato foi empregado em algumas experimentações iniciais. Seus resultados são analisados adiante neste texto, em conjunto com os resultados obtidos pela segunda abordagem para a geração de regras de adaptação.

Nesta segunda abordagem são levadas em conta as informações descritas no contexto semântico obtido com a integração de dados de uso e ontologia de domínio. Nesta abordagem as regras são obtidas não apenas pela utilização das páginas ainda não acessadas pelo usuário em uma sessão, mas pelo emprego das relações descritas na ontologia de domínio. Em cada contexto semântico são detectadas algumas relações predominantes, sendo que a ontologia é consultada em busca destas relações. Com isso são identificados possíveis itens de interesse do usuário. Alguns exemplos de resultados obtidos são a identificação de relações de composição ou complementariedade e sua sugestão.

A partir dos contextos semânticos obtidos com a integração da mineração de uso da Web e da ontologia de domínio, são realizadas duas operações para a geração de adaptações de estrutura do *site* Web. A primeira operação é a verificação de similaridade entre os padrões iniciais para que possam ser agrupados os padrões com similaridade acima de um limite arbitrário. A segunda operação é a identificação e comparação de sessões de usuários, durante a visita destes ao *site* Web. A identificação é realizada a partir do mecanismo de armazenamento de informações em *cookies*, o que facilita para o sistema esta tarefa. A comparação da sessão de um usuário com os demais padrões somente ocorre após a transformação dos dados desta sessão para um contexto semântico. Depois desta etapa é gerado um valor para a similaridade entre os padrões, para proporcionar o seu agrupamento.

Com as informações de similaridade o sistema pode realizar ações de adaptação que empregam as informações da seqüência de acessos e também as relações descritas na ontologia. Assim a adaptação pode utilizar as relações predominantes no contexto semântico ou então pode buscar por relações complementares descritas na ontologia e relacionadas com os itens da sessão do usuário.

Cada adaptação gerada é anotada de forma diferenciada caso seja acessada pelo usuário, desta forma gerando uma base para comparações entre os acessos feitos à estrutura geral do *site* Web e os acessos realizados à estrutura modificada pela adaptação.

7.2.5 Resumo e análise do processo definido

A etapa final deste processo de integração consiste na geração de padrões mais significativos, a partir dos padrões de uso e das informações descritas na ontologia. Para tal é seguido o processo da figura 7.2, que relaciona os padrões de acesso freqüente descobertos com as relações descritas na ontologia.

Os padrões obtidos são mais abrangentes do que os padrões relacionados apenas com a utilização do *site* Web. As associações destes novos padrões com regras específicas de adaptação permitem a geração de resultados mais expressivos. Alguns exemplos destes resultados são a identificação de usuários buscando material sobre um determinado tópico, de usuários acessando o *site* para obter uma visão geral dos conteúdos, acessando material descritivo sobre tarefas, acessando exercícios ou conteúdos complementares.

O acompanhamento de uso da Web é realizado a partir da coleta dos dados com uso de codificação específica e de *cookies*, sendo o pré-processamento dos dados modificados para que possam ser integradas as informações da ontologia de domínio da aplicação. Deste modo, o tratamento de informações de percursos freqüentes obtidas pela análise das sessões de usuários pode ser ampliado, observando-se a informação semântica de cada página presente no percurso.

Dados internos:

Lista de acessos do usuário

Lista de padrões freqüentes de acesso

Ontologia de domínio

Entrada: URL de página acessada

Saída: Conteúdo da página e lista de sugestões de adaptação

Roteiro:

Para cada acesso

Recebe URL com identificador da página

Atualiza o registro de acessos

Busca a identificação da sessão do acesso atual

Compara percurso da sessão com percursos freqüentes

Identifica o tópico correspondente na ontologia

Recupera relações do percurso da sessão na ontologia

Consulta regras de adaptação

Gera opções de adaptação

Busca conteúdo da página solicitada

Integra conteúdo e adaptações

Envia resultado ao usuário

Figura 7.13: Integração de informações de uso com semântica

A figura 7.13 indica o roteiro do processo geral seguido no trabalho, com a definição da seqüência de etapas, dados de entrada e resultados. Os dados internos considerados envolvem a lista de acessos dos usuários (sendo que cada sessão pode ser diferenciada com a consulta ao valor do *cookie* de identificação utilizado), a lista de padrões freqüentes de acesso e, por fim, a ontologia de domínio utilizada na aplicação em questão. A cada acesso de usuários a URL correspondente à solicitação gerada é empregada como o valor de entrada a ser considerado no processo. O resultado de saída esperado é o conteúdo normal da página solicitada, acrescido da lista de sugestões de adaptação. Para a geração deste resultado são observados os passos resumidos na figura 7.3.

7.3 Abordagem geral para a prototipação

O mecanismo de adaptação implementado está organizado em torno das etapas gerais envolvidas neste processo de adaptação. Para facilitar a experimentação e sua possível adoção foi utilizada, em uma primeira abordagem, uma composição com um sistema gerenciador de conteúdo Web disponível a partir de código aberto (Joomla)[67]. Junto a este sistema foram incluídos módulos para coleta dos dados de acesso e para a geração de adaptações. O processo de Mineração de Uso da Web utiliza um componente executado periodicamente para a geração de percursos freqüentes, possibilitando assim o tratamento de um volume grande de dados sem impacto para o sistema na Web. O processo de integração com as informações semânticas e geração de regras é realizado nos mesmos moldes. A anotação semântica do conteúdo é realizada com o software Protégé, a partir de atualizações do conteúdo do *site* Web. A seguir são detalhadas estas etapas.

Na figura 7.14 está ilustrado este processo de forma geral. A partir de uma requisição enviada pelo usuário do *site* Web, esta é armazenada e compõe o "*log* de acessos" que será utilizado na Mineração de Uso da Web. Após o pré-processamento destes dados e da geração de percursos freqüentes é realizada a integração destes percursos e das informações semânticas, descritas na ontologia de domínio. Com estas duas fontes de informações são identificadas não apenas as seqüências de páginas acessadas, mas os conceitos e as relações entre elas, o que é avaliado e usado para gerar as regras de adaptação que serão empregadas no sistema. Com a utilização do *cookie* identificador do usuário é possível verificar os dados de sua sessão e descobrir a ocorrência de alguma sobreposição com percursos freqüentes já identificados. Esta informação e as informações do conjunto de regras de adaptação serão utilizadas para a geração do resultado final.

O resultado final é composto pelo conteúdo originalmente solicitado, pela organização padrão da interface do *site* e por elementos da mesma que podem receber novos apontadores de acordo com a detecção do padrão de acesso. A organização do sistema de gerenciamento de conteúdo utilizado permite a definição de padrões gerais para a interface gerada. Nestes padrões podem ser especificadas áreas que poderão receber algum conteúdo adaptado. Assim o resultado obtido pelo usuário será sempre o conteúdo normal do *site* Web e, de acordo com sua navegação, sugestões de apontadores nestas áreas de adaptação.

---

67 http://www.joomla.com.br/

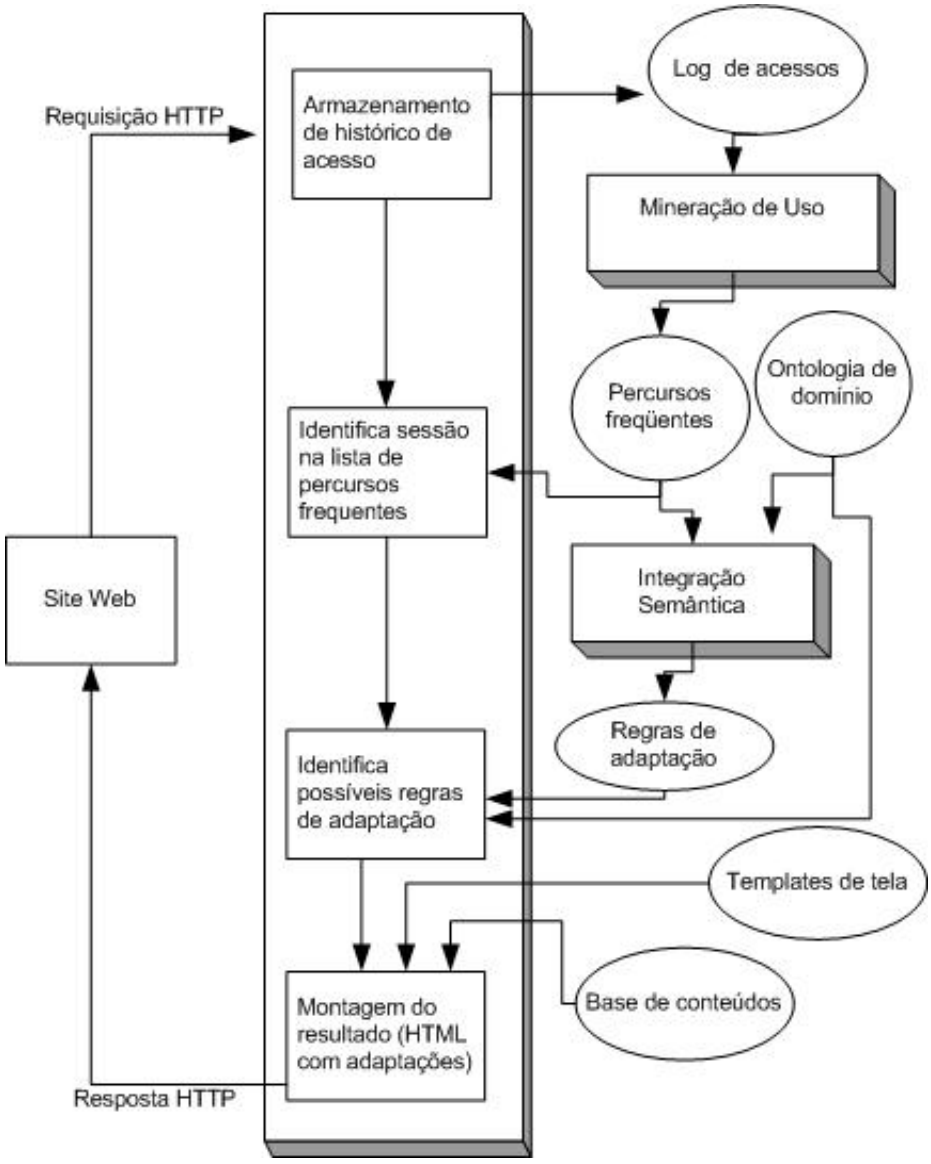


Figura 7.14: Elementos da arquitetura implementada

Os resultados da utilização das regras de adaptação podem ser observados na figura 7.15 abaixo. Nela estão identificadas informações que são acrescidas à estrutura original do *site* Web a partir da coleta e processamento dos dados de uso. De acordo com as informações já coletadas e processadas, o sistema implementado possui acesso à descrição de percursos freqüentes, regras de adaptação e ainda a descrição de relações específicas a partir da ontologia de domínio. A partir do comportamento observado para um determinado usuário, estas informações são utilizadas como complementares na estrutura original do *site* Web e são publicadas nas áreas indicadas. Por exemplo, as informações semânticas são obtidas a partir das relações existentes entre os itens na ontologia. Cada relação é avaliada de acordo com as regras em uso e assim são sugeridas ou não adaptações com sua utilização, de acordo com as características da relação e com as regras selecionadas.

No exemplo da figura 7.15 pode ser observada a mensagem “Requisitos para este tópico” seguida de um apontador gerado com este processo. Já na parte esquerda inferior da imagem pode ser observada a mensagem “Links sugeridos”, seguida por um

conjunto de apontadores relacionados com a tarefa identificada pelas regras de adaptação. Para que pudessem gerar subsídios para a avaliação das adaptações, os acessos a estes itens gerados como adaptação de estrutura serão recebidos pelo sistema com um parâmetro adicional. Este é identificado como o elemento "adapta" descrito na figura 7.7 e é aplicado em avaliações posteriores.

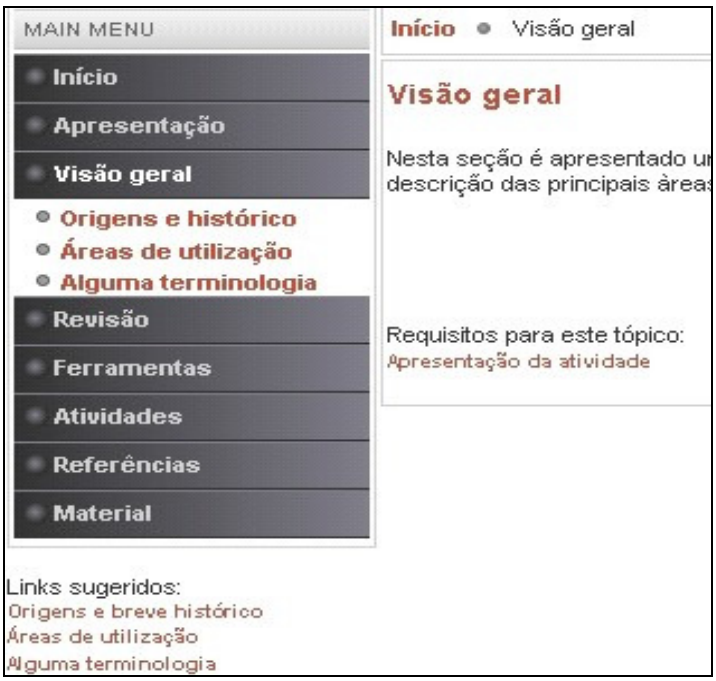


Figura 7.15: Adaptação de estrutura em experimento

O pré-processamento dos dados de uso e a geração dos padrões utilizados em adaptações são feitos periodicamente. Durante as sessões de usuários o sistema detecta, a partir de um histórico recente de acessos contendo apenas os acessos da sessão, alguma sobreposição deste comportamento observado com os padrões anteriormente descobertos. Neste caso o sistema realiza a adaptação associada. Nestes experimentos, a adaptação consiste em modificações da estrutura da página resultante, com o acréscimo de novas possibilidades de navegação, originadas nos padrões descobertos.

7.4 Experimentos para validação da proposta

A seguir são descritos três conjuntos de experimentos realizados para a verificação de características da abordagem desenvolvida na tese.

O primeiro conjunto de experimentos teve como objetivo o tratamento de registros de acesso e a obtenção de padrões freqüentes de acesso. Nos outros dois conjuntos de experimentos os objetivos incluíam a geração de adaptações. Estes se diferenciam pela forma de geração das adaptações.

No segundo conjunto de experimentos o objetivo principal foi o tratamento dos dados de uso da Web e sua utilização na geração de adaptações. Nele não foram realizadas as operações de integração de informações de uso e de informações semânticas previamente. Este experimento apresenta resultados que indicam a possibilidade de utilização conjunta das tecnologias propostas. Foi realizado com a

geração de percursos freqüentes e de agrupamentos. Nas adaptações são utilizadas as informações dos percuros freqüentes, dos agrupamentos e da ontologia para a geração de links adicionais, modificando assim a estrutura das páginas apresentadas.

No terceiro conjunto de experimentos as adaptações baseiam-se nos contextos semânticos obtidos e na geração de sugestões de adaptação a partir de regras obtidas nas relações das ontologias. Nestes experimentos são descritas e analisadas com maior detalhe as opções de integração de recursos de uso e informações semânticas. Estes experimentos foram repetidos em diversos contextos, possibilitando a identificação de facilidades para a utilização de ontologias diversas. As adaptações geradas foram acompanhadas e indicam um percentual satisfatório de acessos aos itens sugeridos pela adaptação.

7.4.1 Descrição dos experimentos de aquisição e processamento de dados de uso

Este conjunto de experimentos foi realizado para avaliação e melhoria do mecanismo de aquisição de dados de uso da Web. Foram realizadas experiências com aquisição de registros de uso em diversos contextos, que incluíam sites de cursos de graduação, *sites* pessoais e *sites* de eventos científicos. Os dados obtidos foram empregados para a descrição de tarefas de pré-processamento necessárias e como fontes para a aplicação de algoritmos de mineração de padrões de acesso freqüentes.

A forma de aquisição de dados de uso empregada foi a descrita no item 7.7.7 e estes experimentos colaboraram na identificação de diversos problemas práticos para o tratamento dos dados de uso. Alguns exemplos destes problemas são: o tratamento de dados originados por acessos de sistemas de recuperação de informações na etapa de aquisição de dados para a geração de índices, o tratamento de situações de uso de memórias *cache,* o tratamento de volumes diferenciados de dados, o processamento eficiente de conjuntos extensos de registros, a geração de dados para utilização em sistemas de mineração de dados já existentes e a geração de formatos diferenciados para a representação dos resultados.

Nestas experiências foram acompanhados os seguintes *sites* Web:

- Curso de graduação – Licenciatura em Educação Física, acessível na URL: http://www.unisinos.br/graduacao/licenciatura/educacao_fisica/. Número de registros tratados: 30.279.
- Curso de graduação – Licenciatura em Filosofia, acessível na URL: http://www.unisinos.br/graduacao/licenciatura/filosofia/. Número de registros tratados: 10.593.
- Curso de graduação - Licenciatura em Física, acessível na URL: http://www.unisinos.br/graduacao/licenciatura/fisica/. Número de registros tratados: 12.728.
- Curso de graduação – Licenciature em História, acessível na URL: http://www.unisinos.br/graduacao/licenciatura/historia/ . Número de registrs tratados: 13.979.
- Curso de graduação – Licenciatura em Letras – acessível na URL: http://www.unisinos.br/graduacao/licenciatura/letras/. Número de acessos tratados: 41.939.

- Curso de graduação – Licenciatura em Matemática, acessível na URL: http://www.unisinos.br/graduacao/licenciatura/matematica/. Número de acessos tratados: 14.952.
- Curso de graduação – Licenciatura em Pedagogia, acessível na URL: http://www.unisinos.br/graduacao/licenciatura/pedagogia/. Número de acessos tratados: 22.254.
- Sistema de busca de informações, acessível na URL: http://www.unisinos.br/busca/index.php?Itemid=87. Acompanhamento de 9.566 acessos, nos quais foram monitoradas informações de acesso e informações de palavras indicadas.
- Site pessoal disponível na URL: http://www.inf.unisinos.br/~rigo. Número de acessos tratados: 20.432 acessos
- Site de evento científico – SBC 2005, acessível na URL: http://www.unisinos.br/congresso/sbc2005/. Número de acessos tratados: 216.854 acessos.

Os diferentes sites acompanhados também permitiram que o mecanismo de aquisição de dados fosse testado com situações diferenciadas quanto à organização do sistema de publicação de conteúdo. Em alguns dos sites (os de cursos de graduação) foi utilizado um sistema de gerenciamento de conteúdo Web (Mambo). No site de evento científico foi utilizada uma aplicação Web proprietária e no site pessoal foi utilizado um mecanismo simples de publicação. Os resultados das coletas de dados indicaram facilidades para a integração do mecanismo em diferentes contextos. Deve se ressaltar que todos os casos utilizavam a mesma plataforma Web, porém casos de plataformas diferenciadas podem ser tratados com a utilização de serviços Web (*web services*).

Em todos os experimentos, foram repetidas algumas das etapas da metodologia proposta, com a inclusão de um módulo de aquisição de dados de uso e o acompanhamento periódico dos resultados obtidos. A cada período de análise foram geradas informações sobre padrões de acesso freqüentes. Estes padrões foram avaliados manualmente, com a finalidade de validação dos resultados. Como em alguns casos os *sites* foram acompanhados por períodos extensos, com mais de seis meses, também foram realizadas operações de comparação dos padrões gerados nos diferentes períodos de tempo.

7.4.2 Descrição do experimento com aquisição de dados de uso e adaptação

Este segundo experimento foi definido para possibilitar a experimentação do mecanismo de aquisição de dados de uso da Web, principalmente. Além disso, foram incluídas como tarefas a utilização dos resultados da mineração como subsídios para a adaptação de estrutura de *sites* Web. Para a realização deste trabalho foi implementado um *site* Web usando a versão 4.5.2 do gerenciador de conteúdo Web Mambo, disponível em http://www.mamboserver.com. Os motivos da escolha deste gerenciador de conteúdo foram a disponibilidade de acesso ao código fonte deste sistema, a facilidade de uso do banco de dados MySQL[68] empregado no sistema, e a sua

[68] http:// www.mysql.com/

implementação em linguagem PHP[69], em função da experiência anterior da equipe de desenvolvimento.

