PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL

FACULDADE DE INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Mecanismos de Apoio a Interpretação e Recuperação de Padrões do Uso da *Web* Baseados em Ontologia de Domínio

MARIÂNGELA VANZIN

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre, pelo Programa de Pós Graduação em Ciência da Computação da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Profª. Drª. Karin Becker

Porto Alegre

2004

AGRADECIMENTOS

À minha orientadora, professora Karin Becker, por toda a sua dedicação, ensinamentos e conselhos. Saiba que os constantes desafios lançados me fizeram crescer muito como pessoa.

Alcides Vanzin, Merce Bergamin Vanzin, Emerson Vanzin, Renata Gheno e Andressa Vanzin. O carinho e compreensão de todos vocês que sempre me apoiaram, mesmo longe.

A professora e amiga Mara Abel por todos os conselhos e pensamentos positivos. A mente é mesmo poderosa.

Ao professor Marcelo Blois pelo intenso apoio, compreensão e paciência nos momentos em que tudo parecia desabar. Obrigada também pelas críticas que contribuíram para o meu crescimento.

Ao grupo MIGAS (Taisa Carla Novello, Elceni Gelain, Laura Mastella). Cada uma de vocês foi fundamental para esta conquista. Aprendemos juntas que é o atrito que nos faz andar, que as dificuldades nos fazem crescer. E como crescemos nestes últimos anos.

Aos colegas Giliane Redolfi e Cristiano Bertolini por terem compartilhado muitos momentos agradáveis e também estressantes durantes o período do mestrado. Momentos estes que ficarão registrados na memória e no coração.

Ao André da Fonte Lopes, bolsista do projeto financiado pela FAPERGS, pelo comprometimento, dedicação e excelente trabalho realizado no desenvolvimento do protótipo definido com parte deste trabalho.

Ao Convênio Dell-PUCRS por viabilizarem a bolsa de estudos durante os quase dois anos de mestrado.

Ao Programa de Pós-Graduação em Ciência da Computação e a todos os professores dos quais pude conviver durante estes dois anos.

Ao departamento de Educação a Distância (PUCRS Virtual), por fornecer os dados e assim, permitir o desenvolvimento deste trabalho.

"No auge da dificuldade, você está a um passo de sua meta. Cada vez que fracassar, lembre-se que está mais perto da concretização de seu sonho".

Do livro A verdade da Vida, vol 20 – Masaharu Taniguchi

RESUMO

O processo de Mineração do Uso da *Web* (MUW) permite extrair padrões de navegação a partir de arquivos de *log* armazenados nos servidores *Web*. O processo de MUW tem demonstrado utilidade para os domínios mais diversificados, porém existem problemas que comprometem a sua efetividade.

O processo de MUW é composto pelas fases de Preparação de Dados, Mineração de Dados e Análise de Padrões. Esta pesquisa enfoca a fase de Análise de Padrões, que tem por objetivo identificar padrões relevantes para o domínio da aplicação dentre os retornados da fase de Mineração de Dados. Problemas encontrados nesta fase referem-se à dificuldade de interpretação e recuperação de padrões. Muitas vezes, os padrões são incompreensíveis para o analista devido à falta de semântica na representação destes, formados geralmente por URL's que nem sempre expressam intuitivamente os eventos de domínio disponíveis no site. A dificuldade encontrada quanto à recuperação de padrões deve-se à grande quantidade de padrões resultantes de algoritmos tradicionais de Mineração de Dados, na maioria das vezes desinteressantes e redundantes.

Perante estes problemas identificados, este trabalho propõe mecanismos de apoio à interpretação e à recuperação de padrões do uso da *Web*, através da exploração do conhecimento representado por Ontologia de Domínio. Os padrões considerados neste trabalho são padrões seqüenciais de navegação. Os mecanismos de interpretação propostos permitem: representar os padrões seqüenciais através de padrões conceituais, que expressem os eventos de domínio envolvidos; e permitir a análise exploratória e interativa destes padrões aprofundando a compreensão e explorando padrões relacionados. Os mecanismos de recuperação visam: a geração de agrupamentos de padrões restringindo o escopo da busca; definir filtros de acordo com o interesse do analista, utilizando a Ontologia de Domínio como apoio; e finalmente recuperar padrões similares ao interesse especificado nos filtros.

Para avaliação da abordagem proposta, foi desenvolvido um ambiente de apoio à fase de Análise de Padrões que incorpora os mecanismos de interpretação e recuperação de padrões. Este ambiente foi utilizado num estudo de caso que aplica o processo de MUW ao domínio da Educação a Distância.

Palavras-Chaves: Análise de padrões, Interpretação e recuperação de padrões do uso da *Web*, Mineração do Uso da *Web*.

ABSTRACT

Web Usage Mining (WUM) aims to extract navigation usage patterns from Web server log files. While WUM techniques were proven to be useful, many problems need to be solved for their effective application.

The WUM process is composed by three generic phases: preprocessing, mining and pattern analysis. This research focuses on the pattern analysis phase, which aims at identifying, from the patterns yielded by the mining phase, the relevant ones for the application domain. Problems found in this phase are related to pattern interpretation and retrieval. Usually, patterns are incomprehensible for the analyst because there is semantic gap between URLs and the events performed by users in a site. Pattern retrieval is critical because mining algorithms yield a huge number of patterns and most of them are useless and redundant.

This research proposes ontology-based mechanisms targeted at the interpretation and retrieval of sequential navigation patterns. The interpretation approach allows: a) the representation of patterns in a more intuitive form; b) interactive pattern rummaging for improving the comprehension of the meaning of a pattern, as well as discovering related patterns. The retrieval approach allows: a) the definition of filters based on conceptual, structural and statistical constraints established over the concepts of the ontology; b) the search for patterns that either match the user-specified filter or are similar to it in some degree; and c) the clustering of related patterns to set focus on the interpretation activity.

The ontology-based mechanisms constitute a supporting environment for the pattern analysis phase, for which a prototype was developed. The use of these mechanisms is illustrated and analyzed in a case study in the Distance Education domain.

Key-words: Web Usage Pattern Analysis, Web Usage Pattern interpretation and retrieval, Web Usage Mining;

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CLF *Common Log Format*

EAD Educação a Distância

ELF *Extend Log File Format*

GVSM *Generalized Vector Space Model*

HTML *HyperText Markup Language*

HTTP *Hypertext Transfer Protocol*

KDD *Knowledge Discovery in Database*

MUW Mineração do Uso da Web

OWL *Web Ontology Language*

OLAM *On-line Analytical Mining*

OLAP *On-line Analytical Processing*

PUCRS Pontifícia Universidade Católica do Rio Grande do Sul

RDFS *Resource Description Framework Schema*

URL *Uniform Resource Locate*

WebCT *Web Course Tool*

WUM *Web Utilization Miner*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contexto Geral

O fluxo incessante de acessos às páginas da Internet via *Web* reflete os conteúdos mais diversos, bem como costumes e necessidades pessoais ainda mais distintos, resultando em padrões de utilização extremamente ricos e diversificados. Compreender estes padrões de navegação que impulsionam os usuários durante a navegação em um *site* tem motivado grande quantidade de pesquisadores em áreas tão diversas como redes de computadores, banco de dados, inteligência artificial, entre outras.

A Mineração do Uso da *Web* (MUW) é a área que se dedica à extração de padrões que revelam o comportamento de navegação dos usuários na *Web*. Estes padrões são obtidos, principalmente, a partir da análise de *logs* de acessos mantidos em servidores *Web*. Os *logs* registram URLs referente às páginas *Web* e arquivos acessados pelos usuários durante a visita a um *site*.

O processo de MUW é composto por três etapas distintas: Preparação de Dados, Mineração de Dados e Análise de Padrões [COO99]. Esta pesquisa enfoca a fase de Análise de Padrões, a qual aborda a identificação de padrões relevantes ao domínio da aplicação dentre os obtidos através da aplicação de técnicas de Mineração de Dados.

Padrões relevantes podem, por exemplo, auxiliar as organizações a planejar estratégias de marketing de venda de produtos, a conhecer o tempo de vida dos seus clientes, efetivar campanhas promocionais, etc [COO97]. Estes padrões também fornecem subsídios aos projetistas na tomada de decisões referentes à estrutura ou topologia utilizada na estruturação do *site*. Sem o conhecimento descoberto a partir da MUW, o projeto de um *site* dependeria apenas das suposições dos projetistas em relação às expectativas e modelos comportamentais dos usuários. Ainda, os padrões de navegação podem ser utilizados com o intuito de propor melhoras no conteúdo disponibilizado em um *site*.

Assim, a MUW torna-se útil quando padrões que agregam valor ao domínio da aplicação são identificados (i.e. conhecimento), o que não constitui é uma atividade trivial. Ela depende da interpretação dos padrões e da recuperação dos relevantes ao domínio. Entende-se por interpretação de padrões as atividades executadas para o entendimento das

informações expressas por um padrão. Já recuperação de padrões refere-se às atividades realizadas para encontrar padrões em meio a um conjunto destes.

Analistas enfrentam dificuldade na interpretação dos padrões descobertos na MUW. Estes são usualmente representados por coleções de URLs que nem sempre expressam de forma evidente e intuitiva os serviços e conteúdos que impulsionam a navegação dos usuários pelas páginas *Web*. Desta forma, a interpretação dos padrões de navegação pode ser prejudicada uma vez que não há interesse em padrões formados por URLs e sim em padrões que expressem o conteúdo e serviço envolvidos neles. Neste contexto, pesquisas vêm sendo realizadas (e.g. [OBE03, DAI02, BER00]) visando associar a MUW com representações do conhecimento do domínio que especificam a semântica das requisições às páginas feitas pelos usuários. O objetivo principal destas abordagens é obter resultados com maior semântica e facilitar o processo de inspeção e análise dos padrões interessantes.

Além da falta de representação semântica dos padrões, outro problema na análise refere-se à dificuldade de recuperação dos padrões relevantes devido à existência de um grande número destes, resultantes das técnicas de Mineração de Dados. Em meio a tantos padrões, muitos são irrelevantes por representarem um conhecimento de senso comum e ainda diversos são redundantes. Para amenizar este problema, abordagens propõem diferentes métodos para redução do número de padrões de acordo com o interesse especificado, sendo de responsabilidade do analista definir o que é relevante ao domínio da aplicação através de medidas de interesse objetivas (e.g [AGR93, AGR94a]), crenças do domínio (e.g [SIL96, COO03, POH03]) e filtros (e.g [KLE94, SPI98]).

Perante os problemas identificados na fase de Análise de Padrões, este trabalho tem como objetivo propor mecanismos que facilitem as atividades de interpretação e recuperação de padrões ao escopo das aplicações de MUW através da exploração do conhecimento representado por Ontologia de Domínio. Resultados preliminares desta pesquisa foram relatados em [BEC03, VAN04, VAN04a].

1.2 Objetivo do Trabalho

O objetivo principal deste trabalho é propor mecanismos que facilitem a interpretação e recuperação de padrões seqüenciais de navegação através da utilização de Ontologia de Domínio disponibilizada previamente. Estes mecanismos referem-se a duas dificuldades

principais encontradas na fase de Análise de Padrões: a grande quantidade de padrões resultantes da aplicação de algoritmos para a busca de padrões seqüenciais e a falta de semântica neles representada.

Os objetivos específicos são:

- propor mecanismos que facilitem a interpretação de padrões através da representação de padrões seqüenciais de URLs em padrões conceituais;
- propor mecanismos que facilitem a interpretação dos padrões conceituais através da análise exploratória da semântica destes padrões conceituais;
- propor mecanismos que auxiliem a recuperação de padrões conceituais através da definição de filtros com o uso de Ontologia de Domínio;
- definir um ambiente de apoio à fase de Análise de Padrões que incorpore estes mecanismos, permitindo uma avaliação sobre a utilidade dos mesmos.