O *site* Web implementado foi disponibilizado na Internet para possibilitar a geração dos testes de uso com um número suficiente de usuários. O conteúdo do *site* foi selecionado entre materiais educacionais disponibilizados de forma gratuita na Internet, com objetivo de possibilitar a simulação de um ambiente Web para disponibilização de material de ensino, delimitando assim o escopo do trabalho a este domínio específico. Esta escolha determinou a descrição da ontologia de domínio utilizada para experimentos de adaptação. A figura 7.16 ilustra a interface implementada.

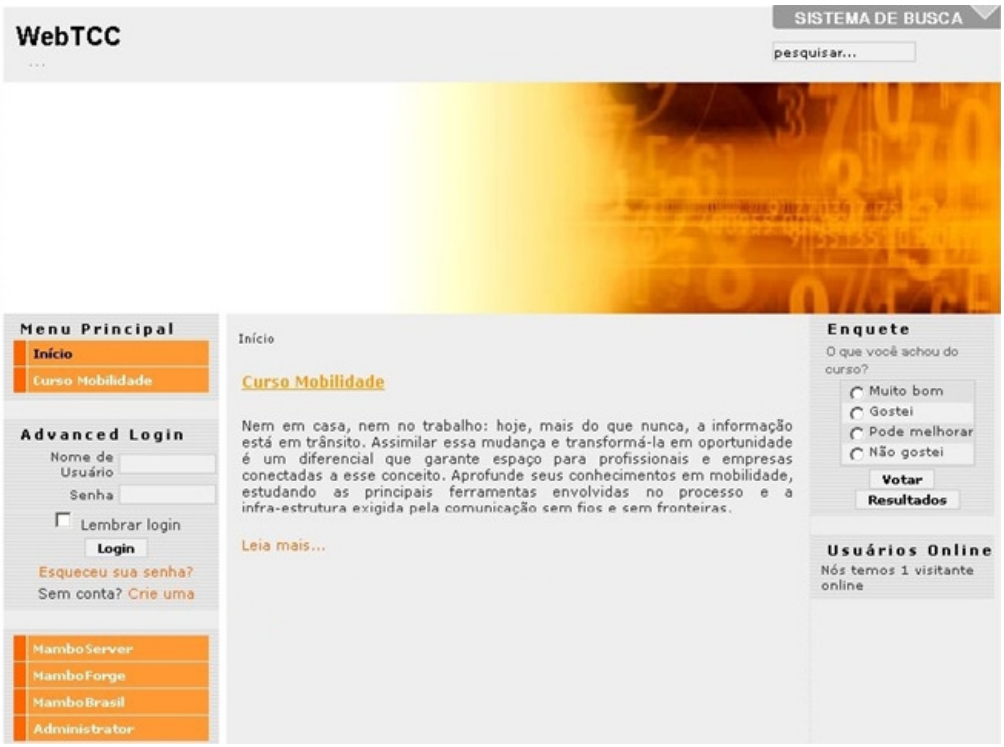


Figura 7.16: Interface do site criado com o gerenciador de conteúdo Web

O gerenciador de conteúdo Web Mambo é um sistema com ferramentas de gerenciamento que podem ser acessadas com um navegador web, permitindo a realização remota de todo o processo de criação, manutenção e arquivamento do conteúdo. Esta ferramenta apresenta um conjunto de funcionalidades para o administrador do ambiente, tais como: gerenciamento de usuários e de grupos de usuários, acompanhamento de uso dos usuários e das seções do *site*, implementação de algumas características relacionadas com performance, com opções de ativação de mecanismos de *cache*, busca de informações, mecanismos de localização durante a navegação, envio de email, entre outras. Como a descrição detalhada deste sistema não é um objetivo deste trabalho, serão descritos apenas os itens de importância para o acompanhamento e avaliação do trabalho realizado.

O conteúdo educacional utilizado no *site* Web apresenta um curso onde os usuários podem aprofundar seus conhecimentos em Mobilidade, estudando as principais

[69] http://www.php.net/

ferramentas envolvidas no processo e a infra-estrutura exigida pela comunicação sem fios e sem fronteiras. O curso está dividido em 10 capítulos, os quais podem ser acessados pelo aluno independentemente de ordem. Cada capítulo está dividido em seções e possui uma lista de questões desenvolvidas para testar o conhecimento adquirido pelo aluno. Além disso, cada módulo pode conter informativos, artigos e links externos, que complementam o aprendizado sobre o assunto em questão.

Na estrutura deste gerenciador de conteúdo Web, seções são áreas compostas por uma ou mais categorias. As categorias, por sua vez, são compostas por um ou mais itens. Itens são os artigos que formam o conteúdo publicado. Na figura 7.17 pode ser observado o menu de criação de itens de conteúdo, sendo exibidas as seções "mobilidade" e "extras", criadas através da ferramenta de administração do gerenciador de conteúdo Web, junto com a opção de adição de categorias para a seção "extras".

Figura 7.17: Menu de administração do conteúdo do gerenciador de conteúdo Web

Dentro da seção "Mobilidade", foram criadas diversas categorias, cada uma correspondendo a um capítulo do curso em questão. Cada capítulo está subdividido em seções, informativos, casos de sucesso e uma lista de questões. Todos foram criados como itens das categorias, o que pode ser visualizado através da Figura 7.18 a seguir. Nesta figura também podem ser observadas as seguintes informações, associadas a cada item de conteúdo:

- Opção de publicação ou não do item;
- Opção para publicação do item na primeira página;
- Ordenação do item dentro da categoria;
- Tipo de acesso (livre ou com usuário registrado);
- Categoria;
- Autor;
- Data de criação.

Figura 7.18: Lista de itens de conteúdo criados

A edição do conteúdo com o gerenciador de conteúdo Web Mambo não requer que o usuário possua conhecimento da linguagem HTML. Todas as informações sobre o conteúdo são cadastradas através de um formulário. As principais são: título, descrição, seção e categoria. As duas últimas informações podem ainda funcionar como metadados. A barra de ferramentas no topo da área de descrição permite a modificação do conteúdo com operações de edição de textos. Estas informações podem ser visualizadas na figura 7.19 a seguir.

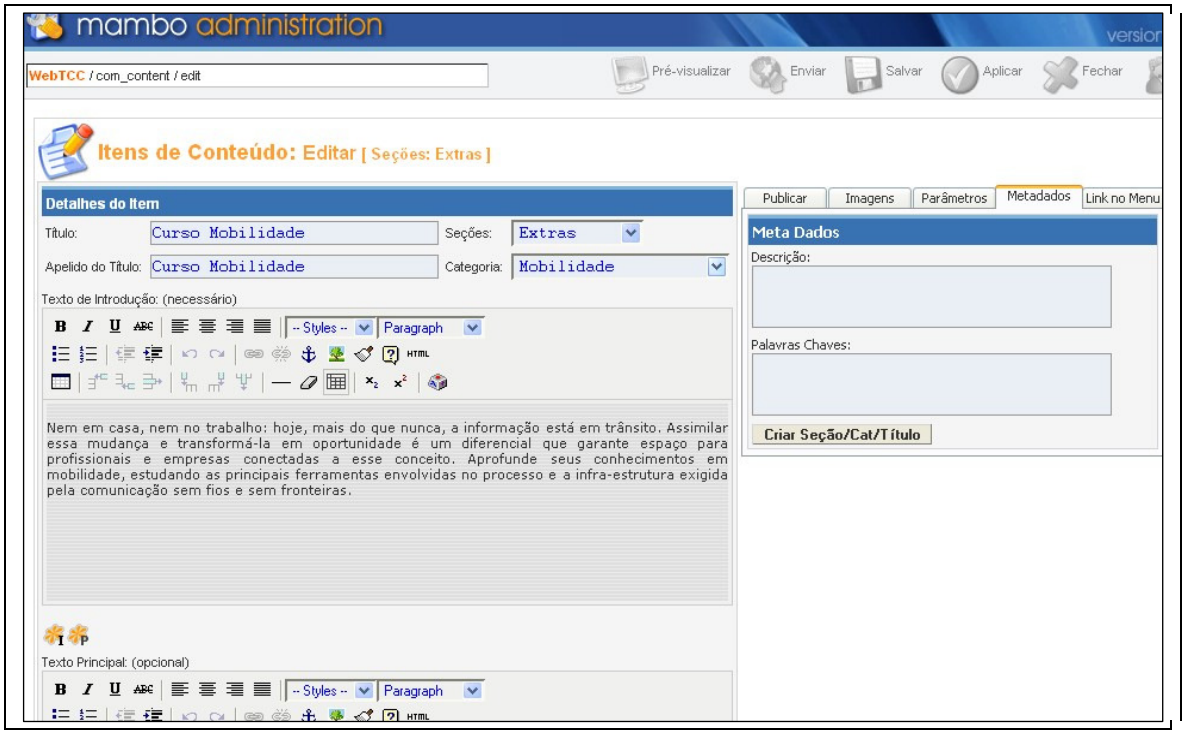


Figura 7.19: Interface de edição de conteúdo Web

Nesta interface de edição de um item de conteúdo ainda há uma série de guias com informações adicionais relacionadas à forma de publicação, data de publicação e retirada do conteúdo, responsáveis pela criação, datas de criação e última modificação, imagens associadas, parâmetros específicos e ainda uma opção para o cadastro de metadados. A cada item de conteúdo pode ser atribuído um status, como publicado, suspenso, em edição e pode ainda ser submetido a um *workflow* de aprovação para publicação, com uso de componentes específicos. Na figura 7.19 pode ser observada a possibilidade de registro de metadados na parte direita da figura.

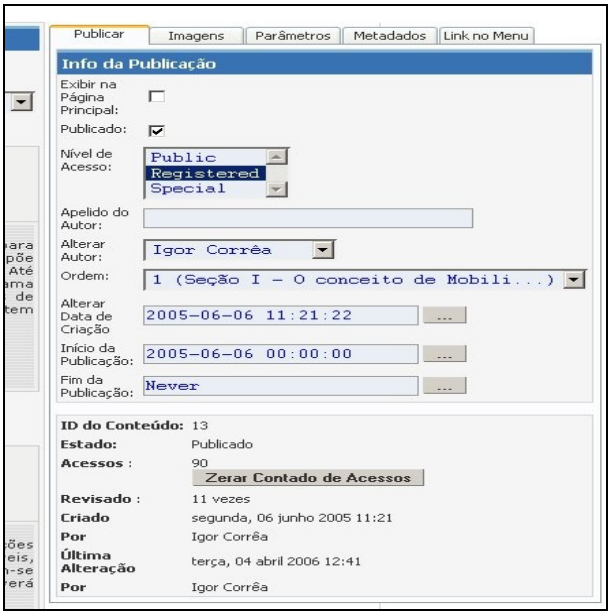


Figura 7.20: Guias com informações adicionais relacionadas ao conteúdo Web

Esta série de guias mencionada pode ser observada com maiores detalhes na figura 7.20 acima, na qual encontra-se selecionada a aba de dados de publicação.

7.4.2.1 Coleta de dados

Algumas estatísticas sobre a navegação do usuário já são disponibilizadas pela ferramenta de administração do Mambo. No entanto, estas informações não são suficientes para acompanhar as escolhas do usuário enquanto acessa as diferentes páginas publicadas. Por esse motivo foi feita uma adaptação no código-fonte do gerenciador de conteúdo Web, sendo acrescentado um módulo para o gerenciamento de um *cookie* e gravação de informações de acesso. Essas informações serão usadas para o processo de Mineração de Uso da Web desenvolvido. Exemplos de informações de uso obtidas pelo gerenciador de conteúdo Web Mambo são descritas a seguir, na figura 7.21, onde podem ser observadas estatísticas sobre os navegadores Web detectados e estatísticas sobre o acesso às páginas publicadas.

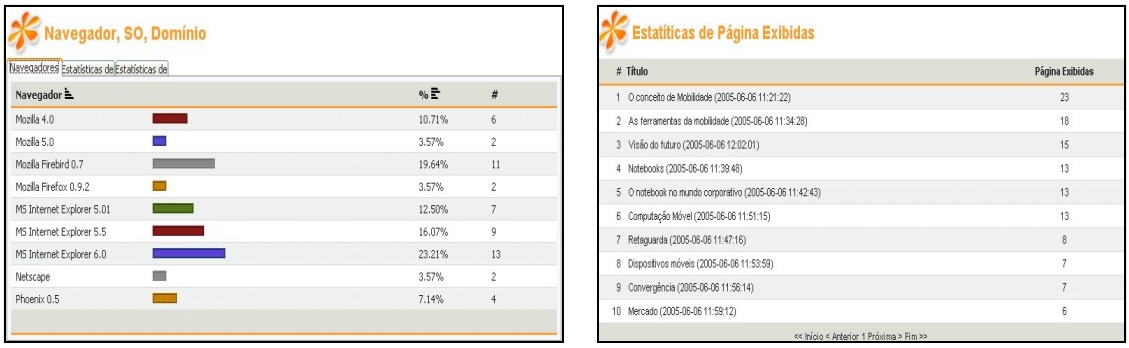


Figura 7.21: Informações de uso padrão: navegadores e páginas acessadas

As informações obtidas com a ferramenta de gerenciamento e publicação de conteúdo Web representam um bom exemplo das informações tradicionalmente disponibilizadas por este tipo de sistema. Elas caracterizam-se como insuficientes para o trabalho desenvolvido, pois não existe a possibilidade de identificação de padrões de acesso gerais e nem de seqüências específicas de páginas acessadas. Assim, foi utilizado um mecanismo específico para obtenção dos dados, implementado com um módulo específico adicionado às páginas geradas pelo *site* Web. O funcionamento deste módulo pode ser resumido nas seguintes etapas:

1. Verificação de existência do *cookie* de identificação;
2. Criação do *cookie* (caso não seja verificada sua existência na máquina do usuário);
3. Armazenamento de informações de acesso em arquivos com formato XML.

As informações armazenadas a cada acesso de um usuário são descritas no item 7.2.2 e resumidas a seguir: número IP[70] (*Internet Protocol*) da origem do acesso, URL[71] (*Universal Resource Locator*) acessada, parâmetros de acesso, navegador utilizado, data e horário de acesso, identificador (*cookie*), identificação de adaptação.

O procedimento de identificação de sessões de usuários será descrito na sessão seguinte.

7.4.2.2 Pré-processamento dos dados de uso Web

Para a geração dos percursos freqüentes, os dados capturados são processados com determinadas heurísticas, de modo a obter seqüências que descrevem o percurso de cada usuário na sua visita ao *site* Web. Estas heurísticas estão relacionadas basicamente com o tratamento de situações ligadas ao uso da Web. Como existe, por exemplo, a possibilidade de um determinado usuário acessar uma determinada página e deixar o navegador com esta página aberta enquanto outras tarefas são realizadas, este fato pode gerar discrepâncias quanto ao tempo de acesso. Para minimizar estas ocorrências são admitidos tempos máximos de permanência a partir dos quais se supõe que a atitude do usuário não é mais a leitura da página.

O tratamento das informações capturadas possibilita a obtenção de percursos, identificados como uma seqüência de acessos a um mesmo *site* que está relacionada a um mesmo identificador (mantido pelo *cookie* descrito anteriormente). Com isso é

---

70 http://www.ietf.org/rfc/rfc0791.txt

71 http://www.ietf.org/rfc/rfc1738.txt

possível, em análise posterior, reunir os percursos idênticos observados. A seguir, na figura 7.22, estão relacionados dois trechos destes resultados, descrevendo um conjunto de percursos gerais e um conjunto de percursos resumidos. Como indicado anteriormente, no item 7.2.2, cada percurso pode ser gerado com a informação da seção do *site* e do tempo de permanência nesta seção. Assim, a figura ilustra um trecho com alguns exemplos obtidos com a monitoração e tratamento dos dados de acesso ao *site* Web utilizado nesta experiência.

```
    Percursos Gerais Obtidos:

Percurso [1] O conceito de Mobilidade (3s) Primeiras iniciativas (172s) Bluetooth
(9s) As ferramentas (113s)

    Percurso [2] O conceito de Mobilidade (42s) As ferramentas (6s) Notebooks (316s)
Computação Móvel (14s) Mercado (8s)

    ...

    Percursos Resumidos:

    Percurso Resumido [1] contagem = 23 Tempo total médio=196s [0] O conceito de
Mobilidade [1] Dispositivos móveis [2] Visão do futuro
```

Figura 7.22: Exemplo da geração de percursos gerais e de percursos resumidos

7.4.2.3Geração de percursos freqüentes

A mineração de padrões seqüenciais trata da análise de dados observados em seqüências de transações. Pode ser utilizada em situações bastante diversas, tais como a análise de dados de DNA, de compras realizadas por clientes e de padrões de dados geofísicos, entre outros exemplos. No presente trabalho está contextualizada em relação à análise de dados obtidos como acompanhamento de acessos a *sites* Web. Assim os padrões seqüenciais tratados representam seqüências de acessos a páginas do *site* Web (ou *page-views*). A partir de um limiar de repetição arbitrário estes padrões são considerados como importantes para o tratamento de adaptações.

A geração de seqüências de acessos freqüentes utilizou módulo desenvolvido com a linguagem C++ e com uso do *parser* Expat[72], como forma de tratar de modo eficiente os dados obtidos. A implementação utilizada está baseada no algoritmo Spade e em melhorias descritas na literatura, conforme indicado anteriormente.

7.4.2.4 Geração de agrupamentos

Os percursos gerais obtidos são utilizados, com métodos específicos para o seu pré-processamento, como dados de entrada para o uso de algoritmos de geração de agrupamentos. A implementação destes algoritmos de agrupamento não é escopo deste trabalho, sendo utilizados a partir do sistema de Mineração de Dados disponibilizado pela Universidade de Waikato (WEKA[73]).

São tratadas as seqüências de percursos gerando-se informações discretizadas com características diversas, tais como o acesso a diferentes páginas, o acesso e o tempo de permanência, ou o acesso, tempo de permanência e número de visitas realizadas à mesma página. O pré-processamento definido permite que sejam indicados os atributos

---

72 http://expat.sourceforge.net/

73 www.cs.waikato.ac.nz/ml/weka/

desejados e assim é gerado um arquivo no formato utilizado como entrada de dados pela ferramenta Weka.

A partir dos percursos realizados é viável a utilização destes algoritmos, de forma a permitir a obtenção de informações possivelmente complementares às informações de percursos freqüentes. Utilizou-se o algoritmo Simple KMeans para a geração do agrupamento. A seguir, na figura 7.23, está ilustrado um exemplo do formato gerado a partir de seqüências de acesso, para uso do algoritmo de geração agrupamentos. Cada seção do *site* analisado consiste em um atributo que, no caso, é descrito com uma informação booleana indicando simplesmente a existência do acesso. Cada linha do conjunto de dados representa uma seção de usuário, contendo, neste exemplo, apenas a informação de acesso às páginas.

```
@relation siteweb
@attribute 15 {TRUE, FALSE}
@attribute 36 {TRUE, FALSE}
.....
@attribute 36 {TRUE, FALSE}
@data
TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Figura 7.23: Arquivo gerado para uso com a ferramenta Weka

A utilização do sistema Weka permite desta forma a geração de agrupamentos, a partir das seções de navegação dos usuários no *site*. A análise e validação dos agrupamentos foi tratada de forma manual. Cada agrupamento descreve um conjunto de páginas acessadas em conjunto e, portanto, representa interesses de grupos de usuários.

7.4.2.5 Ontologia de domínio

Foi empregada uma ontologia de domínio para a verificação da adequação do seu uso como forma de complementar as informações para adaptação da estrutura do *site* Web, juntamente com as informações de percursos freqüentes e de agrupamentos de páginas. Apesar do modelo ser simples, as relações descritas atendem ao objetivo designado para esta experimentação, que seria permitir identificar itens de conteúdo e suas relações com os demais. Na figura 7.24 abaixo está descrito um trecho com o esquema simplificado da ontologia utilizada, na linguagem OWL.

Na ontologia exemplificada parcialmente na figura 7.24 foram descritos os termos importantes do curso. Esta descrição parte do pressuposto de identificação dos itens de conteúdo disponibilizados no *site* Web. São definidos os conceitos “curso”, “tópico” e “tipo_de_recurso” e “idMambo” com o objetivo de permitir a descrição simples da estrutura de um determinado curso. Estes conceitos são relacionados com as seguintes propriedades e relações: “ehRequisitoDe”, “descritoEm”, “composto_por_topico”, “parteDe”, “possuiRequisito”, “descritoPorItem”, “descritoPorItemMambo”. A relação “ehRequisitoDe” possibilita a indicação de itens que representam requisitos para um determinado item de conteúdo. Esta relação está associada de forma inversa com a relação “possuiRequisito”. A relação “composto_por_topico” indica relações de composição que possibilitam descrever itens de conteúdo que integram uma unidade maior. Esta relação é associada como inversa da relação “parteDe”. Já “descritoPorItem” e “descritoPorItemMambo” são utilizadas para

a anotação semântica dos conteúdos, possibilitando que cada item de conteúdo publicado no sistema gerenciador de conteúdo Web tenha uma representação em uma instância da ontologia.