1.3 Método de Pesquisa

A Figura 1 representa as principais atividades desenvolvidas na condução deste trabalho. Inicialmente foram estudadas as principais fontes sobre o processo de MUW, fornecendo um entendimento geral. Posteriormente, o estudo restringiu-se à fase de Análise de Padrões, onde o problema motivador para esta pesquisa foi identificado. Ele refere-se à dificuldade de interpretação e recuperação padrões relevantes ao domínio.

Identificado o problema, foi possível definir o objetivo principal deste trabalho e direcionar a pesquisa as abordagens relacionadas ao problema. Primeiramente o estudo enfocou os trabalhos que propõem suporte à recuperação de padrões relevantes ao domínio da aplicação através da especificação de medidas e filtros. Com este estudo verificou-se que a atividade de identificar padrões interessantes relaciona-se à redução do número destes, o que nem sempre é eficiente pois padrões relevantes podem ser desconsiderados. Outro ponto verificado é que estas abordagens não forneciam suporte a interpretação de padrões, sendo de responsabilidade do analista, utilizar seu conhecimento sobre o domínio para interpretar e avaliar os padrões. Desta forma, a continuidade da pesquisa focou-se em abordagens que propunham a integração do conhecimento do domínio ao processo de MUW tornando a

atividade de Análise dos Padrões menos dependente do conhecimento do domínio detido pelos analistas.

Estas abordagens forneceram subsídios para a definição dos mecanismos de apoio a interpretação e recuperação de padrões do uso da Web. Posteriormente à definição dos mecanismos de suporte à fase de Análise de Padrões, o passo seguinte consistiu no desenvolvimento de um ambiente de apoio através de um protótipo que implementa estes mecanismos. Visando avaliá-lo, este ambiente foi utilizado num estudo de caso no domínio da Educação a Distância (EAD) que possibilitou a uma comparação do processo de MUW aplicado no domínio da EAD utilizando o ambiente de apoio à fase de Análise de Padrões, com outro processo de MUW no mesmo domínio, porém sem apoio a esta fase.

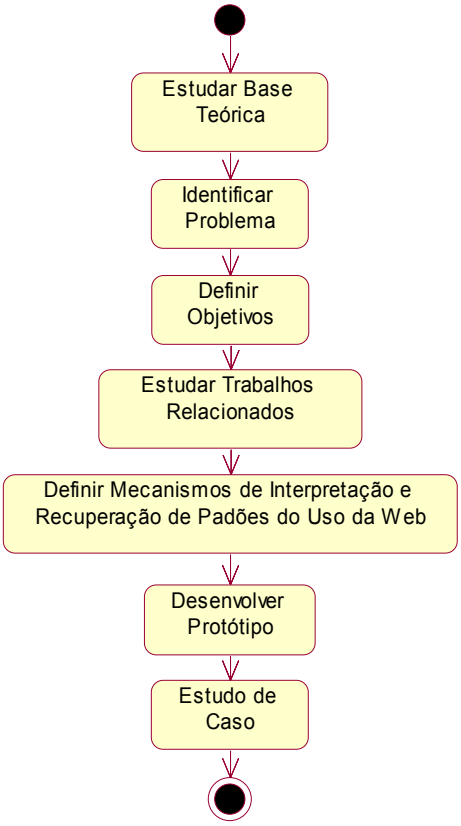


Figura 1: Atividades desenvolvidas na condução da pesquisa

1.4 Estrutura do Trabalho

Este trabalho está dividido em 9 capítulos. O Capítulo 2 apresenta as fases que compõem o processo de MUW, focando-se na fase de Análise de Padrões, cujos principais problemas são relatados. O Capítulo 3 discorre sobre as principais abordagens que visam auxiliar a interpretação e recuperação de padrões.

O Capítulo 4 apresenta os principais objetivos da abordagem proposta, assim como os requisitos para a representação da Ontologia de Domínio para a interpretação e recuperação de padrões seqüenciais. Também são apresentadas algumas particularidades quanto às fases do processo de MUW.

O Capítulo 5 e 6 descrevem os mecanismos propostos para auxiliar a fase de Análise de Padrões. O Capítulo 5 descreve os mecanismos de interpretação de padrões que se referem aos padrões seqüenciais conceituais e a análise exploratória destes. Já o Capítulo 6 apresenta os mecanismos voltados à recuperação de padrões. Estes possibilitam a geração de agrupamentos de padrões focando o escopo da busca; definição de filtros de interesse, utilizando a Ontologia de Domínio como apoio; e finalmente a definição de mecanismos de busca por padrões, envolvendo ou não medidas de similaridade.

O Capítulo 7 descreve o ambiente de apoio proposto para avaliar os mecanismos de interpretação e recuperação de padrões durante a fase de Análise de Padrões. O Capítulo 8 apresenta um estudo de caso realizado no contexto da Educação a Distância para avaliar os mecanismos propostos. Este último capítulo também apresenta um comparativo entre dois processos de MUW no ambiente de EAD, sendo um com apoio à fase de Análise de Padrões.

O Capítulo 9 discorre sobre as conclusões, limitações e trabalhos futuros. Posteriormente, encontram-se as referências bibliográficas pesquisadas e os demais anexos.

2 MINERAÇÃO DE DADOS DA *WEB*

Este capítulo apresenta as áreas distintas da Mineração de Dados da Web, destacando a Mineração do Uso da Web (MUW). As fases que compõem o processo de MUW são detalhadas, focando-se na fase de Análise de Padrões, cujos principais problemas são relatados.

A Mineração de Dados (*Data Mining*) é considerada parte de um grande processo de descoberta de conhecimento em banco de dados (KDD – *Knowledge Discovery in Database*). KDD corresponde à exploração e análise, por meio automático ou semi-automático, de grande quantidade de dados, com o propósito de descobrir regras e padrões significativos [BER97]. A partir dos esforços da Mineração de Dados associados com a *Web,* surgiu uma nova área de aplicação denominada de Mineração da *Web* (*Web Mining*), a qual visa utilizar técnicas de Mineração de Dados para descoberta e análise de informações úteis da *Web.* As técnicas da Mineração de Dados da *Web* visam descobrir conhecimento novo e relevante dos dados da *Web,* onde a partir das informações descobertas seja possível demonstrar características, comportamentos, tendências e padrões de navegação do usuário da *Web* [COO99, SRI00].