```
<owl:Class rdf:ID="tipo_de_recurso"/>
 <owl:Class rdf:ID="topico"/>
 <owl:Class rdf:ID="curso"/>
 <owl:Class rdf:ID="idMambo"/>
 <owl:ObjectProperty rdf:ID="ehRequisitoDe">
  <rdfs:domain rdf:resource="#topico"/>
  <rdfs:range rdf:resource="#topico"/>
 </owl:ObjectProperty>
 <owl:ObjectProperty rdf:ID="descritoEm">
  <rdfs:domain rdf:resource="#topico"/>
  <rdfs:range rdf:resource="#idMambo"/>
 </owl:ObjectProperty>
 <owl:ObjectProperty
          rdf:ID="composto_por_topico">
   <owl:inverseOf>
     <owl:ObjectProperty rdf:ID="parteDe"/>
   </owl:inverseOf>
   <rdfs:domain rdf:resource="#curso"/>
   <rdfs:range rdf:resource="#topico"/>
 </owl:ObjectProperty>
 <owl:ObjectProperty rdf:ID="possuiRequisito">
  <rdfs:domain rdf:resource="#topico"/>
  <rdfs:range rdf:resource="#topico"/>
 </owl:ObjectProperty>
 <owl:ObjectProperty rdf:about="#parteDe">
  <rdfs:domain rdf:resource="#topico"/>
  <owl:inverseOf
     rdf:resource="#composto_por_topico"/>
  <rdfs:range rdf:resource="#curso"/>
 </owl:ObjectProperty>
 <owl:DatatypeProperty
      rdf:ID="descritoPorItemMambo">
  <rdfs:domain rdf:resource="#topico"/>
  <rdfs:range
    rdf:resource=".../XMLSchema#string"/>
 </owl:DatatypeProperty>
```

Figura 7.24: Trecho da ontologia de domínio utilizada

Um exemplo de uso para a descrição de conteúdos e relações entre os componentes é indicado na figura 7.25 a seguir, onde podem ser observadas as relações de composição, indicação de requisitos e associação de itens de conteúdo cadastrados no gerenciador de conteúdo Web utilizado (Mambo). Nela pode ser observada, por exemplo, a indicação de que o capítulo 2 possui como requisito o capítulo 1 e que está descrito no item “I16”.

```
<curso rdf:ID="mobilidade">
....
   <composto_por_topico>
    <topico rdf:ID="capitulo02_ferramentas">
     <possuiRequisito>
      <topico rdf:ID="capitulo01_conceito">
       <descritoEm>
        <idMambo rdf:ID="I16"/>
       </descritoEm>
       <descritoPorItemMambo rdf:datatype="http://.../XMLSchema#string">
                          16</descritoPorItemMambo>
       <parteDe rdf:resource="#mobilidade"/>
       <ehRequisitoDe rdf:resource="#capitulo02_ferramentas"/>
      </topico>
     </possuiRequisito>
     <descritoEm>
```

Figura 7.25: Exemplo de instância da ontologia de domínio utilizada

As consultas à ontologia foram possíveis com o uso da linguagem RDQL, para identificar itens de interesse. Para as consultas foi utilizada a biblioteca RAP[74], que implementa módulos necessários para o uso da linguagem RDQL. A ontologia foi criada e mantida com uso da ferramenta Protégé[75]. A seguir encontra-se um exemplo de busca de itens relacionados, usando a ontologia descrita e a linguagem RDQL.

```
SELECT ?x  WHERE (v:'.$capitulo.', v:ehRequisitoDe, ?x)
  USING v FOR <http://www.site.com/tcc/logs/curso.owl#>';
```

Figura 7.26: Trecho de consulta em RDQL

Na figura 7.26 a variável "$capitulo" representa o capítulo acessado pelo usuário. O termo "v:ehRequisitoDe" representa as propriedades contidas na ontologia que identifica itens que são requisitos do item atual. A consulta retorna como resultado os itens relacionadas por esta propriedade com o item atual. Este exemplo de consulta foi utilizado na experimentação de possibilidades adicionais de adaptação providas pela ontologia.

7.4.2.6 Adaptação de estrutura do site Web

Neste experimento foram utilizadas as seguintes possibilidades de adaptação: percursos freqüentes de acesso, agrupamentos de percursos freqüentes e relações descritas na ontologia. No caso dos percursos freqüentes foi considerado um mecanismo de identificação do percurso que mais se aproxima do percurso observado na sessão atual de um usuário. Com isso as páginas relacionadas no percurso freqüente padrão são indicadas como sugestão de navegação. Estas páginas são sugeridas abaixo do menu de navegação, localizado na porção esquerda da interface disponibilizada.

Os agrupamentos possibilitam a identificação de grupos de páginas acessadas freqüentemente em conjunto, porém sem necessariamente uma manutenção da mesma ordem de acesso. Assim indicam assuntos de interesse para determinado perfil de usuários. Foram utilizados para adaptação como sugestão destes assuntos de interesse. Estas páginas são sugeridas na parte central da interface, abaixo do conteúdo correspondente ao item selecionado. Já as relações descritas na ontologia são utilizadas para a sugestão de adaptações com base no item acessado e suas relações de possível interesse, como por exemplo, as relações que indicam pré-requisitos entre os conteúdos.

A adaptação de estrutura do *site* Web pode ser compreendida como a etapa final de um processo maior descrito na figura 7.27. Nela podem ser identificadas as etapas anteriores, onde são coletados os dados de uso, realizadas as tarefas de pré-processamento, utilização dos mecanismos de geração de percursos freqüentes e de geração dos agrupamentos, organização dos resultados de forma que possam ser utilizados pelo mecanismo de adaptação. Também podem ser identificadas a criação e uso da ontologia de domínio do *site*, juntamente com a informação de acesso atual gerado pelo usuário.

---

74 http://sourceforge.net/projects/rdfapi-php/

75 http://protege.stanford.edu/

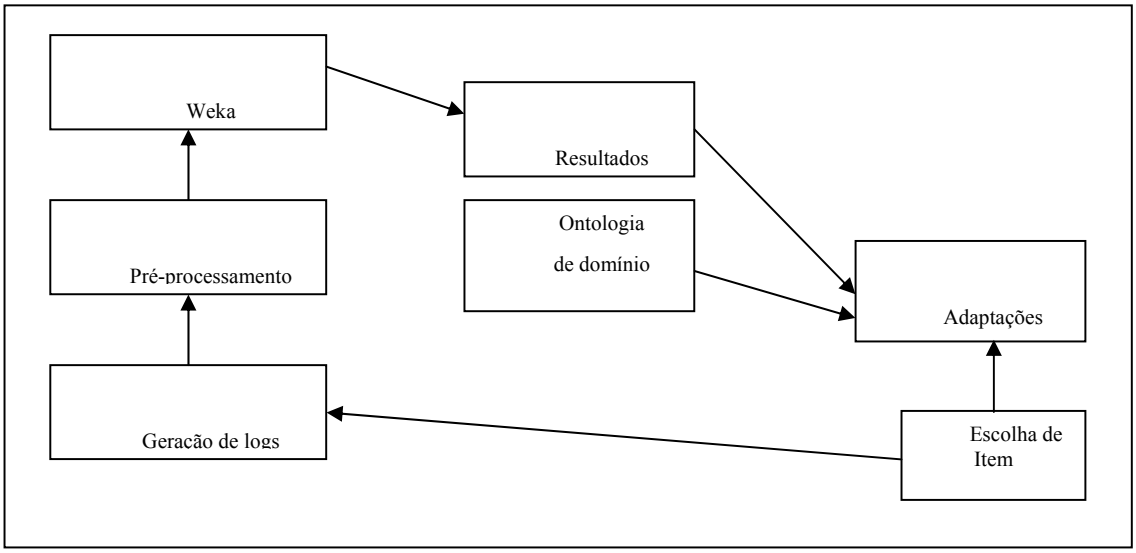


Figura 7.27: Processo geral de acompanhamento de uso e adaptação

A utilização deste mecanismo de adaptação junto ao sistema de gerenciamento de conteúdo Web é possível, pois o mesmo é disponibilizado como ferramenta de código aberto. Sendo assim, foi realizada a inserção do módulo de código que implementa o tratamento da etapa de adaptação.

A marcação para as áreas de adaptação deste experimento podem ser visualizados na figura a seguir (7.28). No menu ao lado esquerdo da página está prevista a inserção dos resultados obtidos com a lista de percursos freqüentes. Para isso, são exibidos nesta área os itens do percurso freqüente para com o qual a sessão atual apresenta similaridade. Abaixo do conteúdo central pode ser vista uma área reservada para a indicação dos itens selecionados pelo mecanismo de agrupamento e logo abaixo deste, um espaço reservado para os resultados obtidos com a consulta à ontologia, que retornará neste caso as relações do tipo requisito.

Figura 7.28: Formato geral dos resultados de adaptação

7.4.3 Descrição dos experimentos com integração semântica

Além do experimento indicado no item 7.4.2, também foram executados outros experimentos na mesma área, porém com características diversas. A seguir estão descritas estas experiências. Os objetivos principais do experimento anterior foram a aquisição e tratamento de dados de uso Web, a geração de padrões de percursos freqüentes e sua utilização em tarefas de adaptação. Nos experimentos aqui descritos o foco principal foi a integração de informações semânticas com informações de uso e a geração de regras de adaptação. Além disso, estes experimentos serviram para indicar a possibilidade de aplicação da metodologia desenvolvida em diferentes contextos de aplicações Web.

Em todos os experimentos realizados foram adotadas as mesmas etapas gerais, descritas na metodologia proposta. Estes experimentos trataram da implementação de *sites* voltados para a divulgação de material educacional, durante períodos variados. Cada uma das experiências levou em conta a descrição de uma ontologia contendo a identificação de conceitos e relações de interesse para possibilitar a anotação semântica dos conteúdos publicados. Estas ontologias foram descritas com o software Protégé e exportadas em arquivos segundo a linguagem OWL. A criação das instâncias das ontologias foi realizada neste editor, com base nas informações de conteúdo geradas para cada caso. Os conteúdos foram publicados utilizando-se o sistema gerenciador de conteúdo Web Joomla[76]. Este sistema é disponibilizado sob licença GPL[77] e utiliza a linguagem PHP e banco de dados Mysql. Desta forma existem algumas facilidades para a inclusão do mecanismo de aquisição dos dados de uso e também para a realização de ações de adaptação, pois a forma de geração dos resultados neste ambiente permite o tratamento diferenciado de formatos gerais de apresentação.

A seguir são apresentadas algumas informações sumarizadas sobre os experimentos realizados.

- Experimento: disciplina de Computação Gráfica II, curso de Graduação em Comunicação Digital.
  - Período: agosto de 2006 até dezembro de 2006.
  - Principais conceitos da ontologia de domínio: tópicos, itens e tipos de recursos.
  - Número de instâncias da ontologia de domínio: 8 tópicos, 19 itens, 7 tipos de recurso.
  - Número de acessos tratados: 1.889 acessos.
  - Número de percursos identificados: 122 percursos freqüentes.
  - Número de contextos semânticos identificados: 67 contextos.
- Experimento: disciplina de Banco de Dados, curso de Graduação em Comunicação Digital.
  - Período: agosto de 2007 até dezembro de 2007.
  - Principais conceitos da ontologia de domínio: tópicos.
  - Número de instâncias da ontologia de domínio: 24 tópicos.
  - Número de acessos tratados: 1.308 acessos.
  - Número de percursos identificados: 179 percursos freqüentes.

---

[76] http://www.joomla.org/

[77] http://www.gnu.org/copyleft/gpl.html

    - Número de contextos semânticos identificados: 84 contextos.
- Experimento: disciplina de Algoritmos e Programação em Linguagem C++, curso de Graduação Desenvolvimento de Jogos Digitais.
    - Período: julho de 2007 até dezembro de 2007.
    - Principais conceitos da ontologia de domínio: tópicos, tipo de material, tipo de conteúdo.
    - Número de instâncias da ontologia de domínio: 24 tópicos.
    - Número de acessos tratados: 12.729 acessos.
    - Número de percursos identificados: 103 percursos freqüentes.
    - Número de contextos semânticos identificados: 34 contextos.
- Experimento: disciplina de Algoritmos e Programação em Linguagem C++, curso de Graduação Desenvolvimento de Jogos Digitais.
    - Período: fevereiro de 2008 até abril de 2008.
    - Principais conceitos da ontologia de domínio: tópicos, tipo de material, tipo de conteúdo.
    - Número de instâncias da ontologia de domínio: 53 itens, 6 tipos de material, 11 tipos de conteúdo.
    - Número de acessos tratados: 4.343 acessos.
    - Número de percursos identificados: 155 percursos freqüentes.
    - Número de contextos semânticos identificados: 47 contextos.

Para estes experimentos também foram acompanhadas as escolhas de links sugeridos como adaptação. Assim foi possível traçar alguns comparativos entre o número de links gerados automaticamente acessados. Observa-se que este número é bastante significativo nos casos de ambientes que já possuem um histórico de acompanhamento e padrões de percurso freqüentes e contextos semânticos gerados. Nos dois últimos casos acompanhados observa-se que à medida em que o tempo de utilização avança, o percentual de links adaptados acessados passa de um percentual de 7% para um percentual de 23%, indicando assim uma adequação à possíveis interesses dos usuários.

A seguir será descrito em maiores detalhes o último experimento, sendo que o formato geral seguido pelos experimentos anteriores é o mesmo. O processo envolve inicialmente a descrição da estrutura de um *site* Web contendo o material instrucional ser publicado. Em geral este material é acompanhado também por itens de conteúdo que servem ao propósito de apoio em necessidades complementares. Nestes casos estudados exemplos típicos deste material de apoio são: a apresentação do *site* e da disciplina, a descrição de crongrama de atividades, enunciados de atividades, listas de material de leitura complementar, listas de URLs com sugestões de ferramentas de software ou outros aplicativos de apoio.

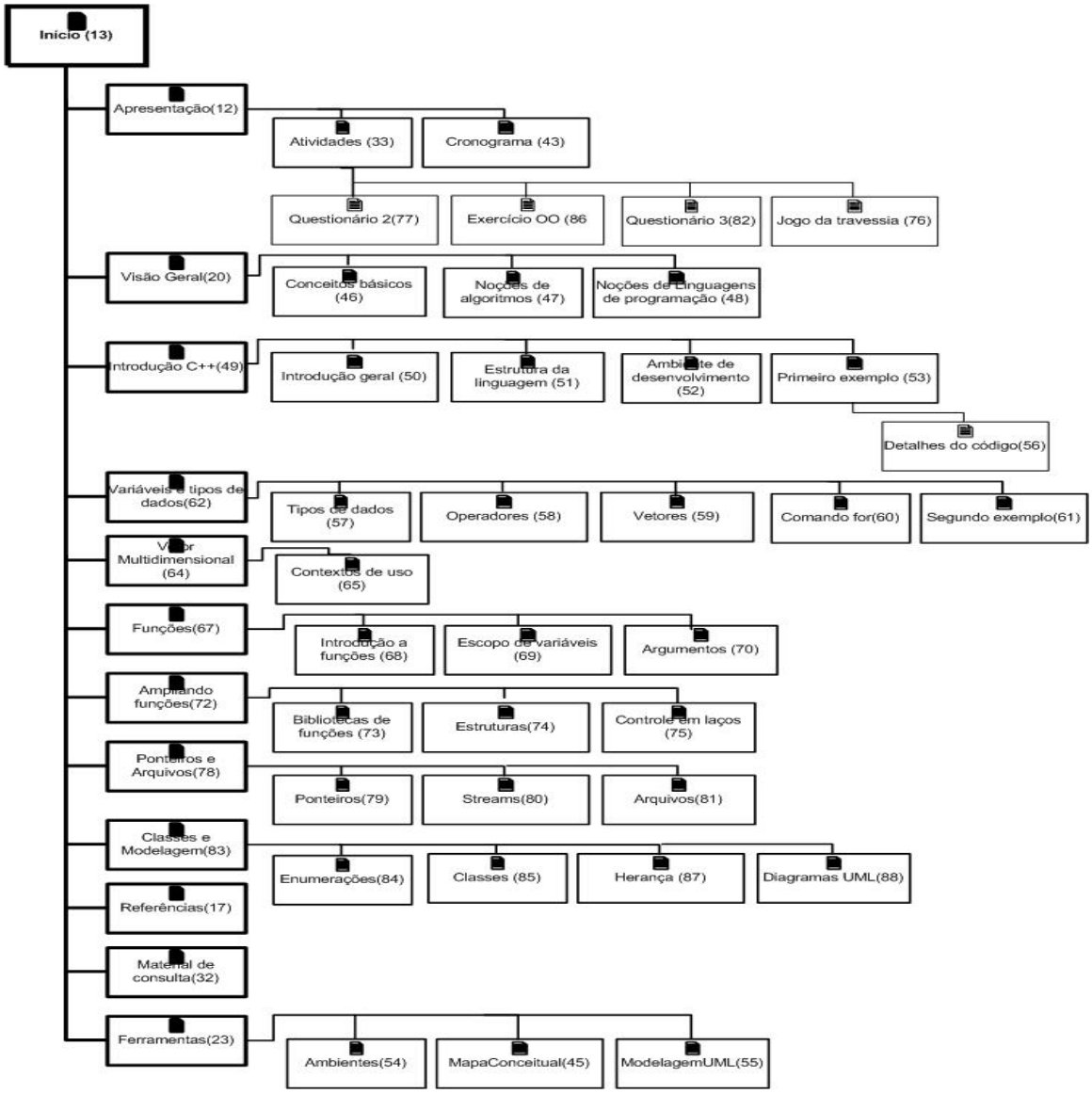


Figura 7.29: Estrutura geral do site Web de um experimento

A figura 7.29 ilustra a estrutura detalhada do *site* Web último experimento realizado. Ela consiste de um mapa de componentes resumido, no qual cada um dos itens é identificado pelo título de um elemento de conteúdo publicado no sistema de gerencimento Web. O elemento "Início" identifica a página inicial ("*Home*") do ambiente montado. Cada um dos itens ligados diretamente ao item "Início", que formam a primeira coluna da figura, corresponde a um item do menu principal do ambiente. Os elementos ligados a cada um destes itens de menu correspondem às opções para cada entrada do menu. Em alguns casos existem subníveis de navegação, como no caso dos itens "Atividades (33)" ou "Primeiro exemplo" (53). Deve ser ressaltado que nem todos os itens possuem o mesmo objetivo, sendo que alguns são material contendo texto ou documentos de aula, outros contém exercícios, leituras complementares, material de apoio ou ainda são considerados como elementos de apresentação do ambiente.

As informações publicadas no sistema gerenciador de conteúdo Web, que são a origem desta organização descrita na figura 7.29, podem ser representadas na ontologia de domínio utilizada nestes experimentos para promover a integração de informações de uso e de informações semânticas. As relações mais claramente identificadas na descrição desta figura são as informações de estrutura, que possibilitam a organização das seqüências sugeridas de conteúdos e também os agrupamentos possíveis. Além desta pode ser relacionada a diferenciação dos itens de acordo com sua função no ambiente, que permite organizar os itens de conteúdo em tópicos como "material de aula", "material de apoio" e "apresentação", por exemplo. Outra possibilidade é a indicação de relações do tipo "pré-e requisitos" ou "complemento" entre itens de conteúdo que não estejam diretamente associados com as relações estruturais. Por fim, os conteúdos principais de cada item publicado podem ser anotados em relações específicas, que posteriomente podem ser utilizadas em tarefas de mineração.

A ontologia descrita para estes experimentos levou em conta este cenário geral. Assim foram definidas as classes (conceitos) "curso", "topico", "tipo_de _material" e "tipo_de_conteudo". Cada elemento de conteúdo publicado corresponde a uma instância do tipo "tópico" e deve estar associado a um tipo de conteúdo e um tipo de material. A figura 7.30 ilustra um resumo das principais classes da ontologia e suas relações.

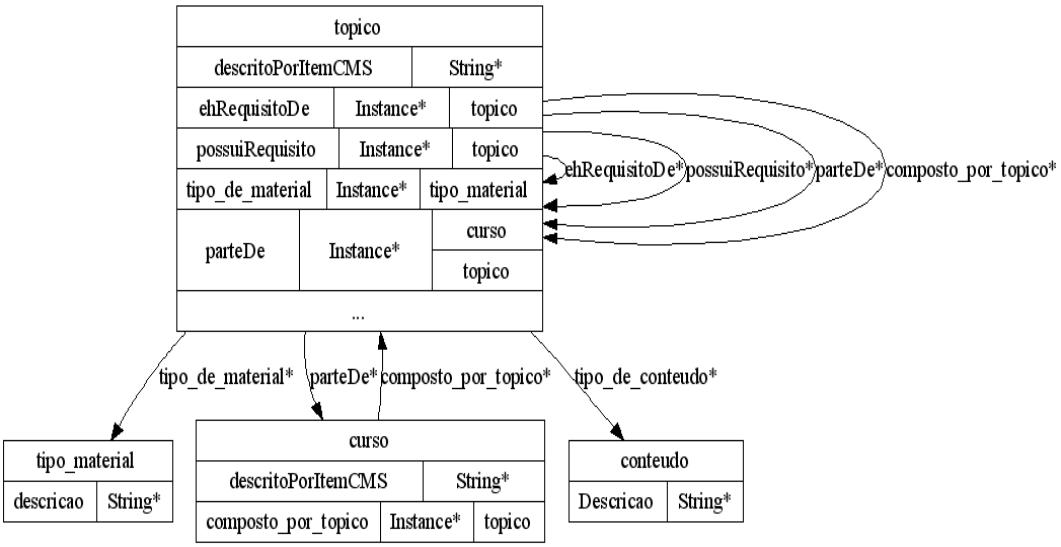


Figura 7.30: Trecho da ontologia utilizada para experimentação

Um pequeno trecho da ontolgia contendo as instâncias de uma parte dos elementos do *site* Web pode ser vista na figura 7.31 a seguir. Nela estão exemplificadas as relações estabelecidas entre o item de código 12 ("Apresentação") e os itens de código 20 ("Visão Geral") e 33 ("Atividades"). Pode-se observar que estes três itens são do tipo "Apoio a Organização", indicando itens de conteúdo de suporte ao ambiente. Estes disponibilizam mensagens gerais de apresentação da disciplina, uma visão geral dos conteúdos e um panorama das atividades. As relações destacadas entre eles permitem a detecção de uma diferença entre os mesmos, dado que entre os itens 12 e 33 existe uma relação do tipo "ParteDe", que indica composição. Entre os itens 12 e 20 existe uma relação de requisito ("ehRequisitoDe").

Ainda na figura 7.31, a partir do item 20 está indicada uma relação de composição com o item 47 ("Noções de Algoritmos") que corresponde a um item do tipo "Conteúdo

de aula". O item 33, por sua vez, está relacionado com diversos outros itens, como o 77 ("Questionário 2"), 86 ("Exercício OO") e 76 ("Jodo da travessia"), todos do tipo "Apoio a organização". Neste caso também pode ser observada a relação de composição entre os elementos. Deste modo, neste exemplo pode ser identificado que existem dois conjuntos de itens de conteúdo associados por relações de composição ("parteDe"). O primeiro ocorre com os itens 20 e 47. O segundo ocorre com os itens 33 e outros três itens (77, 86 e 76). Também pode ser observada na ontologia a relação de composição entre o item 12 e o item 33, que pode levar à conclusão de composição também entre os itens anteriormente citados (77, 86 e 76) para com o item 12.

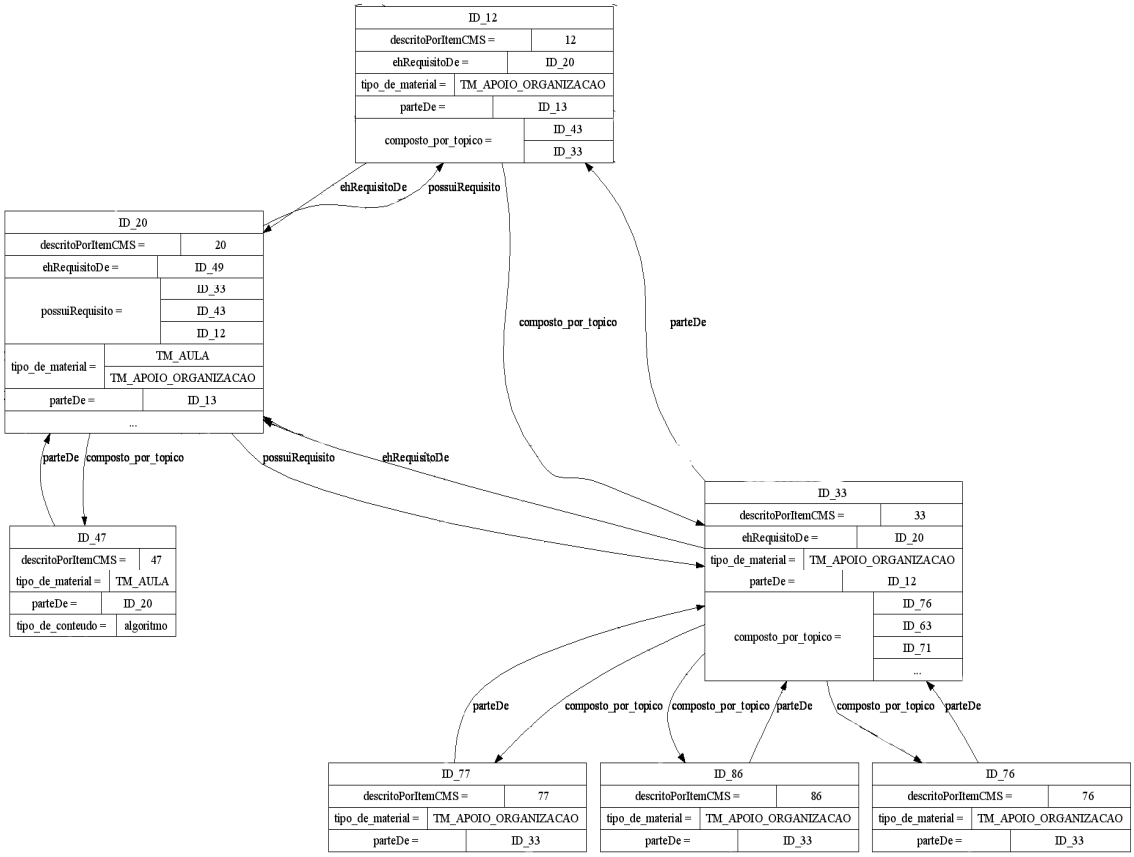


Figura 7.31: Exemplo de algumas instâncias e suas relações

Tendo como base estas informações de publicação de conteúdos e a anotação semântica na ontologia, o experimento delimitou um período inicial de utilização do ambiente, no qual os dados são apenas coletados e não ocorrem adaptações. Após este período são gerados os percursos freqüentes que serão integrados com as informações semânticas descritas na ontologia. Conforme indicado anteriormente, cada item de conteúdo publicado recebe uma identificação única no sistema de gerenciamento de conteúdo Web. Esta identificação é utilizada para a montagem das requisições HTTP durante a navegação dos usuários. O mecanismo de captura de dados de acesso, conforme indicado no item 7.2.2, armazena estas referências, junto com outros dados, tais como o número IP, data e hora de acesso e identificação do *cookie*. Durante a geração dos percursos freqüentes são empregados os códigos de acesso na descrição de cada padrão. Por fim, durante a integração de padrões de uso do *site* e informações

semânticas, estes códigos são utilizados nas consultas feitas à ontologia, tal como definido no item 7.2.3.

Cada percurso gerado possui um formato simplificado no momento em que é realizada a integração com informações semânticas. Assim os elementos do percurso identificam apenas a ocorrência daquela seqüência de acesso indicada pelos itens do percurso. Entretanto, durante a geração dos percursos, são obtidas algumas outras informações que podem ser utilizadas em tarefas de qualificação destes. Por exemplo, podem ser obtidas as informações de número de acessos a um item do percurso e também são gerados os tempos médios de permanência em cada item do percurso. Estas informações podem ser empregadas como atributos em alguns mecanismos de delimitação de pesos para os acessos e podem auxiliar métricas diversas.

Tabela 7.1: Relações de padrões de acesso e contextos semânticos

N	Padrões acesso	Contextos semânticos
1	13, 12, 33	(composto_por_topico) (ehRequisitoDe) [type - topico] [tipo_de_material - TM_APOIO_ORGANIZACAO] (composto_por_topico) [type-topico] [tipo_de_material-TM_APOIO_ORGANIZACAO]
2	12, 33, 44	(composto_por_topico) [type - topico] [tipo_de_material - TM_APOIO_ORGANIZACAO] (composto_por_topico) [type - topico] [tipo_de_material - TM_APOIO_ORGANIZACAO]
3	49, 50, 51	(composto_por_topico) [type - topico] [type - topico] [parteDe - ID_49] [tipo_de_material - TM_AULA] [tipo_de_conteudo - linguagem_de_programacao]
4	20, 46, 47, 48	(composto_por_topico) [type - topico] [tipo_de_material - TM_AULA] [type - topico] [parteDe - ID_20] [tipo_de_material - TM_AULA] [tipo_de_conteudo - algoritmo] [type - topico] [parteDe - ID_20] [tipo_de_material - TM_AULA] [tipo_de_conteudo - linguagem_de_programacao]
5	49, 50, 51, 52	(composto_por_topico) [type - topico] [type - topico] [parteDe - ID_49] [tipo_de_material - TM_AULA] [tipo_de_conteudo - linguagem_de_programacao] [type - topico] [parteDe - ID_49] [tipo_de_material - TM_AULA] [tipo_de_conteudo - linguagem_de_programacao]
6	62, 64, 67, 72	(ehRequisitoDe) [parteDe - ID_13] [tipo_de_material - TM_APOIO_ORGANIZACAO] (ehRequisitoDe) [parteDe - ID_13] [tipo_de_material - TM_APOIO_ORGANIZACAO] (ehRequisitoDe) [parteDe - ID_13] [tipo_de_material - TM_APOIO_ORGANIZACAO] [ehRequisitoDe - ID_78]

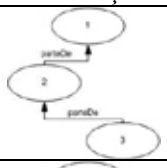
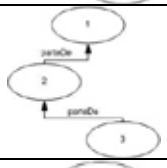
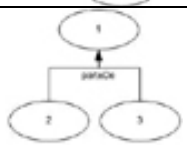
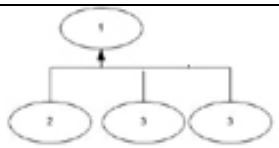
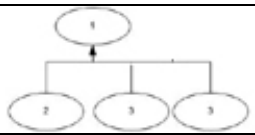
Após a obtenção dos percursos freqüentes de acesso, quanto estes já passaram por uma análise automática através do cálculo de suporte, o mecanismo de integração desenvolvido possui à sua disposição um conjunto de tuplas nas quais estão indicados os códigos dos elementos de conteúdo e sua seqüência de ocorrência em cada padrão. Seguem abaixo, na tabela 7.1 alguns exemplos destas tuplas, obtidas nas experimentações realizadas. Nela estão dispostos alguns padrões de acesso e os respectivos contextos semânticos, obtidos com a integração da informação de seqüência de acesso com as informações da ontologia de domínio. Em cada contexto estão relacionados os itens de conteúdo entre si. Assim, na linha 1 está colocada a informação de que o item 13 é composto pelo item 12, que o item 13 é requisito do item 12, que estes dois itens são elementos do tipo "tópico" e do tipo de material de "apoio à organização do *site* Web". Nesta mesma linha está colocada a informação de que o item 12 é composto pelo tópico 33, que estes dois itens são elementos do tipo "tópico" e do tipo de material de "apoio à organização do *site* Web".

Uma interpretação dos contextos semânticos indicados na tabela 7.1 pode ser vista na tabela 7.2 abaixo. Na segunda coluna são identificados os itens do percurso e entre eles são descritos arcos indicando relações de composição. Assim é possível obter-se uma imagem que corresponde à organização destes itens na estrutura do *site* Web. Esta avaliação inicial pode indicar já algumas vantagens deste contexto em relação ao padrão

seqüencial apenas. Entretanto, caso as relações utilizadas sejam apenas estas, algumas distorções ocorrerão, como no caso dos padrões 4, 5 e 6, que apresentam semelhanças nesta interpretação inicial, porém possuem função bastante diferenciada no *site*, sendo que os itens do padrão 6 são voltados para o apoio à organização do material, enquanto que os padrões 4 e 5 são voltados ao material de aulas. Também ocorreria uma distorção entre o padrão 6 e o padrão 1, que não apresentam semelhanças na interpretação inicial, baseada nas imagens ilustrativas de cada um. Entretanto, como os dois padrões possuem conteúdo semelhante, o índice de similaridade entre eles é bastante alto. Já em outros casos, como entre os padrões 1 e 2, existe uma semelhança na interpretação visual e também um índice de similaridade alto. O mesmo ocorre entre os padrões 3 e 4, que possuem índice alto de similaridade e semelhanças na interpretação visual, mesmo contendo um número de elementos diferentes.

Com estes exemplos, busca-se ilustrar os ganhos obtidos com a geração do contexto semântico para os padrões de acesso freqüentes. No caso de utilização apenas das informações de acesso, nenhuma destas semelhanças e diferencas poderiam ter sido detectadas. Além disso, esta detecção ocorre sem necessidade de aquisição de dados de usuários identificados e o processo de geração pode ser automatizado, gerando sempre uma atualização de contextos. Estes contextos podem então ser utilizados para a etapa seguinte prevista na metodologia, que é a adaptação de estrutura do *site* Web.

Tabela 7.2: Visualização e interpretação de contextos semânticos

N	Visualização	Interpretação
1		A seqüência de acessos indica três páginas que estão organizadas com a relação “parteDe” entre si.
2		A seqüência de acessos indica uma relação composição semelhante ao padrão de número 1. A similaridade entre eles é alta.
3		A seqüência de acessos indica uma relação de composição entre o primeiro e segundo item. O segundo e terceiro compartilham a relação de composição com o primeiro, indicando um segundo nível de agrupamento de conteúdos. A similaridade entre este padrão e cada um dos dois anteriores é baixa.
4		A seqüência de acessos indica uma relação de composição entre o primeiro e segundo item. O segundo, terceiro e quarto compartilham a relação de composição com o primeiro, indicando um segundo nível de agrupamento de conteúdos. A similaridade entre este padrão e o de número 3 é maior do que este e os de número 1 e 2.
5		Semelhante ao padrão anterior. Similaridade muito alta entre este padrão e o padrão 3.
6		Padrão que indica navegação em um primeiro nível de conteúdos. Possui similaridade baixa com os padrões 3, 4 e 5, em função do tipo de material.

A etapa de adaptação da estrutura do site Web utiliza os contextos semânticos em conjunto com a identificação dos acessos de um determinado usuário durante sua sessão. Esta identificação é obtida pela utilização de um *cookie* que armazena um identificador gerado aleatoriamente quando é detectado um novo acesso. Assim não existe necessidade de identificação pessoal do usuário do *site* Web. Para a geração de alternativas são utilizadas as relações da ontologia e o contexto semântico que for identificado como semelhante à sessão do usuário. Assim podem ser utilizadas relações como requisitos de conteúdos, seqüências de acesso freqüente ou relações complementares às relações observadas no perfil semântico.

No experimento realizado foram utilizadas duas formas de modificações na estrutura do *site* Web. Na primeira são identificadas relações de interesse na ontologia, tais como relações entre itens de conteúdo que possuem pré-requisitos para com o item de conteúdo sendo acesado. Este tipo de adaptação é gerado na parte central do conteúdo, abaixo do texto principal exibido a cada acesso. A figura 7.32 ilustra esta situação. Nela pode ser observado que o item atualmente selecionado "Ponteiros e Arquivos") possui a indicação de que existem dois itens que são pré-requisitos, indicados na parte central do texto ("Funções" e "Ampliando funções").

Figura 7.32: Exemplo de adaptação de estrutura (área de conteúdo)

Na segunda forma são verificadas diversas possibilidades e identificadas as relações mais adequadas. Esta adaptação gera seus resultados na área de menu de opções do *site* Web, como pode ser visto no exemplo da figura 7.33 abaixo. As relações predominantes no contexto semântico identificado como similar à sessão do usuário são utilizadas em conjunto com as informações de seqüência de acesso. Neste caso exemplificado, o item atualmente selecionado ("Referências") está relacionado com os dois itens indicados na parte inferior do menu à esquerda da imagem ("Material" e "Ferramentas").

Figura 7.33: Exemplo de adaptação de estrutura (área de menu)

7.5 Arquitetura baseada em Web Semântica

No capítulo 6 deste trabalho foram rapidamente comentados alguns métodos para o uso de recursos da Web semântica na descrição de aplicações Web, sendo possível assim a utilização de modelos diversos tais como modelo de domínio, de navegação ou de adaptação. Os trabalhos SHDM (*Semantic Hypermedia Design Method*) (Lima, 2003) e sua expansão ASHDM (*Adaptive SHDM*) (Assis, 2005), são utilizados como referência para a continuação deste trabalho, com a descrição de uma aplicação adaptativa onde o modelo de domínio e o modelo de navegação permitam a melhoria da proposta de integração de informações de uso com informações semânticas como forma de geração de recursos para adaptações. O próprio modelo indicado possui trabalhos associados no sentido de implementação de protótipos de aplicações, como em Szundy et al. (2004), com o uso de RDF e da bilioteca Jena[78], ou em Nunes e Schwabe (2006), onde é descrito um ambiente de prototipação rápida combinando projeto orientado a modelos e DSL (*Domain Specific Languages*). Como a descrição do modelo SHDM é feita a partir da linguagem OWL, pode ser utilizado como referência para a descrição dos modelos de domínio, navegação, apresentação e adaptação, utilizadas como base para informações a serem tratadas no processo de atendimento de requisições a uma aplicação de Hipermídia Adaptativa.

Os experimentos descritos no item 7.4 foi realizados utilizado um sistema de gerenciamento de conteúdo Web, para implementação de aplicação para testes da metodologia desenvolvida. Para a proposta de continuidade das experimentações desta tese, foi desenvolvida uma arquitetura voltada para a implementação de aplicações de Hipermídia Adaptativa em um ambiente semântico. Detalhes desta arquitetura e experimentos realizados estão descritos a seguir.

Foi tomada como premissa a geração de adaptações sem a identificação do usuário. Para isso as informações de uso são aproveitadas como base, junto com informações de perfis de grupos de usuários obtidos com o acompanhamento do uso. São empregadas duas ontologias, uma contendo informações de hierarquia das páginas

78 http://jena.sourceforge.net

pertencentes ao *site* Web (Ontologia da Aplicação) e a outra com a definição da interface que cada página pode possuir (Ontologia de Apresentação). O sistema mantém uma base de dados para os conteúdos das páginas, que armazena todos os conteúdos em formato textual, relacionando-os a partir da Ontologia da Aplicação e permitindo representar diferentes conteúdos para uma mesma página, com informações que podem ser usadas na tarefa de adaptação. Todos os acessos dos usuários são mantidos de forma resumida em uma base de dados. As regras de adaptação baseiam-se na utilização do *site* e na estrutura descrita nas ontologias. Elas relacionam informações do perfil do usuário com a estrutura do *site*, indicando possibilidades de adaptação utilizadas em conjunto com informações descritas na Ontologia de Apresentação. As possibilidades de adaptação implementadas foram definidas como acréscimos à estrutura do *site,* na forma de *hyperlinks* adicionais, e como ajustes de conteúdos, na forma de exibição de material adicional.

Para proporcionar flexibilidade na implementação da arquitetura, a mesma foi separada em dois módulos: validação e adaptação. Seu funcionamento geral, descrito na figura 7.34, pode ser resumido como: recebimento de solicitação de acesso e gravação da informação correspondente ao modelo do usuário; consulta da página solicitada na Ontologia da Aplicação; consulta do *template* correspondente na Ontologia de Apresentação; montagem do resultado intermediário (ainda sem adaptação) em código HTML; verificação do padrão comportamental através do modelo do usuário; aplicação de regras de adaptação; geração do código final (com adaptações) em HTML; envio da resposta para o usuário.

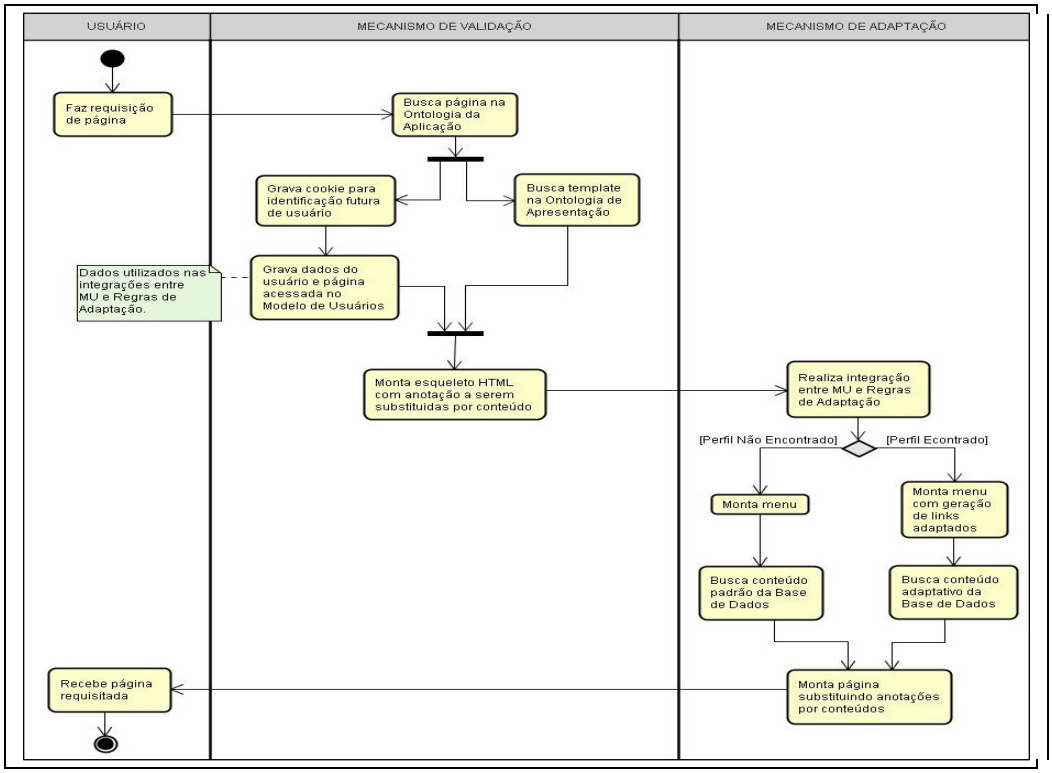


Figura 7.34: Funcionamento geral da arquitetura desenvolvida

O mecanismo de validação é responsável pela montagem inicial do código HTML da página requisitada e pela gravação de informações que possibilitam

adaptações de conteúdo em requisições futuras. Utiliza-se um *cookie* para identificar acessos da sessão de um usuário. Esta abordagem foi escolhida por facilitar a identificação de informações da sessão, apesar de sua desvantagem no caso de desativação junto ao navegador do usuário. Cada acesso gera um registro que compõe o modelo de usuários, contendo dados sobre a navegação, tais como a página acessada e o tipo de conteúdo. Algumas destas informações são geradas pela requisição recebida e outras são complementadas com uma consulta à ontologia da aplicação. Nela estão descritas a estrutura da aplicação e os detalhes de cada página, permitindo a categorização do acesso com informações semânticas como o tipo de conteúdo e as relações desta página com as demais.

Estas ontologias foram descritas manualmente por especialistas no domínio da aplicação, com uso do editor de ontologias Protégé[79], tendo sido utilizada a linguagem OWL[80] para sua representação. Parte da Ontologia da Aplicação está descrita brevemente na figura 7.35, na qual podem ser observadas algumas relações, como a "possuiSubpagina". Além destas relações são mantidos alguns atributos definidos nesta ontologia, como "tipoConteudo" e "template", e outros reutilizados de outras terminologias conhecidas, como o Dublin Core[81], no caso de "dc:description" e "dc:creator".

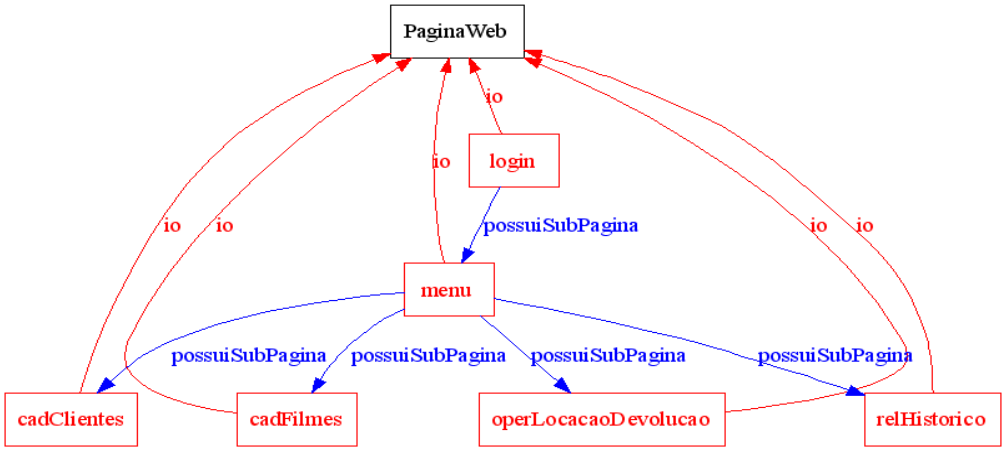


Figura 7.35: Trecho da estrutura da Ontologia da Aplicação

A Ontologia de Apresentação descreve características da interface a ser definida. Cada elemento desta ontologia indica especificamente um tipo de formato, no elemento "template". Estes formatos são descritos, na ontologia de apresentação, a partir dos conceitos "template", "elemento" e "sub-elemento", que permitem a descrição de dados da página e possibilitam a geração do código HTML básico com anotações de conteúdo para substituições. Nas seções onde devem ser feitas substituições por conteúdos adaptados são utilizados marcadores específicos. A figura 7.36 ilustra a descrição de uma típica área de interface, descrita como "TPL01conteudo" e composta por elementos com informações para a geração do HTML ("tag") e informações para substituição ("swah:text"). Estas informações são utilizadas em conjunto com outros elementos da codificação padrão necessária para a montagem de uma página em HTML.

---

79 http://protege.stanford.edu

80 http:// www.w3.org/2004/OWL

81 http://dublincore.org/