A Mineração da *Web* se divide em três categorias de acordo com a parte da *Web* a ser minerada: Mineração de Conteúdo (*Web Content Mining*), Mineração de Estrutura (*Web Structure Mining*) e Mineração do Uso (*Web Usage Mining*) [KOS00, SRI00]. A distinção entre estas categorias está representada na Figura 2, extraída de Berendt *et al.* [BER02a].

Figura 2: As três áreas da Mineração de Dados da *Web*

A Mineração do Conteúdo trata da descoberta de informações úteis referente ao conteúdo, dados, documentos e serviços da *Web*. Cabe salientar que o conteúdo da *Web* não se restringe a texto ou hipertexto, também abrangendo uma ampla variação de tipos de dados, tais como áudio, vídeo, dados simbólicos, metadados e vínculos de hipertexto.

Já Mineração da Estrutura da *Web* foca-se nas informações que existem de forma implícita entre os documentos, procurando descobrir um modelo sobre a estrutura de *links* da *Web*. O modelo é baseado na topologia de *hyperlinks*, podendo ser utilizado para categorizar conjuntos de páginas *Web* e ser útil na geração de informações similares e relacionadas entre diferentes *sites Web*. Assim, este tipo de mineração busca encontrar a estrutura de *hyperlinks* interna à própria *Web*.

A Mineração do Uso da *Web* (MUW) centra-se na descoberta de padrões de uso da *Web*. A MUW é descrita em detalhe na seção seguinte por ser o foco principal desta pesquisa, mas cabe ressaltar que ela relaciona-se diretamente com as demais áreas da Mineração de Dados da *Web,* afinal o comportamento de navegação é dependente da estrutura do *site Web* e do conteúdo disponibilizado nele [COO03].

2.1 Mineração do Uso da *Web*

A MUW centra-se na aplicação de técnicas que possam detectar padrões de comportamento dos usuários enquanto eles interagem com *sites* disponíveis na *Web* [KOS00, SRI00]. A descoberta de padrões de navegação pelas páginas *Web* proporciona um entendimento mais aprofundado do comportamento dos usuários bem como da estrutura e do conteúdo das páginas *Web* envolvidos na interação dos usuários com o *site Web* [MOB96, SRI00].

Freqüentemente as organizações desenvolvem seus *sites* da forma que seus projetistas consideram mais apropriada para os usuários. A coleta e posterior análise dos dados referentes aos acessos podem esclarecer a natureza do tráfego no *site*, auxiliando na compreensão do comportamento dos usuários, e permitindo assim verificar se o *site* está eficientemente projetado e organizado. Segundo Cooley *et al.* [COO99], a mineração do uso da *Web* proporciona um equilíbrio entre a visão do projetista de como o *site* deveria ser usado em contraste com a maneira como os usuários navegam através dele.

A análise dos dados obtidos através da aplicação de técnicas de mineração do uso da *Web* tem demonstrando ser eficiente nos mais variados domínios, abrangendo desde o comércio eletrônico (e.g. [SRI00, KOS00, COO99]) até a Educação a distância (e.g. [MAC03, ZAI01]). Por exemplo, analisando os padrões descobertos na área do comércio eletrônico é possível auxiliar as organizações a planejar estratégias de marketing de venda de produtos, a conhecer o tempo de vida dos seus clientes, efetivar campanhas promocionais, entender a motivação dos usuários durante a navegação, construir *sites* adaptativos, etc. Já no contexto da Educação a Distância, a MUW pode ser utilizada para sugerir melhorias quanto ao conteúdo e estrutura de um curso, assim como avaliar a efetividade do projeto de um *site* de acordo com os diferentes processos de aprendizagem.

2.2 Processo de Mineração do Uso da *Web*

O processo de MUW é composto por três etapas distintas, cada uma com suas próprias características, métodos, entradas e saídas [COO99]. São elas:

- Preparação de dados: inclui seleção e limpeza de dados, identificação de usuários, sessões e transações, complemento do caminho de acesso às páginas *Web* entre outras atividades;
- Mineração de Dados: aplicação de algoritmos de Mineração de Dados gerando regras, padrões e estatísticas;
- Análise de Padrões: descoberta de regras e padrões interessantes.

A Figura 3, adaptada de [COO97], ilustra cada uma das fases, assim como os principais elementos que compõem o processo de MUW. Cabe salientar que o processo de MUW é altamente iterativo e interativo, podendo envolver contínuos retornos a uma ou mais fases.

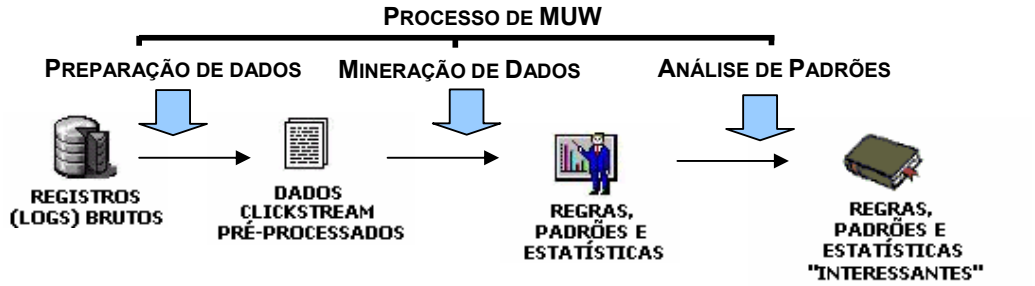


Figura 3: Fases da Mineração do Uso da *Web*.

2.2.1 Preparação dos Dados

Técnicas de mineração de uso da *Web* são aplicadas principalmente sobre conjunto de sessões ou transações de usuários, informações estas contidas principalmente em arquivos de *log* armazenados nos servidores *Web.* Uma sessão de usuário é composta por todas as páginas acessadas por um determinado usuário durante uma visita ao *site*. Uma transação é um agrupamento semanticamente significativo de páginas contidas em uma sessão. Outras fontes de dados compreendem os formulários de registro de visitantes, os dados oriundos de scripts e as informações da autenticação de usuários [COO97].