```
<TPL01_conteudo rdf:ID="TPL01conteudo">
      <ordem rdf:datatype="...#int">3</ordem>
      <possuiElemento>
            <Elemento rdf:ID="TPL01conteudo_div">
                  <tag rdf:datatype="...#string">div</tag>
                  <possuiSubElemento>
                        <SubElemento rdf:ID="TPL01conteudo_pre">
                        <tag rdf:datatype="...#string">swah:text</tag>
                        </SubElemento>
                  </possuiSubElemento>
                  <class rdf:datatype="... #string">TPL01conteudo</class>
                  </Elemento>
      </possuiElemento>
</TPL01_conteudo>
```

Figura 7.36: Parte da ontologia de apresentação

O mecanismo de adaptação realiza a análise da codificação HTML gerada pelo mecanismo de validação e substitui suas anotação por conteúdos. Para determinar os conteúdos correspondentes a cada anotação é feita uma análise entre o modelo de usuário e as regras de adaptação, em busca de um padrão adequado. As regras são descritas a partir de condições e ações. As condições correspondem às informações do comportamento percebido do usuário e as ações correspondem à utilização do conteúdo para substituições. Estas regras são geradas pelo acompanhamento do comportamento dos usuários do site Web.

Foram definidos três tipos de regras, para o menu de navegação, associadas ao perfil do usuário e ao conteúdo. As regras de menu são as mais simples e são compostas de uma condição indicando a página acessada e diversas ações, que correspondem às páginas sugeridas em hyperlinks junto ao menu gerado pela aplicação. As regras associadas ao perfil do usuário são descritas com condições onde são indicadas as páginas mais acessadas por determinado grupo de usuários. As regras de conteúdo integram informações do perfil do usuário, página acessada e conteúdo a ser adaptado. São compostas de condições associadas ao tipo de perfil e à página acessada. As ações são associadas às áreas da interface que possibilitam adaptação, como por exemplo, o topo da página, o conteúdo principal e a área destinada a conteúdos complementares. A definição do tipo de perfil do usuário leva em consideração o número de vezes que cada página associada às regras que definem este perfil foi acessada. A geração do perfil pode ser feita de forma manual, com a edição de regras que o definem, ou de forma semi-automática, com a análise das sessões dos usuários. A cada perfil de usuário definido correspondem um grupo com comportamentos similares e são associadas sugestões de hyperlinks para o menu e a substituição de conteúdo por informações adaptativas. Caso nenhuma regra válida seja encontrada, a página a ser apresentada ao usuário será montada com seu conteúdo padrão.

Para facilitar a criação e publicação de páginas por usuário sem conhecimentos em tecnologias da Web Semântica, foi desenvolvida uma ferramenta de administração de conteúdos criados através da arquitetura, denominada SWAH (Semantic Web / Adaptive Hypermedia). A interface disponibiliza telas de cadastros de páginas e inserção de conteúdos, telas para armazenamento de imagens e regras, além de telas de configuração geral dos parâmetros do ambiente. A Figura 7.37 exibe algumas telas do ambiente de administração.

Figura 7.37: Telas da interface SWAH

O item à esquerda da Figura 7.38 apresenta a tela de criação de páginas, que mantém a Ontologia da Aplicação atualizada com cada página atualizada pelo usuário. Nesta tela constam informações de cadastro como nome do arquivo, título da página, autor, descrição e tipo de conteúdo, *template* escolhido, tipo de menu a ser utilizado (adaptativo, inferido ou fixo) e página superior na hierarquia (página "pai"). Após o cadastro de uma página é necessário inserir conteúdo para a mesma. O item à direita da Figura 7.38 mostra a tela de inserção de conteúdos, onde é editado o conteúdo padrão da página, ou seja, conteúdo que é apresentado caso não existam regras de adaptação para a mesma. Nesta tela podem ser observadas as informações gerais obtidas da tela anterior, no cadastro da página, seguida pela área de edição do conteúdo onde está disponível uma barra de ferramentas que busca facilitar a interação entre os usuários e a edição de páginas. No caso da inserção de conteúdo adaptativo, além destas informações é indicado um nome para o tipo de conteúdo da página (que também deverá ser usado na montagem das regras), bem como inserido o conteúdo correspondente. Os dados cadastrados pela interface SWAH são armazenados em um banco de dados MySQL[82], sendo que o acesso às ontologias é realizado com uso da biblioteca RAP[83], sendo utilizada a linguagem PHP 5[84] para a implementação dos sistema.

7.5.1 Experimentação da arquitetura implementada

Para a aplicação das funcionalidades implementadas na arquitetura foi desenvolvido um caso de uso com base nos requisitos para uma tabacaria, possuidora de três categorias de produtos: livros, revistas e DVDs. Devido à facilidade de relacionamento de áreas entre essas categorias, foram feitos testes com regras para sugestão de hyperlinks no menu de páginas distintas e com a substituição ou complementação de conteúdos. Para o teste foram criados quatro perfis de usuários, identificados como adulto, jovem, infantil e técnico. A primeira etapa para a criação do site foi o cadastro de suas páginas, tendo sido criadas dezessete páginas, todas relacionadas às três categorias de produtos citadas. Foram criadas várias subcategorias para distinguir os tipos de produtos, como por exemplo, livros técnicos, revistas

82 http://www.mysql.com/

83 http://sites.wiwiss.fu-berlin.de/suhl/bizer/rdfapi/

84 http://www.php.net/

infantis, DVDs de filmes, entre outros. A seguir foram inseridos os conteúdos das páginas e criados conteúdos adaptativos.

Figura 7.38: Página com conteúdos adaptados

O exemplo apresentado na figura 7.38 é de uma página de revistas técnicas. Essa página teve todas suas áreas adaptadas devido às regras de adaptação e o comportamento do usuário em sua navegação entre as páginas do portal, o que resultou na avaliação de seu perfil como perfil técnico. Segundo as regras criadas para a o *site* de experimentação, um usuário de perfil técnico acessa com maior freqüência categorias como livros de auto-ajuda e livros técnicos, revistas de filmes, jogos e de assuntos técnicos, além de DVDs de jogos. Estas informações podem ser observadas na figura 7.39, onde a regra descrita com ao atributo type=’profile’ e name=’tecnico’ descreve estas possibilidades de adaptação.

As regras cadastradas podem ser verificadas na figura 7.39. É importante observar a indicação dos conteúdos adaptativos nas regras com atributo type=’content’. A regra indica que se o usuário possui o perfil técnico e estiver acessando a página com nome “tabacariaRevistasTecnicas”, o topo deve usar o conteúdo “topRevTecnicasPT”, a área central o conteúdo “dcRevTecnicaPT” e a área lateral o conteúdo “ocRevTecnicaPT”. Esses nomes são atribuídos aos conteúdos na sua criação, através da interface de administração, na tela de inserção de conteúdos adaptativos. Também é possível ver na Figura 5 a adaptação de *hyperlinks* do menu de acordo com as regras anotadas com atributo type=’menu’. As regras de menu indicam que, ao acessar a página com nome “revistasTecnicas”, sejam mostrados os *hyperlinks* das páginas de livros técnicos, revistas de jogos, DVDs de filmes e DVDs de jogos.