As informações contidas no arquivo de *log* são adicionadas automaticamente quando o usuário realiza uma requisição ao servidor *Web.* Por exemplo, ao visitar uma página *Web,* as informações sobre o acesso são adicionadas no arquivo de *log*. Arquivos de *log* geralmente seguem um formato padronizado, chamado CLF (*Common Log Format*), ou uma variação deste formato, chamada ELF (*Extend Log File Format*) [W3C03]. O arquivo no formato CLF registra todo o histórico das páginas e arquivos acessados pelos usuários. Cada registro deste histórico contém as seguintes informações: endereço IP que gerou a requisição; data e horário da requisição; método da requisição (*Get* ou *Post*); resultado da requisição (sucesso, falha, erro e etc); tamanho dos dados em número de *bytes*; URL da página acessada; e identificação do usuário. A Figura 4 ilustra uma amostra dos dados extraídos de um *log* armazenado em um servidor *Web.*

Arquivo de *Log*

```
200.176.25.110 - aluno1 [10/Jan/2002:00:00:06 -0200] "GET /ESP_SE_01130/competencia/07_01/conselhos.doc HTTP/1.1" 200 31744
200.176.8.249  - -      [10/Jan/2002:00:10:17 -0200] "GET / HTTP/1.1" 200 189
200.176.8.249  - -      [10/Jan/2002:00:10:18 -0200] "GET /webct/public/home.pl HTTP/1.1" 200 1977
200.176.8.249  - aluno2 [10/Jan/2002:00:10:39 -0200] "GET /webct/homearea/homearea HTTP/1.1" 200 20032
200.176.8.249  - -      [10/Jan/2002:00:11:20 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/dropbox_stud_home.pl HTTP/1.1" 401 899
200.248.5.164  - -      [10/Jan/2002:00:11:21 -0200] "GET /webct/homearea/homearea HTTP/1.1" 401 866
200.176.8.249  - aluno2 [10/Jan/2002:00:11:25 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/dropbox_stud_home.pl HTTP/1.1" 200 28552
```

Figura 4: Amostra de dados do *log* de acesso no formato CLF

No protocolo HTTP (*HyperText Transfer Protocol*) [W3C03], um acesso a uma simples página *Web* provoca o registro de várias entradas de *log* no servidor considerando os diversos arquivos necessários à visualização da página, sendo estes imagens e estilos, scripts e outros arquivos carregados juntamente com a página. Em geral, somente as entradas de *log* associadas aos acessos às páginas HTML (*HyperText Markup Language*) serão de interesse para o processo de MUW, pois os demais arquivos, especialmente imagens, não são

explicitamente solicitados pelo usuário. Neste contexto insere-se o conceito de visão de página que é definida como sendo todos os arquivos que contribuem para compor uma página tal como visualizada pelo usuário, como resultado de um único click do usuário. A Figura 5 representa uma visão de página que gerou quatro entradas no *log* armazenado no servidor *Web,* sendo a primeira relevante para a MUW por referenciar a página *Web.*

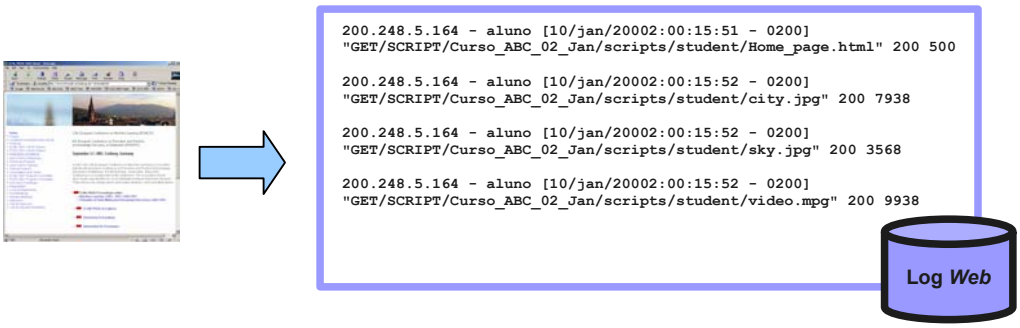


Figura 5: Visão de Página *Web* e *Log*

Os dados contidos em um *log* de servidor *Web* não representam com total confiabilidade os acessos dos usuários. Isso não se deve apenas ao fato da presença de grande número de itens irrelevantes, mas também pela freqüente ausência de identificação dos usuários, a inexistência de registros referentes a visitas a inúmeras páginas e a dificuldade de identificar com precisão o início e o fim de uma sessão de usuário. O uso do *cache* e servidores *proxy* estão entre os fatores que contribuem para esta falta de confiabilidade [COO99].

Perante estas inconsistências, as fontes de dados necessitam passar pela fase de preparação de dados, incluindo o desenvolvimento de um modelo de dados para os *logs* de acesso; a filtragem e a limpeza dos dados brutos; a identificação de usuários, sessões e transações; o complemento do caminho de acesso às páginas *Web.*

Outro aspecto relevante observado nas informações armazenadas nos *logs* é a falta de representatividade semântica das URLs em relação aos serviços e conteúdos oferecidos pelas páginas *Web.* Por exemplo, a URL *"/SCRIPT/Curso_ABC/scripts/student/serve_bulletin?COMPOSE+Main*" não expressa claramente qual o evento ocorrido no *site*, e como este relaciona-se com os conteúdos e serviços oferecidos. Este conhecimento geralmente pertence aos projetistas do *site* ou aos especialistas do domínio. Para superar o problema, tipicamente a fase de Preparação de Dados inclui o enriquecimento semântico dos

dados, em particular sobre as páginas acessadas, de acordo com o conhecimento extraído do domínio (e.g. OBE03, DAI02, BER02a). Oberle *et al.* [OBE03] chama estes *logs* de *logs* semânticos. O objetivo principal é facilitar a interpretação dos resultados do processo de MUW.

Cabe citar que o esforço despendido nesta fase pode chegar até 80% do esforço total no processo sendo os resultados das fases subseqüentes altamente dependentes da maneira como os dados são preparados. No tocante ao enriquecimento dos dados, os principais objetivos são obter padrões mais representativos, a facilitar a interpretação dos resultados da fase de Mineração de Dados.