```
<rule id='8' type='menu'>
        <condition>revistasTecnicas</condition>
        <action>tabacariaLivrosTecnicos</action>
        <action>tabacariaRevistasJogos</action>
        <action>tabacariaDvdsFilmes</action>
        <action>tabacariaDvdsJogos</action>
</rule>
...
<rule id='101' type='profile' name='tecnico'>
        <condition>tabacariaLivrosAutoAjuda</condition>
        <condition>tabacariaLivrosTecnicos</condition>
        <condition>tabacariaRevistasFilmesMusicas</condition>
        <condition>tabacariaRevistasJogos</condition>
        <condition>tabacariaRevistasTecnicas</condition>
        <condition>tabacariaDvdsFilmes</condition>
        <condition>tabacariaDvdsJogos</condition>
</rule>
...
<rule id='201' type='content'>
        <condition type='profile'>tecnico</condition>
        <condition type='content'>tabacariaRevistasTecnicas</condition>
        <action type='top'>topRevTecnicasPT</action>
        <action type='content'>dcRevTecnicaPT</action>
        <action type='other'>ocRevTecnicaPT</action>
</rule>
```

Figura 7.39: Regras para adaptação de conteúdo

A arquitetura desenvolvida possibilita esse tipo de comportamento para qualquer página criada com base em sua estrutura, utilizando suas ontologias e regras de adaptação.

Resumo do capítulo:

Neste capítulo é descrita a arquitetura geral do sistema desenvolvido, sendo detalhadas as etapas gerais e as abordagens específicas de cada uma. São identificados os componentes da integração de dados de uso com semântica. São detalhados dois experimentos para a aquisição, tratamento e integração de dados de uso e semânticos em contextos voltados para a Educação. Nestes exemplos é utilizada a integração da metodologia descrita com um sistema de gerenciamento de conteúdo Web. Também é descrita a implementação de um mecanismo de gerenciamento de conteúdo Web baseado em informações semânticas, a partir de ontologias de domínio. Este ambiente permite a descrição de relações entre os conteúdos já no momento de autoria, facilitando a integração de recursos para adaptação Web.

8 CONCLUSÃO

A seguir são apresentados os comentários sobre as avaliações de resultados obtidos e as considerações finais do trabalho. Por fim são indicadas possibilidades de trabalhos futuros.

8.1 Avaliações dos resultados

Abordagens para sistemas de Hipermídia Adaptativa apresentam algumas dificuldades quanto à sua avaliação, tanto para avaliação do desempenho como da qualidade final. O sistema proposto pode ser avaliado nos dois quesitos. Como o objetivo principal é identificar melhorias na qualidade das adaptações, baseadas nas informações de classes de usuários, alguns testes foram realizados para avaliar a quantidade e qualidade das adaptações geradas e a quantidade de acessos a estas sugestões. O experimento em discussão foi realizado ao longo de diversos períodos variando entre seis e dez meses onde o material esteve disponível para acesso, com as informações de adaptação sendo geradas. O contexto de uso foi o suporte para a interação e disponibilização de material para disciplinas de curso de graduação. Alguns testes específicos foram feitos inicialmente para a geração de uma base de dados sintética (isolada) que permitisse uma validação preliminar com respeito às expectativas de resultados.

Os resultados indicam que alguns padrões freqüentes estão relacionados com comportamentos específicos. Um destes casos é a navegação geral, onde o usuário acessa os tópicos principais disponíveis no *site* Web. Esta informação é obtida com a análise dos padrões de percursos freqüentes que retornam, após a integração com informações semânticas, uma relação com um mesmo conceito em nível de abstração superior, normalmente a página de início do *site* Web. O tipo de relação verificada na ontologia neste caso é a relação "*parteDe*", sendo que todos os itens de navegação dos percursos se relacionam com um mesmo tópico na ontologia. Outro caso importante é a situação na qual os itens no padrão freqüente são relacionados entre si com a relação "*parteDe*", porém como antecedente e subseqüente na relação. Neste caso o comportamento detectado é descrito como uma navegação na qual o usuário acessa itens internos e relacionados com um tópico inicial de interesse.

Na figura 8.1 são identificadas algumas destas situações, exemplificando o ganho obtido com a integração de informação entre as duas fontes. A observação realizada tomou como base a constatação de diversos percursos freqüentes compostos por um número diferenciado de elementos. A análise de exemplos de percursos diferentes, com o mesmo número de elementos, permite identificar contextos completamente diferenciados a partir das relações descritas na ontologia. Estes contextos não seriam obtidos com a informação de uso apenas. A figura 8.1 resume alguns destes casos. Nela podem ser identificados elementos (elipses) que

correspondem aos itens acessados em percursos freqüentes, numerados segundo a ordem de acesso. Os arcos entre estes elementos indicam as relações obtidas na ontologia de domínio para os itens do percurso freqüente correspondente.

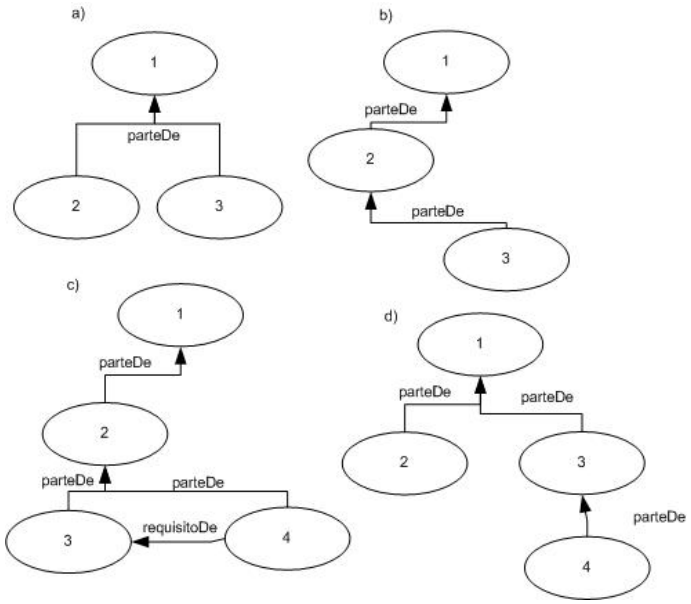


Figura 8.1: Contexto semântico de percursos freqüentes

Observando-se os itens “a” e “b” da figura 8.1 percebe-se que no primeiro houve uma navegação entre níveis equivalentes quanto ao aprofundamento no conteúdo. Já no segundo item percebe-se que a navegação seguiu o caminho de detalhamento para um determinado tópico. Desta forma, podem ser determinadas regras de adaptação diferenciadas. Nos resultados de acesso os padrões observados são mais extensos, porém foram utilizados na figura exemplos com padrões com número menor de elementos. Nos padrões mais extensos este comportamento é igualmente verificado. Também pode ser observado na figura 8.1, nos itens “c” e “d”, outro exemplo da descoberta de contextos diferenciados para padrões de acesso freqüentes. No primeiro caso a navegação iniciou pelo nível geral e foi logo direcionada para um nível mais detalhado. Já no segundo houve inicialmente uma navegação mais extensa no nível geral e depois a escolha por um nível mais detalhado.

Conforme detalhado na tabela 7.2, a integração das informações do percurso de acesso com as informações semânticas descritas na ontologia de domínio possibilita a detecção de semelhanças e diferenças entre padrões de acesso. Estas semelhanças não estão baseadas na ordem de acesso apenas. Apesar deste fator ter sido levado em consideração na metodologia, são as relações descritas na ontologia de domínio e as anotações semânticas descrevendo o conteúdo publicado que permitem que os resultados do acompanhamento de uso e da mineração de padrões de acesso sejam aprofundados. Esta melhoria obtida no contexto semântico pode ser aproveitada na geração de adaptações de forma automática e coerente, utilizando as relações já descritas na ontologia.

Foram realizadas experiências com diversas situações, na quais as descrições das relações na ontologia e a quantidade de acessos monitorados apresentaram variações significativas. Entretanto considera-se que os resultados obtidos são adequados mesmo

com uma quantidade pequena de acessos. As relações descritas na ontologia permitem que a qualidade dos contextos semânticos melhore, possibilitando identificação de padrões de modo mais completo.

No processamento destes percursos de acordo com as informações semânticas observa-se o agrupamento de diversos deles em um mesmo contexto, o que permite supor a identificação de uma tarefa comum aos diversos usuários. Por exemplo, em conjunto de percursos de freqüência alta são identificados contextos semânticos diferentes, porém em número significativamente menor. Estes fatos apontam para uma boa possibilidade de utilização desta abordagem na identificação de comportamentos navegacionais acrescidos de semântica. Uma vantagem adicional é a possibilidade existente no sistema para a descrição de regras de adaptação que utilizem este contexto semântico e não apenas a informação estatística.

As adaptações sugeridas são monitoradas e o acesso a estes itens pode ser comparado com os acessos aos itens não modificados do *site* Web. Neste caso, avalia-se que o método proposto favorece a geração de adaptações significativas, relacionando informações de uso e semânticas. Existem variações nos resultados dependendo da quantidade de acessos e tempo de uso do sistema. A quantidade de acessos também está relacionada com a necessidade de um período de uso do sistema, para a identificação de necessidades de alguns grupos de usuários. Conforme referido, ao tomarmos o conjunto total de acessos, o número de utilizações das opções de navegação adaptadas é de cerca de 7 %. Entretanto ao utilizarmos períodos de acesso próximos ao final do experimento este valor é de cerca de 23%. Isso confirma uma tendência esperada pela dinâmica do processo implementado.

8.2 Considerações finais

Com a enorme quantidade de documentos disponíveis atualmente na Web, o acesso e a coleta da informação desejada podem se tornar tarefas difíceis e originar resultados de baixa qualidade. A adaptação de *sites* Web permite minorar este problema, apresentando os conteúdos ou a estrutura dos sites de acordo com o perfil de uma classe de usuários. Este recurso de adaptação pode ser observado em diversos sistemas desenvolvidos em âmbito de pesquisa, mas também em alguns recursos comerciais, sendo que as áreas de utilização são bastante amplas.

Componente importante dos sistemas conhecidos para adaptação, a geração e manutenção do perfil dos usuários pode ser realizada em diversas maneiras. Este perfil do usuário pode ser composto por informações válidas em períodos curtos ou longos, sendo estas duas abordagens respectivamente associadas às tarefas executadas e dados cadastrais. Normalmente as informações com prazo de validade longo necessitam a identificação do usuário para sua correta aquisição. Em alguns contextos esta identificação do usuário é importante e fundamental, porém em outros pode não ser desejada pelos usuários e pode não ser fundamental. Nestes casos, uma abordagem possível é o tratamento das adaptações a partir da perspectiva de classes de usuários. Acredita-se ser possível a obtenção de bons resultados tratando-se a adaptação do *site* Web a partir desta perspectiva de uma classe de usuários e não de um usuário específico. Esta identificação de classe de usuários pode estar associada a interesses ou tarefas em comum e pode ser detectada sem necessidade de identificação do usuário, a partir do acompanhamento de sua interação com *sites* Web.

Outro fator a destacar em relação às informações de longo prazo descritas no perfil do usuário é uma possível tendência à desatualização, devido à diversos fatores alheios ao sistema de adaptação e cuja identificação somente pode ser realizada no caso de uma ação explícita do usuário interagindo com o sistema e atualizando as informações. A identificação de necessidades de curto prazo, como uma tarefa realizada em uma sessão de interação com um *site* Web, não está associada à necessidade de identificação ou de informação adicional do usuário, podendo ser efetivada com acompanhamento do uso da Web.

Esta identificação de metas e interesses mais imediatos dos usuários, a partir do acompanhamento do uso Web, pode vir a atender uma demanda de melhorias nos mecanismos de geração e manutenção de perfis de usuários.

A mineração de uso da Web possibilita a captura e análise de características do comportamento dos usuários de *sites* Web, inclusive para utilização em mecanismos voltados à personalização e adaptação. Tipicamente podem ser obtidos padrões de comportamento com técnicas como mineração de percursos freqüentes, regras de associação ou agrupamentos. Estes padrões podem guiar ações de ajustes em conteúdo ou estrutura dos documentos apresentados.

Esta abordagem por ser expandida, com o uso de informações semânticas associadas às informações de uso, sendo este o objetivo do presente trabalho. Em geral existem duas formas de encaminhamento desta melhoria, ou seja, o uso das informações semânticas na etapa de pré-processamento, enriquecendo a geração de padrões ou então o uso de informações semânticas em etapa posterior, junto com a adaptação propriamente dita.

As informações semânticas referidas podem ser descritas a partir de conjuntos de recursos sendo disponibilizados pela iniciativa da Web Semântica, que oferece suporte tanto para a anotação semântica dos documentos e conteúdos, em uma abordagem mais geral, como para a descrição de modelos para a aplicação Web em suas diversas etapas, tais como a descrição do conteúdo, da interface, da apresentação e mesmo a adaptação. Esta possibilidade foi evidenciada neste trabalho, a partir da descrição breve de iniciativas para o desenvolvimento de métodos para descrição semântica de aplicações Web.

Neste contexto, este trabalho apresenta uma metodologia de integração de recursos de mineração de uso web e modelagem semântica de aplicações voltada para a aquisição e tratamento automático de perfis de usuários para a utilização em uma aplicação de Hipermídia Adaptativa, cujo objetivo é possibilitar a adaptação de estrutura e conteúdo Web. Tratando de etapas importantes no processo de adaptação de sites, como o acompanhamento de uso e a modelagem da aplicação, esta proposta vem contribuir no sentido da aquisição automática de informações de uso Web e geração de classes de usuários e no processo de utilização destas informações para a geração de adaptações.

Os experimentos realizados e experiências anteriores, bem como o trabalho previsto em continuidade a este apresentado, indicam a possibilidade de obtenção de padrões significativos de acesso, relacionando páginas de determinado *site* Web, somente a partir da navegação de usuários. Estes padrões podem ser tratados com a tradução de determinadas necessidades ou metas, utilizadas como o identificador de classes de usuários. Também é evidenciado que a utilização de uma ontologia de domínio, na qual as páginas podem ser associadas a conceitos específicos ou com etapas

de processos repetitivos no *site* é mais abrangente que a utilização apenas das informações de acesso, sem o seu relacionamento com informações semânticas. Esta linha de trabalho pode ser aprofundada quando a aplicação possuir uma descrição formal, com uso de ontologias e associada a modelos específicos que auxiliem, por exemplo, na identificação de conceitos associados ao domínio da aplicação ou a etapas de tarefas rotineiras.

8.3 Trabalhos futuros

Durante o trabalho desenvolvido foram identificados alguns fatores que apresentam possibilidades de aprofundamento, disponibilizando mais recursos para o tratamento das adaptações. Também foram identificados problemas, relacionados com as escolhas realizadas no trabalho. Assim estas situações são relacionadas e comentadas abaixo, como sugestões de trabalhos futuros.

Um deles é o uso de elementos com granularidade mais fina. No modelo atual são tratados os acessos a páginas Web, sendo este o menor elemento considerado pelo mecanismo. Entretanto, na modelagem de aplicações com metodologias baseadas em semântica, é viável a identificação de elementos internos a uma página Web, que fazem parte de sua composição com finalidades distintas e cujo acesso pode ter um significado importante. Tecnologias como AJAX[85] podem ser testadas para prover o tratamento dos elementos das páginas Web com maior nível de detalhe e com maior interatividade.

Outro trabalho possível é a integração das informações de uso com informações sobre os conteúdos, porém de forma automática, independentemente da anotação semântica manual. Desta forma, seria possível ampliar o conjunto de fatores levados em conta na elaboração das adaptações. Para esta abordagem fazem-se necessários subsídios que podem ser obtidos com tecnologias ligadas ao Processamento de Linguagem Natural. Um destes recursos é a identificação automática de termos e entidades, bem como suas relações. Outro é a obtenção automática de ontologias e termos em documentos, o que possibilitaria a identificação de tópicos em documentos.

A integração da arquitetura baseada em Web Semântica com a aquisição de dados de uso e com os contextos semânticos obtidos possibilita que as adaptações sejam geradas levando em conta outros fatores, tais como a estrutura da aplicação e dos componentes de apresentação dos resultados. Este tipo de abordagem integrada facilita o desenvolvimento de aplicações capazes de adequação dos resultados também a diferentes dispositivos ou à situações de uso de mídias auxiliares, como a informação sonora.

---

85 http://ajaxpatterns.org/

REFERÊNCIAS

ADRIANS, P.; ZANTINGE, D. **Data mining**. Harlow: Addison-Wesley, 1997. 158 p.

AGGARWAL, C. C.; YUU, P. S. An automated system for web portal personalization. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES, 28., 2002, Hong Kong, China. **Proceedings…** [S.l.:s.n.], 2002. p. 1031-1040.

AGRAWAL R.; SRIKANT R. Fast algorithms for mining association rules. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES, 20., 1994, Santiago, Chile. **Proceedings**… [S.l.:s.n.], 1994. p. 487-99.

AGRAWAL R.; SRIKANT R. Mining sequential patterns: generalizations and performance improvements. In: EXTENDING DATABASE TECHNOLOGY, 5., 1996, Avignon, France. **Proceedings…** [S.l.:s.n.], 1996. p. 3-17.

ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster analysis.** Beverly Hills, CA: Sage, 1984. 88 p.

ALEXAKI, S.; CHRISTOPHIDES, V.; KARVOUNARAKIS, G. The ICS-forth RDF suite: managing voluminous RDF descriptions bases. In*:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, WWW, 10., 2001, Hong Kong, China. **Proceedings…** [S.l.:s.n.], 2001. p. 1-13.

ANDREASEN, T. et al. Ontological extraction of content for text querying. In: INTERNATIONAL CONFERENCE ON APPLICATIONS OF NATURAL LANGUAGE TO INFORMATION SYSTEMS, 6., 2002, Stockholm, Sweden. **Revised Papers**. Berlin: Springer, 2002. p. 123-136.

AROYO, L.; POKRAEV, S.; BRUSSE, R. Preparing SCORM for the semantic web. In: INTERNATIONAL CONFERENCE ON ONTOLOGIES, DATABASES AND APPLICATIONS OF SEMANTICS, 3., 2003, Catania, Sicily. **Proceedings…** [S.l.:s.n.], 2003. p. 621-634.

AROYO, L. et al. Ontology-based personalization in user-adaptive systems. In: INTERNATIONAL WORKSHOP ON WEB PERSONALIZATION, RECOMMENDER SYSTEMS AND INTELLIGENT USER INTERFACES, 2., 2006, Dublin, Ireland. **Proceedings**... [S.l.:s.n.], 2006. p. 87-97.

ASSIS, P. S. **Arquitetura para adaptação e meta-adaptação de sistemas hipermídia**. 2005. Tese (Doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro.

ASSIS, P. S. et al. Meta-models for adaptive hypermedia applications and meta-adaptation. In: ADAPTIVE HYPERMEDIA AND ADAPTIVE WEB-BASED SYSTEMS, AH, 2., 2004, Eindhoven, Netherlands. **Proceedings…** [S.l.:s.n.], 2004. p. 433-436.

ASSIS, P. S. et. al. Model-driven adaptation and meta-adaptation. In: ADAPTIVE HYPERMEDIA AND ADAPTIVE WEB-BASED SYSTEMS, AH, 4., 2006, Dublin, Ireland. **Revised Papers**. Berlin: Springer, 2006. p. 213-222. **(**Lecture Notes in Computer Science, v. 4018).