2.2.2 Mineração de Dados

A fase de Mineração de Dados, também conhecida como Descoberta de Padrões, compreende a aplicação de técnicas de mineração sobre os dados pré-processados resultantes da fase anterior [SRI00]. Diversas são as técnicas de Mineração de Dados disponíveis. Dentre elas cabe citar regras associativas, agrupamentos, classificação e padrões seqüenciais. A técnica que enfoca a descoberta de padrões seqüenciais foi a selecionada para esta pesquisa por revelar padrões de acesso às páginas *Web* obedecendo a uma determinada seqüência temporal.

Um padrão seqüencial é formado por um conjunto de itens que obedecem a uma seqüência temporal. Tipicamente, padrões seqüenciais são associados a uma medida de suporte que corresponde ao percentual de seqüências que contêm um determinado padrão [AGR94a, SRI95]. No domínio da *Web,* os itens que compõem um padrão seqüencial geralmente são representados por URLs que correspondem a acessos às páginas *Web,* e o suporte é dado pelo percentual de sessões de usuários que contêm um determinado caminho de navegação.

A Figura 6 representa um padrão seqüencial extraído de um *site* de uma livraria on-line. Este padrão é composto por 4 URLs, e está associado a um suporte de 80%, isto é, 80% das sessões de navegação existentes no *log* indicam que usuários acessaram a página principal, requisitaram informações referentes ao item 12 e posteriormente ao item 45, e finalmente confirmam a compra destes produtos. O acesso a estas URLs aconteceu nesta ordem, mas não são necessariamente consecutivos. Cabe ressaltar que devido à pobreza de

representação semântica das URLs, a interpretação do padrão seqüencial neste caso, depende do auxílio de um especialista de domínio ou projetista do *site*.

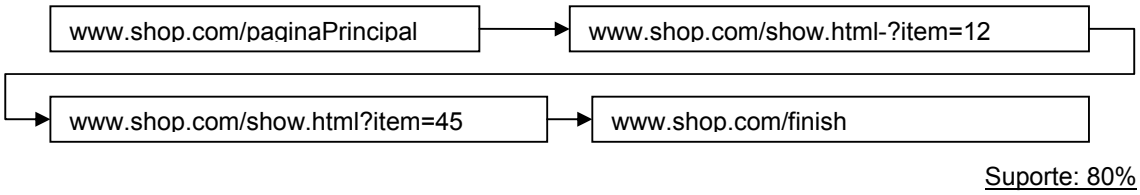


Figura 6: Padrão Seqüencial

Nas propostas de Agrawal e Srikant [AGR94a, SRI95] e Mannilla *et al*. [MAN95], a geração de padrões seqüenciais é feita sobre um banco de transações, visando encontrar padrões seqüenciais que ocorrem com uma certa freqüência. Spiliopoulou *et al.* [SPI98] propõem um algoritmo para geração de padrões seqüenciais voltado às especificidades da *Web.*

O algoritmo selecionado para esta pesquisa foi o *AprioriAll* proposto em [AGR94a] por ser um algoritmo tradicional de Mineração de Dados, já aplicado em diversos domínios. Uma implementação deste algoritmo se encontra disponível na ferramenta *Intelligent Miner* [IBM04]. O algoritmo é descrito com detalhes na próxima subseção.

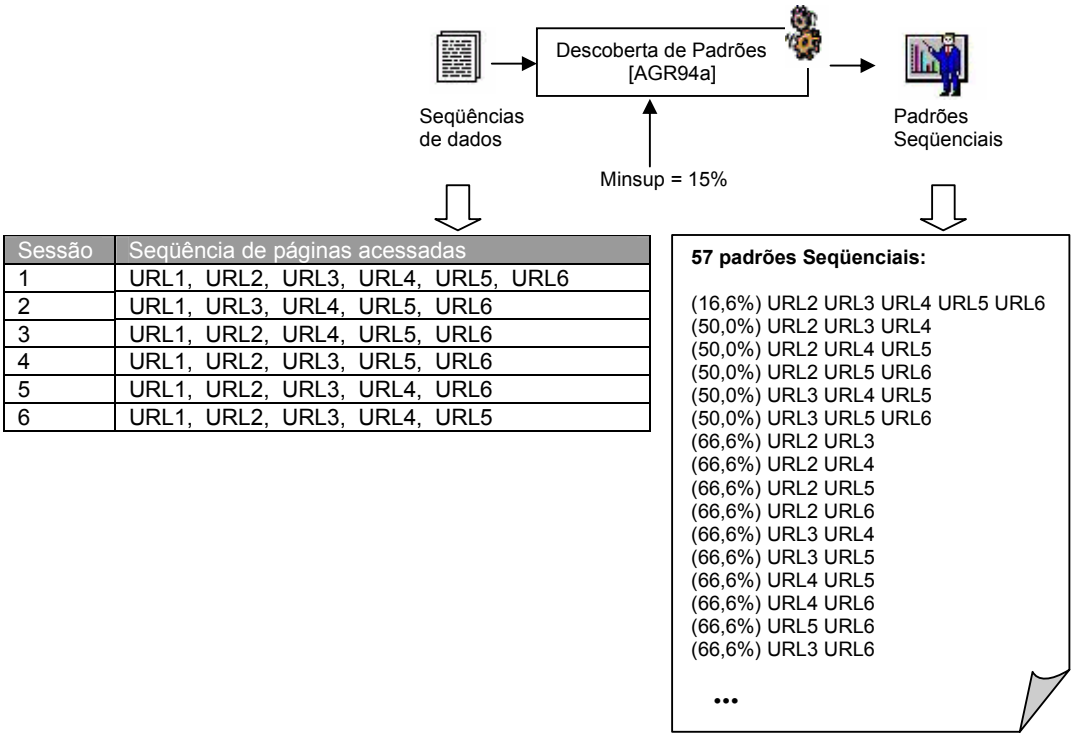
2.2.2.1 *Algoritmo* AprioriAll

O algoritmo *AprioriAll* proposto por Agrawal e Srikant [AGR94a] recebe como entrada as chamadas seqüências de dados, formadas por um ou mais itens. Para o processo de MUW, estes itens representam URLs acessadas pelos usuários cujas seqüências constituem sessões de navegação. Também, é necessário que o usuário especifique um valor mínimo para o suporte (*minsup*), lembrando que suporte é considerado como o percentual de seqüências de dados que contêm um determinado padrão. Desta forma, os padrões seqüenciais resultantes possuem um suporte maior ou igual ao valor do *minsup* especificado.