BAADER, F. **The description logic handbook** : theory, implementation and applications. Cambridge, UK: Cambridge University Press, 2002.

BAEZA-YATES, R. **Modern information retrieval.** New York, N.Y.: Addison-Wesley, 1999. 513 p.

BAEZA-YATES, R.; POBRETE, B. A. Content and structure website mining model. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, WWW, 15., 2006, Edinburgh, Scotland. **Proceedings**… [S.l.:s.n.], 2006. p. 957-958.

BAILEY, C. et al. Towards open adaptive hypermedia. In: INTERNATIONAL CONFERENCE ON ADAPTIVE HYPERMEDIA AND ADAPTIVE WEB BASED SYSTEMS, AH, 2., Malaga, Spain. **Revised Papers**. Berlin: Springer, 2002. p. 36-46. (Lecture Notes in Computer Science, v. 2347).

BALABANOVICH, M.; SHOAM, Y. Content-based, collaborative recommendation. **Communications of ACM**, New York, v. 40, n. 3, p. 66-73, Mar. 1997.

BALDONI, M.; BAROGLIO, C.; PATTI, V. **Structureless, intention-guided web sites:** planning based adaptation. Torino: Dipartamento de informatica – Università Degli Studio di Torino, 2003. Disponível em: <http://www.di.unito.it/~argo/papers/2001_UAHCI01.pdf> . Acesso em: jun. 2008.

BAMSHAD M.; COOLEY R.; SRIVASTAVA J. Automatic personalization based on web usage mining. **Communications of the ACM**, New York, v. 43, n. 8, p. 142-151, Aug. 2000.

BARAGLIA, R.; PALMERINI, P. Suggest: a web usage mining system. In: INTERNACIONAL CONFERENCE ON INFORMATION TECHNOLOGY : CODING AND COMPUTING, 2002, Las Vegas, USA. **Proceedings…** [S.l.:s.n.], 2002. p. 282 – 287.

BARAGLIA, R.; SILVESTRI, F. An online recommender system for large web sites. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, WI, 2004, Beijing, China. **Proceedings**… [S.l.:s.n.], 2004. p. 199-205.

BARAGLIA, R. Personalization of web sites without user intervention. **Communications of the ACM,** New York, v. 50, n. 2, p. 63-67, Feb. 2007.

BECHHOFER, S.; GOBLE, C. Towards annotation using DAML+OIL. In: K-CAP WORKSHOP ON KNOWLEDGE MARKUP AND SEMANTIC ANNOTATION, Victoria, Canadá. **Proceedings**… [S.l.:s.n.], 2001. p. 235 – 246.

BELEW R. K. **Finding out about – a cognitive perspective on search engine technology and the www**. Cambridge, UK: Cambridge University Press, 2000.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web: a new form of web that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, May 2001. Disponível em: <http://www.sciam.com/article.cfm?id=the-semantic-web>. Acesso em: jun. 2008.

BERRIOS, D.; KELLER, R. M. Developing a web-based user interface for semantic information retrieval. In*:* INTERNATIONAL SEMANTIC WEB CONFERENCE, ISWC, 2., 2003, Sanibel Island, Florida. **Proceedings…** [S.l.:s.n.], 2003. p. 65 - 70.

BERTRAM, C. et al. Semantic association identification and knowledge discovery for national security applications. **Journal of Database Management**, Lincoln, USA, v.16, n.1, p. 33 – 53, Jan. 2005.

BRIN, S.: PAGE, L. The anatomy of a large-scale hypertextual web search engine. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, WWW, 7., 1998, Brisbane, Austrália. **Proceedings…** [S.l.:s.n.], 1998. Disponível em: http://www-db.stanford.edu/pub/papers/google.pdf>. Acesso em: jun. 2008.

BRUSILOVSKY, P. User modeling and user-adapted interaction. **Adaptive Hypermedia,** Netherlands, v. 11, n. 1, p. 87 - 110, Apr. 2001.

BRUSILOVSKY, P. Adaptive navigation support. In: BRUSILOVSKY, P.; KOBSA, A.; NEIDL, W. (Ed.). **The adaptive web:** methods and strategies of web personalization. Berlin: Springer-Verlag, 2007. p. 263-290. (Lecture Notes in Computer Science, v. 4321).

BRUSILOVSKY, P. Methods and techniques of adaptive hypermedia. **User Modeling and User-Adapted Interaction,** Berlin, v**.** 6, n. 2, p. 87-129, Mar. 1996.

BRUSILOVSKY, P. Adaptive and intelligent technologies for web-based education. **Künstliche Intelligenz,** Berlin, v. 4, n. 1, p.19-25, Mar. 1999.

BRUSILOVSKY, P.; KARAGIANNIDIS, C.; SAMPSON, D. Layered evaluation of adaptive learning systems. **International Journal of Continuing Engineering Education and Lifelong Learning**, [S.l.], v.14, n. 4/5, p. 402 – 421, 2004.

BRUSSE R.; ALBERNIK, M.; VEENNUSTRA, M. **Using semantic web technology for e-learning**. Disponível em: <https://doc.telin.nl/dscgi/ds.py/Get/File-22688/semanticwebsci2002.pdf>. Acesso em: mar. 2008.

CARMEL, D. et al. An extension of the vector space model for querying XML documents via XML fragments. In: XML AND INFORMATION RETRIEVAL, 2002, Toronto. **Proceedings...**Toronto, Canadá: ACM SIGIR, 2002.

CERI, S. et al. **Designing data-intensive web applications**. San Francisco, EUA: Morgan Kaufmann, 2003. 596 p.

CERI, S.; FRATERNALI, P.; BONGIO, A. Web modeling language (webml): a modeling language for designing web sites. **Computer Networks**, Amsterdam, v.33, n. 1-6, p. 137-157, June 2000.

CHEN, S. Y.; MAGOULAS, G. D. **Adaptable and Adaptive Hypermedia Systems**. Hershey, PA: IRM Press, 2005. 342 p.

CHISHMAN, R.; RIGO, S. J.; VIEIRA, R. Uso de ontologias específicas para busca em domínio. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL, 4., 2003, Campinas. **Proceedings...** Campinas: Sociedade Brasileira de Ciências, 2003. p. 37-47.

CHRISTOPHER, D. STAFF: The Hypercontext Framework for Adaptive Hypertext. In: CONFERENCE ON HYPERTEXT AND HIPERMEDIA, 13., 2002, Maryland. **Proceedings...** Maryland: ACM, 2002. p. 11-20.

CIORASCU, C.; CIORASCU, I.; STOFFEL, K. Knowler – ontological support for information retrieval systems. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE, 26., 2003, Toronto. **Proceedings...** Toronto: ACM SIGIR, 2003. p. 20 – 28.

COHEN, S. et al. XSEarch: a semantic search engine for XML. In: VLDB CONFERENCE, 29., 2003, Berlin, Germany. **Proceedings...** Berlin: VLDB, 2003. p. 45 – 56.

COOK, D. J.; HOLDER, L. B. Graph-based data mining. **IEEE Intelligent Systems**, Los Alamitos, v. 15, n. 2, p. 32-41, Mar./Apr. 2000.

COOLEY R.; BAMSHAD M.; STRIVASTAVA J. Data preparation for mining World Wide Web browsing patterns. **Journal of Knowledge and information systems**, [S.l.], v. 1, p. 5 – 32, Apr. 1999.

COST, R. et al. ITtalks: a case study in the semantic web and DAML+OIL. **IEEE Inteligent Systems**, Los Alamitos, v. 17, n. 1, p. 40-47, 2002.

CRAMPES M.; RANWEZ, S. **Ontology-suported and ontology-driven conceptual navigation on the world wide web**. San Antonio, Texas: Hypertext, 2000.

DAVIES J.; WEEKS, R.;KROHN, U. QuizRDF: search technology for the semantic Web. In. ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 37., 2004, Big Island, Hawaii. **Proceedings...** Big Island, Hawaii: HICSS, 2004. p. 8 – 17.

DE BRA, P.; ARROYO, L.; CHEPEGIN, V. The next big thing: adaptive web-based systems. **Journal of Digital Information**, Salzburg, Austria, v. 5, n. 1, p. 214 - 216, 2004.

DE BRA, P.; BRUSILOVSKY P.; HOUBEN, G. Adaptive hypermedia: from systems to framework. **Computing Surveys**, New York, v. 31, n. 4, p. 26 - 5, 1999.

DE BRA, P. et al. The adaptive hypermedia architecture. In: CONFERENCE ON HYPERTEXT AND HYPERMEDIA, 40., 2003, Nottinghan, UK. **Proceedings...** New York: ACM, 2003. p. 81-84.

DE BRA, P. et al. Authoring and management tools for adaptive educational hypermedia systems: the AHA! Case study. In: JAIN, L. C.; TEDMAN, R. A.; TEDMAN, D. K. (Ed.). **Evaluation of Teaching and Learning Paradigms in Intelligent Environment.** Berlin: Springer, 2007. p. 285-308. (Studies in Computational Intelligence, 62).

DE BRA, P. STASH, N.; DE LANGE, B. AHA! Adding Adaptive Behavior to Websites. In: NLUUG CONFERENCE, 10., 2003, Ede, Netherlands. **Proceedings...** Ede, Netherlands: NLUUG, 2003. p. 20 – 30.

DE TROYER, O.; LEUNE, C. WSDM: a user-centered design method for web sites. **Computer Networks**, Amsterdam, v. 30, p. 85-94, 1998.

DEMIRIZ, A. Webspade: a parallel sequence mining algorithm to analyze web log data. In: INTERNATIONAL CONFERENCE ON DATA MINING, 2., 2002, Irving, TX, USA. **Proceedings...** [S.l.]:IEEE, 2002. p. 755-758.

DESHPANDE, M.; KURAMOCHI, M.; KARPHYS, G. Frequent sub-structure based approaches for classifying chemical compounds. **IEEE Transactions on Knowledge and Data Engineering**, New York, v. 17, n. 8, p. 1036 – 1050, Aug. 2005.

DIAZ, A. et al. Extending the capabilities of rmm: Russian dolls and hypertext. In: HAWAI INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 30., 1997. **Proceedings…** [S.l.]:IEEE Computer Society, 1997. v. 6, p. 177-186.

DOLOG, P. et al. Personalization in distributed e-learning environments. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 13., 2004, New York. **Proceedings...** New York: ACM Press, 2004. p. 85-94.

EIRINAKI, M. et al. SEWeP: using site semantics and a taxonomy to enhance the web personalization. In: SIGKDD CONFERENCE, 9., 2003, Washington, D.C, USA. **Proceedings...** [S.l.:s.n.], 2003. p. 99 – 108.

EIRINAKI, M. et al. Introducing Semantics in Web Personalization: the role of ontologies. In: ACKERMANN, M. et al. (Org.). **Semantics, Web, and Mining.** Berlin:Springer, 2006. p. 147-162.

EL-SAYED, M.; RUIZ, C.; RUNDENSTEINER, E. A. FS-Miner: efficient and incremental mining of frequent sequence patterns in web logs. In: ANNUAL ACM INTERNATIONAL WORKSHOP ON WEB INFORMATION AND DATA MANAGEMENT, 6., 2004, Washington DC, USA. **Proceedings...** New York: WIDM, ACM Press, 2004. p. 128-135.

ENGELS, R. H. P.; BREMDAL, B. A.; JONES, R. Corporum: a workbench for the semantic web. In: WORKSHOP SEMANTIC WEB MINING; EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICES OF KNOWLEDGE DISCOVERY IN DATABASES, 5., 2001, Freiburg, Germany. **Proceedings...** [S.l.:s.n.], 2001.

ERDMANN, M.; STUDER, R. Ontologies as Conceptual Models for XML Documents. In: KNOWLEDGE ACQUISITION FOR KNOWLEDGE-BASED SYSTEMS WORKSHOP, 12., 1998, Banff, Canada. **Proceedings...** Banff, Canada: Voyager Inn, 1998. p. 37 – 49.

ESPINOZA , F.; HÖÖK, K. An interactive www interface to an adaptive information system. In: USER MODELLING FOR INFORMATION FILTERING ON THE WWW, UM, 1996, Hawaii. **Proceedings...** [S.l.:s.n.], 1996.

FAYAD, U.; PIATETSKY-SHAPIRO.; PADHRAIC. S. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, New York, v. 39, n. 11, p. 27-34, Nov. 1996.

FENSEL, D. **Ontologies**: silver bullet for knowledge management and electronic commerce. Berlin: Springer-Verlag, 2001.

FENSEL, D. Ontology-based knowledge management. **Computer**, New York, v. 35, n. 11, p. 56-59, 2002. Disponível em: <http://www.isi.edu/info-agents/workshops/ijcai03/papers/DIsern-article-ijcai.pdf>. Acesso em: jun 2008.

FENSEL, D. **Ontoweb – Project 2003**. Disponível em: <http://ontoweb.aifb.uni-karlsruhe.de/>. Acesso em: jun. 2008.

FONS, J. J. et al. OOWS: un método de producción de software en ambientes web. **Avances en Comercio Electrónico**, [S.l.], v. 9, p. 119 - 136, 2002.

FREITAS, F. L. G.; Ontologias e a web semântica. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003, Campinas. **Anais...** Campinas: JAI, 2003. v. 8, p. 1-52.

FREITAS, V. et al. Adaptweb: an adaptative web-based courseware. In: INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGIES IN EDUCATION, 2002. **Proceedings...** Badajoz, ES: ICTE, 2002. p. 20 – 23.

GANGEMI, A.; MIKA, P. Understanding the semantic web through descriptions and situations on the move to meaningful internet systems. In: COOPIS, DOA, AND ODBASE - OTM CONFEDERATED INTERNATIONAL CONFERENCES, 2003, Catania, Sicily, Italy. **Proceedings…** Berlin:Springer, 2003. (Lecture Notes in Computer Sciences, v. 2888).

GRABS T.; SCHEK, H. J. Generating Vector spaces On-The-Fly for flexible XML Retrieval. In: WORKSHOP ON XML AND INFORMATION RETRIEVAL, 2002, Athens, Greece; ACM SIGIR, 2002, Athens, Greece. **Proceedings...** [S.l.:s.n.], 2002.

GRUBER, T. R. A translation approach to portable ontologies. **Knowledge Acquisition,** [S.l.], v. 5, n. 2, p. 199-220, 1993. Disponível em: <http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html>. Acesso em: jun. 2008.

GRUBER, T. R. **What is an Ontology?** 1992. Disponível em: <http://www.ksl.stanford.edu/kst/what-is-an-ontology.html>, Acesso em: jun. 2008.

GUARINO, N. Understanding, building and using ontologies. A commentary to using explicit ontologies in kbs development. **International Journal of Human and Computer Studies**, [S.l.], n. 46, p. 293-310, 1997.

HAN, J. et al. Mining frequent patterns without candidate generation: a frequent pattern tree approach. **Data Mining Knowledge Discovery**, [S.l.], v. 8, n. 1, p. 53-87, 2004.

HARMELEN, F.; BROEKSTRA, J.; KAMPMAN, A. Sesame: An architecture for storing and querying rdf data and schema information. In: FENSEL, D. et al. **Spinning the Semantic Web:** Bringing the World Wide Web to Its Full Potential. Cambridge, MA:MIT Press, 2003.

HAYASHI, Y.; TOMITA, J.; KIKUL, G. Searching text-rich XML documents with relevance ranking. In: WORKSHOP ON XML AND INFORMATION

RETRIEVAL ; ACM SIGIR, 2000, Athens, Greece. **Proceedings...** [S.l.:s.n.], 2000.

HEFLIN, J. D. **OWL web ontology language use cases and requirements.** 2004. Disponível em: <http://www.w3.org/TR/webont-req>. Acesso em: jun. 2008.

HEFLIN, J. D. Towards the semantic web: knowledge representation in a dynamic, distributed environment. 2001. **Dissertation** (PhD), University of Maryland.

HENDLER, J.; BERNERS-LEE, T.; MILLER, E. Integrating Applications on the semantic Web. **Journal of the Institute of Electrical Engenieers of Japan**, [S.l.], v. 122, n. 10, p. 676-680, Oct. 2002.

HERLOCKER, J. **Understanding and improving automated collaborative filtering systems**. 2000. Thesis (Doctor) - University of Minessota. Disponível em: <http://web.engr.oregonstate.edu/~herlock/papers.html>. Acesso em: jun. 2008.

HORROCKS, I.; PATEL-SCHENEIDER, P. F. Three theses of representation in the semantic web. INTERNATIONAL WORLD WIDE WEB CONFERENCE, 12., 2003, Budapeste. **Proceedings…** Disponível em: <http://www.cs.man.ac.uk/~horrocks/Publications/download/2003/p50-horrocks.pdf>. Acesso em: jun. 2008.

IEEE. **1484.12.1**: IEEE Standard for Learning Object metadata. New York, 2002.

ISAKOWITZ, T.; STOHR, E. A.; BALASUBRAMANIAN, PRMM: A methodology for structured hypermedia design. **Communications of the ACM**, [S.l.], v. 38, n. 8, p. 34-44, 1995.

JIN, X.; XU, S.; DECKER, S. **Ontowebber:** model-driven ontology-based web site management. Palo Alto, CA: Stanford University, 2001. p. 529-547.

JIN, X.; ZHOU, Y.; MOBASHER, B. Task-oriented web user modeling for recommendation. In: INTERNATIONAL CONFERENCE ON USER MODELING, 10., 2005, Edinburgh. **Proceedings...** Edinburgh:[s.n.], 2005. p. 109 – 118.

KIM, W.; KERSCHBERG, L.; SCIME, A. Learning for automatic Personalization in a semantic taxonomy-based meta search agent. **Electronic Commerce Research and Applications**, Amsterdam, v. 1, n. 2, p. 15-173, 2003.

KLAPSING, R.; NEUMANN, G. Applying the resource description framework to web engineering. In: INTERNATIONAL CONFERENCE ON ELECTRONIC COMMERCE AND WEB TECHNOLOGIES, 1., 2000, London, UK. **Proceedings...** London, UK: Springer, 2000. p. 229-238.

KLAPSING, R. et al. Semantics in web engineering: applying the resource description framework. **IEEE MultiMedia**, [S.l.], v. 8, n. 2, p. 62-68, 2001.

KLEINBERG, J.; SANDLER, M. Using Mixture Models for Collaborative Filtering. In: ACM SYMPOSIUM ON THEORY OF COMPUTING, 36., 2004, Orlando, FL. USA. **Proceedings...** [S.l.]: Academic Press, 2004. p. 49 – 69.

KOHAVI, R. Mining e-commerce data: the good, the bad and the ugly. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 7., 2001, São Francisco. **Proceedings...** São Francisco: ACM, 2001. p. 8-13.

KOSALA, R.; BLOCKEEL, H. Web mining research: a survey. **SIGKDD Explorations**, [S.l.], v. 2. n. 1, p. 1-15, 2000.

KOUTRI, M. ; AVOURIS, N. ; DASKALAKI, S. **A survey on web usage mining techniques for web-based adaptive hypermedia systems:** adaptable and adaptative hypermedia systems. London,UK: Idea Inc., 2004.

LEI, Y.; MOTTA, E. Modelling data-intensive web sites with ontoweaver. In: WORKSHOP ON WEB INFORMATION SYSTEMS MODELLING ; CONFERENCE ON ADVANCED INFORMATION SYSTEMS, 16., 2004, Riga, Latvia. **Proceedings...** [S.l.:s.n.], 2004.

LEI, Y. Ontoweaver: an ontology-based approach to the design of data-intensive web sites. **Journal of Web Engineering**, [S.l.], v. 4, n. 3, p. 244-262, 2005.

LEY I.; DOMINGUE, J. Design of customized web applications with OntoWeaver. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE CAPTURE, 2., 2003, Sanibel Island, USA. **Proceedings...** New York: ACM Press, 2003. p. 54-61.

LELEU, M. et al. GO-SPADE: mining sequential patterns over datasets with consecutive repetitions. In: INT. CONF. MACHINE LEARNING AND DATA MINING, MLDM, 3., 2003, Leipsig, Germany. **Proceedings...** [S.l.:s.n.], 2003. p. 293-306.

LI, J.; PEASE, A.; BARBEE, C. Agent Semantic Communication Service. 2002. Project Report - Teknowledge Corporation. Palo Alto, CA, USA. Disponível em: <http://reliant.teknowledge.com/DAML>. Acesso em: jun. 2008.

LI, J.; ZHONG, N. Mining ontology for automatically acquiring web user information needs. **IEEE Transactions on Knowledge and data engineering,** [S.l.], v. 18, n. 4, p. 554-568, Apr. 2006.

LIMA, F. **Modeling applications for the semantic Web**. 2003. Tese (Doutorado)- Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro.

LIU, F.; YU, C.; MENG, W. Personalized web search by mapping user queries to categories. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE MANAGEMENT, CIKM, 2002, Virginia, USA. **Proceedings....** [S.l.]: McLean, 2002. p. 558 – 565.

LOH, S.; WIVES, L.; OLIVEIRA, J.P.M. Concept –based knowledge Discovery in texts extracted from the web. **SigKDD Explorations**, New York, v. 2, n. 1, p.29-30, 2000.

LUKE, S.; SPECTOR, L.; RAGER, D. Ontology-based knowledge discovery on the world-wide web. In: INTERNET-BASED INFORMATION SYSTEMS, 1996, Portland, USA. **Proceedings...** Portland: AAAI, 1996. p. 96 – 102.

MAEDCHE, A.; STAAB, S. Discovering conceptual relations from text. In: EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE , 2000, Berlim. **Proceedings...** Berlim: ECAI, 2000. p. 321 – 325.

MAEDCHE, A. et al. Seal - tying up information integration and web site management by ontologies. **IEEE Data Engineering Bulletin**, [S.l.], v. 25, n. 1, p. 10-17, 2002.

MAEDCHE, A. et al. Semantic portal: the seal approach. In: FENSEL, D. **Spinning the Semantic Web:** Bringing the World Wide Web to Its Full Potential. Cambridge, MA: MIT Press, 2003.

MAGNINI, B.; SERAFINI, L.; SPERANZA, M. Linguistic based matching of local ontologies. In: WORKSHOP ON MEANING NEGOTIATION ; NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 18., 2002, Edmonton, Canada. **Proceedings...** [S.l.:s.n.], 2002.

MARKELLOU P.; RIGOU, M.; SIRMAKESSIS, S. Mining for web personalization. In: SCIME, A. (Ed.). **Web Mining**: Applications and Techniques. Hershey, PA: IRM Press, 2005. p.27.