A Figura 7 representa os resultados obtidos pela geração de padrões seqüenciais de acordo com o algoritmo *AprioriAll*. Considera-se como entrada um conjunto de 6 seqüências de dados, as quais representam acessos às páginas *Web* de um determinado *site*. Neste exemplo, uma seqüência de dados é composta pelo conjunto ordenado de acessos às páginas *Web* realizadas por um usuário, isto é, uma sessão de navegação de um usuário (ou uma

transação contida nesta). Por exemplo, a seqüência de dados do usuário 6 é composta por 5 URLs, acessadas nesta ordem: “URL1 - URL2 - URL3 - URL4 - URL5”.

Os itens destas 6 seqüências de dados referem-se a 6 URLs distintas, oferecidas pelo *site*. O valor para *minsup* especificado foi 15% (i.e., contido em pelo menos uma sessão), assim muitos padrões foram retornados.



Sessão	Seqüência de páginas acessadas
1	URL1, URL2, URL3, URL4, URL5, URL6
2	URL1, URL3, URL4, URL5, URL6
3	URL1, URL2, URL4, URL5, URL6
4	URL1, URL2, URL3, URL5, URL6
5	URL1, URL2, URL3, URL4, URL6
6	URL1, URL2, URL3, URL4, URL5

Figura 7: Descoberta de Padrões Seqüenciais

O algoritmo *AprioriAll* é baseado na propriedade *Apriori*, desenvolvida originalmente para algoritmos de associação [AGR93]. Na técnica de associação, a propriedade *Apriori* é utilizada para encontrar conjuntos de itens freqüentes, isto é, que possuam suporte acima do mínimo estabelecido pelo usuário. No *AproriAll*, esta propriedade é utilizada para encontrar conjunto de itens freqüentes, bem como seqüências freqüentes. Originalmente, o *AprioriAll* foi proposto para encontrar seqüências em transações de itens comprados por clientes, principalmente do domínio de supermercados, comércio eletrônico, etc.

O algoritmo *AprioriAll* é constituído por 5 fases:

- Fase de Ordenação: Os dados de entrada são ordenados por um atributo agrupador (e.g. cliente, sessão) e pelo momento de ocorrência das transações (e.g. ordem de acesso as URLs);
- Fase dos Itens Freqüentes: Todos os itens ou conjunto de itens que possuam suporte maior ou igual ao valor do *minsup* devem ser identificados, ou seja, para que um item seja freqüente, o percentual de transações nas quais ele está contido deve ser maior que o *minsup*.
- Fase de Transformação: Visando otimizar o tempo de resposta do algoritmo, todos os itens freqüentes das seqüências de dados são mapeados para números inteiros.
- Fase da Seqüência: Todas as seqüências candidatas são geradas através da combinação exaustiva dos itens freqüentes, obedecendo ao critério temporal e o valor de *minsup*. Alguns algoritmos derivados do *AprioriAll* (e.g. *AprioriSome* e *DynamicSome*) apresentam variações quanto à fase de seqüência.
- Fase Maximal: Nesta fase, são encontradas as seqüências maximais que estão contidas no conjunto total de seqüências geradas, de forma a reduzir o número de padrões seqüenciais. Uma seqüência é maximal quando ela não está contida em nenhuma outra seqüência, ou seja, ela não é uma subseqüência de nenhuma outra seqüência. Embora a proposta do *AprioriAll* original inclua esta fase, muitas vezes o usuário pode estar interessado no suporte específico das subseqüências do padrão maximal. Provavelmente, por esta razão, as implementações conhecidas destes algoritmos, inclusive a da própria ferramenta *Intelligent Miner* que disponibiliza o *AprioriAll*, não implementam esta fase.

Aplicando o algoritmo *AprioriAll* sobre os dados de entrada descritos na Figura 7, foram obtidos 57 padrões seqüenciais, o que representa um número elevado, considerando o conjunto de dados de entrada. Vale lembrar que o processo de MUW quando aplicado a domínios reais considera milhares de seqüências de dados como entrada.

A Figura 7 apresenta alguns padrões gerados com o seu respectivo suporte especificado entre parênteses. Por exemplo, o último padrão (“URL3 - URL6”) indica que 66,6% das seqüências de dados de entrada suportam o acesso à URL3 seguido (imediatamente

ou não) pela URL6. Analisando as seqüências de dados observa-se que este padrão é verificado no caminho de navegação das sessões 1, 2, 4 e 5.

Como os padrões seqüenciais são resultantes da combinação entre os itens freqüentes, muitos deles expressam informações redundantes. Nota-se, por exemplo, que o acesso à URL2 seguido pelo acesso à URL4 é uma seqüência contida em 4 dos padrões apresentados. Por outro lado, se apenas os padrões maximais fossem considerados, o analista estaria impossibilitado de analisar informações mais detalhadas. Por exemplo, se a fase maximal fosse aplicada ao exemplo da Figura 7, apenas o padrão "URL2 - URL3 - URL4 - URL5 - URL6" seria retornado por ser maximal. Se por um lado isto reduziria o número de regras, impediria a verificação do suporte específico das várias subseqüências.

Algumas extensões sobre o algoritmo *AprioriAll* são propostas no trabalho de Srikant e Agrawal [SRI95], destacando-se a extensão visando a geração de padrões generalizados. Estes padrões generalizados são obtidos com o uso de taxonomias que definem uma hierarquia entre conceitos e permitem buscar padrões com maior suporte.