MARTELLI, S.; SIGNORI, O. Semantic characterization of links and documents. **Ercim News**, [S.l.], n. 51, 2002. Special Theme: Semantic Web.

MCBRIDE, B. Jena: implementing the rdf model and syntax specification. In: INTERNATIONAL WORKSHOP ON THE SEMANTIC WEB, 2., 2001, Hong Kong, China. **Proceedings...** HongKong: [s.n.], 2001. p. 74 – 83.

MCGUINNESS, D. L.; SILVA, P. P. Infrastructure for Web Explanations. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 2003, Sanibel Island, USA. **Revised papers.** Berlin: Springer, 2003. p.113-129. (Lecture Notes in Computer Science, v. 2870).

MEDJAED, B. et al. Composing web services on the semantic web. **VLDB Journal**, [S.l.], n. 10, 2003.

MEI, Q. et al. Semantic annotation of frequent patterns. **ACM Transactions on Knowledge Discovery from Data**, [S.l.], v.1, n. 3, Dec. 2007.

MIDDLETON, S. E. et al. Exploiting synergy between ontologies ande recommender sytems. In: SEMANTIC WEB WORKSHOP, 2002, Hawaii. **Proceedings...** [S.l.:s.n.], 2002.

MIKROYANNIDIS A.; THEODOULIDIS, B. Web site ontology evolution through web site adaptation, In: POSTGRADUATE RESEARCH CONFERENCE IN ELECTRONICS, PHOTONICS, COMMUNICATIONS AND NETWORKS, AND COMPUTING SCIENCE, 2005, Lancaster, UK. **Proceedings...** [S.l.]:IEEE, 2005. p. 22 – 26.

MILLER, E.; SWITCH, R.; BRICKLEY, D. **Resource Description Framework**. Disponível em: <http://www.w3.org/RDF>. Acesso em: jun. 2008.

MOBASCHER, B.; COOLEY, R.; SRIVASTAVA, J. Automatic personalization based on web usage minning. **Communications of the ACM**, [S.l.], v. 43, n. 8, p. 142-151, 2000.

MOBASHER, B. Web usage mining and personalization. In: SINGH, M. P. (Ed.). **Practical handbook of internet computing**. Lincoln, USA: CRC Press, 2005.

MOBASHER, B.; DAÍ, H. Integrating semantic knowledge with web usage mining for personalization. In: SCIME, A. (Ed.). **Web mining**: applications and techniques. London, UK: Idea Group Publishing, 2004.

MOBASHER, B. Using ontologies to discover domain-level web usage profiles. In: WORKSHOP ON SEMANTIC WEB MINING, 2., 2002. Helsinki. **Proceedings...** Helsinki:[s.n.], 2002.

MRABET, Y. et al. Recognising professional-activity groups and web usage mining for web browsing personalization. In: INTERNATIONAL CONFERENCE ON WEB INTELIGENCE, 2007, Silicon Valley, USA. **Proceedings...** [S.l.]: IEEE, 2007. p. 719 – 722.

MURUGESAN, S.; DESHPANDE, Y. (Ed.). **Web Engineering - Managing the Diversity and Complexity of Web Application Development**. Berlin: Springer, 2001. p. 223-235. (Lecture Notes in Computer Science, v. 2016).

NAEVE, A. **The concept browser- a new form of knowledge management tool**. 2003. Technical Report - Stockolm. Disponível em: <http://cid.nada.kth.se/pdf/CID-159.pdf >. Acesso em: jun. 2008.

NASRAOUI, O. et al. Mining Web Access Logs Using Relational Competitive Fuzzy Clustering. **International Journal on AI Tools**, [S.l.], v. 9, n. 4, p.509 – 526, 2000.

NIELSEN, H. F. **Logging in W3C httpd**. Disponível em: <http://www.w3.org/Daemon/User/Config/Logging.html>. Acesso em: jun. 2008.

NILL, A. Adaptivity and user modeling within the AVANTI Project. In: WORKSHOP ON ADAPTIVE MULTIMEDIA TECHNOLOGIES FOR PEOPLE WITH DISABILITIES, 1995; ACM MULTIMEDIA, 1995, San Francisco, USA. **Proceedings...** [S.l.:s.n.], 1995.

NILSSON, M.; NAEVE, A. Semantic web meta-data for e-learning – some architectural guidelines. In: WORLD WIDE WEB CONFERENCE, 11., 2002, Hawaii. **Proceedings...** [S.l.:s.n], 2002. Disponível em: <http://kmr.nada.kth.se/papers/SemanticWeb/p744-nilsson.pdf>. Acesso em: jun. 2008.

NILSSON, M.; PALMER, M.; BRASE, J. The LOM RDf binding-principles and implementation. In: ARIADNE CONFERENCE, 3., 2003, Leuven, Belgium. **Proceedings…** [S.l.:s.n.], 2003. p. 11-17.

NUNES, D. A.; SCHWABE, D. Rapid prototyping of web applications combining domain specific languages and model driven design. In: INTERNATIONAL CONFERENCE ON WEB ENGINEERING, 2006, Palo Alto. **Proceedings...** [S.l.]: ACM, 2006. p. 153-160.

OBERLE, D. et al. An Extensible ontology software environment. In: STAAB, S.; STUDER, R. **Handbook on ontologies.** Berlin: Springer, 2003. Disponível em: <http://www.aifb.uni-karlsruhe.de/WBS/dob/pubs/handbook2003a.pdf> Acesso em: jun. 2008.

OLIVEIRA, J. P. M.; MUÑOZ, L. S. Adaptive web-based courseware development using metadata standards and ontologies. In: INTERNATIONAL CONFERENCE, 16., 2004, Riga. **Proceedings...** [S.l.:s.n], 2004. p. 37 – 49.

OLIVEIRA , J. P. M.; RIGO, S. J. Mineração de uso em sites web para a descoberta de classes de usuários. In: CLEI, 2006, Santiago, Chile. **Proceedings...** [S.l.:s.n.], 2006. p. 127 – 150.

ORLANDO, S.; PEREGO R.; SILVESTRI, C. CCSM: an efficient algorithm for constrained sequence mining. In: INTERNATIONAL WORKSHOP ON HIGH PERFORMANCE DATA MINING, 6.; INTERNATIONAL SIAM CONFERENCE ON DATA MINING, 30., 2003, San Francisco. **Proceedings...** New York: ACM**,** 2003. p. 540 – 547.

PALMER M.; NILSSON, M.; BRASE, J. Semantic web metadata for e-learning – some architectural guidelines. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 12., 2002, Honolulu, Hawaii. **Proceedings...** [S.l.:s.n.], 2002. p. 7 – 11.

PARAMYTHIS, A.; STEPHANIDIS, C. A generic adaptation framework for web-based hypermedia systems. In: SCIME, A. (Ed.). **Adaptable and adaptive Hypermedia Systems**. London, UK: Idea Group Publishing, 2005. p.80-103.

PASTOR, O. et al. Conceptual modelling of web applications: the OOWS approach, 3540281967, web Engineering. In.: MENDES, E.; MOSLEY, N. **Theory and Practice of Metrics and Measurement for Web Development**. Berlin: Springer, 2005. p. 1-400.

PEI, J. et al. Mining access patterns efficiently from web logs. In: PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, PAKDD, 2000, Kyoto, Japan. **Proceedings…** [S.l.:s.n.], 2000. p.396 – 407.

PETRELLI, D. User-centred design of flexible hypermedia for a mobile guide: reflections on the hyperaudio experience. **User Modeling and User-Adapted Interaction**, Netherlands, v. 15, n. 1, p. 303 – 338, Apr. 2005.

POPOV, B. et al. Towards semantic web information extraction. In: WORKSHOP ON HUMAN LANGUAGE TECHNOLOGY FOR THE SEMANTIC WEB AND WEB SERVICES, 2003, Sanibel Island, Florida, USA; INTERNATIONAL SEMANTIC WEB CONFERENCE, 2., 2003, Sanibel Island, Florida, USA. **Proceedings…** [S.l.:s.n.], 2003.

PRESSMAN, R. S. **Software engineering**: a practitioner's perspective. 6th ed. New York: McGraw-Hill, 2004.

QUARESMA, P.; RODRIGUES, I. P. A natural language interface for information retrieval on semantic web documents. In: ATLANTIC WEB INTELLIGENCE CONFERENCE. 2003, Madrid. **Revised Papers.** Berlin: Springer, 2003. p.142-154. (Lecture Notes in Computer Science, v. 2663).

QUARESMA, P. Using logic programming to model multi-agent web legal systems – an application report. In: ICAIL, 2001, St.Louis, USA. **Proceedings …** [S.l.:s.n.], 2001.

REGGIORI, A.; GULIK, D. W.; BJELOGRLIC, Z. Indexing and Retrieving Semantic Web Resources: The RDFStore model. In: SWAD-EUROPE WORKSHOP ON SEMANTIC WEB STORAGE AND RETRIEVAL, 2003, Amsterdam, Netherlands. Disponível em: <http://www.asemantics.net/presos/SWAD-E/SWADe-rdfstore.html>. Acesso em: jun. 2008.

RIGO, S. J.; OLIVEIRA, J. P. M. Aquisição Automática de Classes de Usuários Integrando Mineração de Uso da Web e Ontologias. In: WORKSHOP EM ALGORITMOS E APLICAÇÕES DE MINERAÇÃO DE DADOS, 2., 2006, Florianópolis. **Proceedings...** [S.l.:s.n.], 2006.

RIGO, S. J.; OLIVEIRA, J. P. M. Aquisição de classes de usuários por mineração do uso da Web. In: SIMPÓSIO BRASILEIRO DE SISTEMAS MULTIMEDIA E WEB, 12., 2006, Natal. **Proceedings...** [S.l.:s.n.], 2006.

RIGO, S. J.; OLIVEIRA, J. P. M. Identifying users stereotypes in educational websites with semantic web resources and web usage mining. 2008. A ser publicado em 2008 na Scientia**,** São Leopoldo.

RIGO, S. J.; OLIVEIRA, J. P. M. Personalização de sites web integrando mineração de uso e ontologias de domínio. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 13., 2007, Gramado, BR. **Proceedings...** [S.l.:s.n.], 2007.

RIGO, S. J.; OLIVEIRA, J. P. M. Uso de semântica e mineração de uso web para identificação de classes de usuários em sites educacionais. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2007, São Paulo. **Anais ...** [S.l.:s.n.], 2007.

RIGO, S. J.; OLIVEIRA, J. P. M.; SCHNEIDER, E. E. Arquitetura baseada em Web Semântica para aplicações de Hipermídia Adaptativa. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, SBSI, 2008, Rio de Janeiro, RJ. **Anais ...** [S.l.:s.n.], 2008.

ROMERO, C. et al. Personalized Links Recommendation Based on Data Mining. In*:* EUROPEAN CONFERENCE ON TECHNOLOGY ENHANCED LEARNING, EC-TEL, 2., 2007, Crete, Greece. **Revised Papers.** Berlin:Springer, 2007. p. 292 – 306. (Lecture Notes in Computer Science, v. 4753).

SAH, M.; HALL, W. Building and managing personalized semantic portals. In*:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, 16., 2007, Banff, Canada. **Proceedings....** [S.l.:s.n], 2007. p. 1227 – 1228.

SAH, M. et al. Semport: a personalized semantic portal. In: CONFERENCE ON HYPERTEXT AND HYPERMEDIA, 18., 2007, Manchester, UK. **Proceedings...** New York: ACM, 2007. p.31-32.

SAIAS J. M. G. **Uma metodologia para a construção automática de ontologias e a sua aplicação em sistemas de recuperação de informação**. 2003. Dissertação (Mestrado) - Universidade de Évora, Portugal.

SAIAS J. M. G.; QUARESMA, P. A methodology to create ontology-based information retrieval systems. In: PORTUGUESE CONFERENCE ON ARTIFICIAL INTELLIGENCE, EPIA, 13., 2003, Beja, Portugal. **Revised Papers**. Berlin: Springer, 2003. p. 424 - 434 (Lecture Notes in Computer Science, v.2902).

SANCHO, P.; FERNANDEZ-MANJON, B. Creating cost-effective adaptative educational hypermedia base don markup Technologies and e-learning

standards. In: INTERACTIVE EDUCATION MULTIMEDIA, 2002. **Proceedings...** Disponível em: <http://www.ub.es/multimedia/iem>. Acesso em: jun. 2008.

SCHEREIBER, A. et al. Ontology-based photo annotation. **IEEE Inteligent Systems**, [S.l.], p. 66-74, May/June 2001.

SCHWABE, D.; ASSIS, P.; BARBOSA, S. Meta-models for Adaptive Hypermedia Applications and Meta-adaptation. In: WORLD CONFERENCE ON EDUCATIONAL MULTIMEDIA, HYPERMEDIA AND TELECOMMUNICATIONS, EDMEDIA, 2004, Chesapeake, VA, USA. **Proceedings…** [S.l.]: AACE, 2004. p. 1720-1727.

SCHWABE , D.; BARBOSA, S. D. J. Systematic hypermedia application design with oohdm. In: ACM CONFERENCE ON HYPERTEXT, 17., 1996, Maryland, USA. **Proceedings...** New York: ACM, 1996. p. 116-128.

SCHWABE, D. et al. Design and implementation of semantic web applications. In: WORKSHOP ON APPLICATION DESIGN, DEVELOPMENT AND IMPLEMENTATION ISSUES IN THE SEMANTIC WEB, 2004, New York. **Proceedings...** New York: [s.n.], 2004. p. 20 – 27.

SCHWABE, D.; ROSSI, G. An object oriented approach to web-based application design. **Theory and Practice of Object Systems**, New York, v. 4, n. 4, 1998.

SCIME, A. (Ed.). **Web mining**: applications and techniques. London, UK: Idea Group Publishing, 2005. 452 p.

SCORM 2003 – Sharable Content Object Reference Model – ADL, the scorm content aggregation model. Version 1.3, Working draft 1. oct. 2003. Disponível em: <http://www.adlnet.gov/scorm/index.aspx> Acesso em: jun. 2008.

SHAHABI, C. et al. Knowledge discovery from user web-page navigation. In: RESEARCH ISSUES IN DATA ENGINEERING, 1997, Birmingham, UK. **Proceedings...** [S.l.:s.n.], 1997.

SILVA, M. J. **The case for a portuguese web search engine**. 2003. Relatório técnico - Universidade de Lisboa, Faculdade de Ciências. Disponível em: <http://www.di.fc.ul.pt/tech-reports/03-3.pdf>. Aceso em: jun. 2008.

SINTEK, M.; DECKER, S. Triple - an RDF Query, Inference, and Transformation Language. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, ISWC, 2002, Sardinia, Italy. **Revised Papers.** Berlin: Springer, 2002. p. 364 – 378. (Lecture Notes In Computer Science, v.2342).

SMITH, M. K.; WELTY, C.; MCGUINNESS, D. L. **OWL web ontology language guide.** Disponível em: <http://www.w3.org/TR/owl-guide>. Acesso em: jun. 2008.

SOUTO, M. A. M. et al. Towards an adaptative web training environment based on cognitive style of learning: an empirical approach In: INTERNATIONAL CONFERENCE ON ADAPTIVE HYPERMEDIA AND ADAPTIVE WEB BASED SYSTEMS, 2., 2002, Malaga. **Proceedings...** [S.l.:s.n.], 2002. p. 338-347.

SOUTO, M. A. M.; VERDIN, R.; OLIVEIRA, J. P. M. Modeling learner's cognitive abilities in the context of a web-based learning environmentadvances in web-based education: personalized learning environments. In: CHEN, S.; MAGOULAS, G. (Org.). **Advances in web-based education**: personalized learning, environments. Hershey, USA: IDEA, 2005.

SOWA, J. F. **Guided tour of ontology.** Disponível em: <http://www.jfsowa.com/ontology/index.htm>. Acesso em: jun. 2008.

SPILIOPOULOU M.; FAULSTICH L.C. WUM: a web utilization miner. In: EDBT WEBDB, 1998, Valencia. **Proceedings ...** [S.l.:s.n.], 1999.

STAAB, S. et al. Seal: a semantic portal with management functionality. In: CURRENT RESEARCH INFORMATION SYSTEMS, 2002, Kassel. **Proceedings ...** Disponível em: <http://www.aifb.uni-karlsruhe.de/~sst/Research/Publications/cris2002.pdf>. Acesso em: jun. 2008.

STASH, N.; CRISTEA, A. I.; DE BRA, P. Adaptation languages as vehicles of explicit intelligence in Adaptive Hypermedia. **J. Continuing Engineering Education and Life-Long Learning**, [S.l.], v. 17, n. 4/5, p. 319–336, 2007.

STASH, N. Adaptation to learning styles in e-l earning: approach evaluation. In: E-LEARN CONFERENCE, 2006, Honolulu, Hawaii. **Proceedings...** Chesapeake, VA: AACE, 2006. p. 284 – 291.

STOJANOVIC, L.; STAAB, S.; STUDER R. eLearning based on the Semanticweb. In: WORLD CONFERENCE ON THE WWW AND INTERNET, 2001, Orlando, USA. **Proceedings...** [S.l.:s.n.], 2001. p. 28 – 41.

STUMME, G.; BERENDT, B.; HOTHO, A. Usage mining for and on the semantic web. In: In: NATIONAL SCIENCE FOUNDATION WORKSHOP ON NEXT GENERATION DATA MINING, 2002. **Proceedings...**Baltimore, USA: [s.n.], 2002. 77-86.

SZUNDY, G. et al. Design and implementation of semantic web aplications. In: Workshop: aplication design, development and implementation issues in the semantic web; WORLD WIDE WEB CONFERENCE, 2004, New York, USA. **Proceedings…** [S.l.:s.n.], 2004.

TODIRASCU, A.; ROMARY, L.; BEKHOUCHE, D. Vulcain – An ontology-based information extraction system. In: INTERNATIONAL CONFERENCE ON APPLICATIONS OF NATURAL LANGUAGE TO INFORMATION SYSTEMS, 6., 2002, Stocholm. **Proceedings...** Berlin: Springer-Verlag, 2002. p. 64 – 75.

TRAN, T.; CIMIANO, P.; ANKOLEKAR, A. Rules for an ontology-based approach to adaptation. In: INTERNATIONAL WORKSHOP ON SEMANTIC MEDIA ADAPTATION AND PERSONALIZATION, 1., 2006, Athens. **Proceedings...** Athens: SMAP, 2006. p.49 – 54.

TRAN, T.; LEWEN, H.; HAASE, P. On the role and application of ontologies in information systems. In INTERNATIONAL CONFERENCE ON RESEARCH, INNOVATION & VISION FOR THE FUTURE, 5., 2007, Hanoi, Vietnam. **Proceedings...** [S.l.]: IEEE, 2007. p. 14-21.

TSANDILAS, T.; SCHRAEFEL, M. D. User-controlled link adaptation. In: CONFERENCE ON HYPERTEXT AND HYPERMEDIA, 15., 2003, Nottingham, UK. **Proceedings...** New York: ACM, 2003. p. 152 – 160.

VASILYEVA, E.; PECHENIZKIY, M.; DE BRA, P. Adaptation of feedback in e-learning systems at individual and group level. In: WORKSHOP ON PERSONALISATION IN E-LEARNING ENVIRONMENTS AT INDIVIDUAL AND GROUP LEVEL, AT THE USER MODELING CONFERENCE, 2007, Corfu, 2007. **Proceedings...** [S.l.:s.n.], 2007. p. 49-56.

VDOVJAK, R. et al. Engineering semantic web information systems in hera. In: **Journal of Web Engineering,** [S.l.], v. 2, n. 1-2, p. 3-26, 2003.

VIEIRA, R.; RIGO, S. J.Busca de informações auxiliada por ontologias. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2002, São Leopoldo. **Anais…** São Leopoldo: Ed. Unisinos, 2002. v. 1, p. 597-600.

WANG, J.; HAN, J.; PEI, J. CLOSET+: searching for the best strategies for mining frequent closed itemsets. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 9., 2003, Washington, DC. **Proceedings...** New York: ACM SIGKDD, 2003. p. 236 – 245.

WANG, K.; XU, C.; LIU, B. Clustering transactions using large items. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 1999, Kansas City, Missouri, United States. **Proceedings...** New York: ACM, 1999. p. 483-490.

WEBER, G.; SPECHT, M. User modeling and adaptive navigation support in www-based tutoring systems. In: INTERNATIONAL CONFERENCE ON USER MODELING, 6., 1997, Sardinia, IT. **Proceedings...** [S.l.:s.n.], 1997. p. 289 – 300.

WIELEMAKER, J.; SCHEREIBER, G.; WIELINGA, B. Prolog-based infrastructure for RDF: scalability and performance. In: SEMANTIC WEB STORAGE AND RETRIEVAL, SWAD-Europe, 2003, Vrije Universiteit, Amsterdam. **Workshop...** Amsterdam:[s.n.], 2003.

WILKS, Y.; BONTCHEVA, K. Tailoring automatically generated hypertext. **User Modeling And User-Adapted Interaction**, [S.l.], v. 15, n. 1-2, p. 135-168, Mar. 2005.

WOON Y. et al. Web Usage mining: algorithms and results. In: SCIME, A. (Ed.). **Web mining**: applications and techniques. London, UK: Idea Group Publishing, 2005. p. 373-394.

WOUKEU, A. et al. **Rethinking web design models**: requirements for addressing the content. 2003. Technical Report - Department of Electronics and Comnputer Science, University of Southampton.

WU, H. **A reference architecture for adaptive hypermedia applications.** Eindhoven: Technische Universiteit Eindhoven, 2002.

YEH, C. Development of an Ontology-based portal for digital archive services. In: INTERNATIONAL CONFERENCE ON DIGITAL ARCHIVE TECHNOLOGIES, 2002, Taipei. **Proceedings...** Taipei: Academia Sinica, Nankang, 2003. Disponível em: <http://www.iis.sinica.edu.tw/APEC02/Program/chingyeh.pdf > Acesso em: jun. 2008.

YERGEAU, F. et al. XML 1.1 - W3C Recommendation. 2004. Disponível em: <http://www.w3.org/XML/Core/#Publications>. Acesso em: jun. 2008.

XIAOQIU, T.; MIN, Y.; MIAOJUN, X. An effective technique for personalization recomendation based on access sequential patterns. In: ASIA-PACIFIC CONFERENCE ON SERVICES COMPUTING, 2006, Washington, DC, USA. **Proceedings...** [S.l.]:IEEE, 2006. p. 42 – 46.

ZAIANE, O. R. Web mining: concepts, practices and research. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 15., 2000, João Pessoa. **Tutorial.** João Pessoa: CEFET-PB; Porto Alegre: PUCRS, 2000. p. 410-474.

ZAKI, M. SPADE: an eficient algorithm for mining frequent sequences. **Machine Learning**, [S.l.], n. 42, p. 31-60, 2001.

ZHONG, N.; LI, Y. Mining Ontology for Automatically Acquiring Web User Information Needs. **IEEE Trans. Knowl Data Eng,** [S.l.], v. 18, n. 4, p. 554-568, 2006.

ZHOU, Y.; MOBASHER, B. Web User segmentation based on a mixture of factor analyzers. In: ECWeb, 2006, Krakow, USA. **Proceedings...** Berlin: Springer, 2006. p. 11 – 20.

ZIMMERMANN, A. et al. Personalization and context management. **User Modeling and User-Adapted Interaction,** [S.l.], v. 15, n.2, p. 275–302, 2005.