Considera-se o exemplo representado pela Figura 8, e o conjunto de dados de entrada descritos na Figura 7. Se uma taxonomia definisse que URL2 e URL4 são especializações do conceito *Produto*, e se esta taxonomia fosse também utilizada como uma entrada ao algoritmo de geração de padrões seqüenciais, outros padrões constituiriam o conjunto de padrões e seriam acrescentados à lista de padrões já mostrados na Figura 7.

Como representado pela Figura 8, os padrões generalizados (representados em negrito) resultantes de acordo com os conceitos definidos pela taxonomia passam a fazer parte do conjunto final de padrões e também apresentam um valor de suporte representativo em relação aos outros padrões descobertos. Por exemplo, o padrão "*Produto* - URL5" representa todas as seqüências de dados que determinam o acesso a URL2 ou URL4 com posterior acesso à URL5. Assim, o valor do suporte é maior quando considerada uma generalização dos itens (URL2 ou URL4) ao invés destes separadamente. Porém, o número de padrões retornados aumenta ainda mais, considerando os diferentes níveis de abstração definidos na taxonomia.

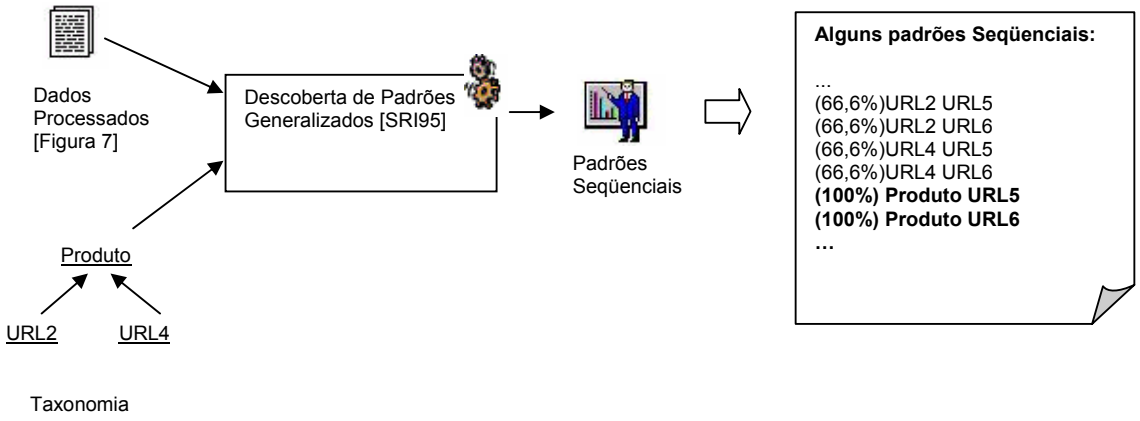


Figura 8: Descoberta de Padrões Seqüenciais com uso de Taxonomia

Ainda, não é possível analisar um padrão generalizado e verificar os padrões existentes em nível detalhado sem consultar a taxonomia definida e filtrar os padrões de acordo com o interesse. Considerando o exemplo citado anteriormente, se o analista tivesse interesse em verificar os padrões que suportam o padrão generalizado "*Produto* - URL6" (ou seja, os padrões "URL2 - URL6" e "URL4 - URL6"), deveria realizar a inspeção manual ou utilizar um filtro para selecionar os padrões de acordo com os conceitos da taxonomia que são especialização do conceito *Produto*.

2.2.3 Análise de Padrões

A análise de padrões, foco principal desta pesquisa, é a última fase da MUW. Ela consiste na identificação de padrões relevantes para o domínio da aplicação dentre os retornados pela fase de Mineração de Dados.

Descobrir padrões interessantes não é uma tarefa fácil uma vez que a definição do que é interessante é muito subjetiva, ou seja, o que é interessante para um usuário pode não ser para outros. De acordo com Fayyad *et al.* [FAY96], um padrão é considerado interessante quando ele é novo, útil, válido e simples. Padrões descobertos são válidos se estes expressam um conhecimento verdadeiro. Eles são considerados novos (pelo menos para o sistema) quando eles contrariam os padrões esperados pelo usuário. Quanto à utilidade, estes padrões devem suportar um conhecimento que possa ser útil ao domínio. Já a simplicidade refere-se à possibilidade de compreensão dos padrões pelos usuários.

O sucesso da fase de Análise de Padrões é dependente das atividades realizadas nas fases anteriores. Freqüentemente nesta fase, o analista se depara com um conjunto elevado de padrões a serem considerados, muitos deles irrelevantes ao domínio. Este é o resultado de aplicação de técnicas tradicionais de Mineração de Dados, como regras associativas [AGR93] e padrões seqüenciais [AGR94a, SRI95], as quais propõem a combinação exaustiva entre os itens que possuem maior freqüência.

Outro aspecto a ser considerado, é que muitos padrões expressam informações redundantes devido às combinações exaustivas de itens (e.g. propriedade *Apriori*). Desta forma, a atividade de identificação dos padrões interessantes ao domínio acaba se tornando uma atividade exaustiva para o analista devido à grande quantidade de padrões a serem analisados e da freqüente redundância entre eles.

Outro problema relacionado à fase de Análise de Padrões, é a dificuldade de interpretação dos padrões seqüenciais gerados. No contexto da MUW, isto se deve ao fato de eles usualmente serem representados por coleções de URLs que nem sempre expressam claramente as intenções dos usuários durante a navegação de um *site Web.*

Considerando o padrão seqüencial apresentado na Figura 6, só o analista com conhecimento do domínio interpreta que 80% das sessões do *log* indicam que usuários acessaram a página de busca por produtos disponíveis no *site*, adicionaram, na cesta de compras, o livro "Hamlet" (item = 12) seguido do livro "Romeu e Julieta" (item = 45) e posteriormente confirmam a compra destes produtos. A Figura 9 representa a interpretação deste padrão seqüencial.

Neste exemplo, a interpretação do padrão torna evidente se o analista possui conhecimento de que o produto de código "12" corresponde ao livro chamado "Hamlet", e o item "45" corresponde ao livro "Romeu e Julieta". Além do mais, para o especialista tem que estar claro que estas URLs estão vinculadas aos eventos subjacentes a uma compra realizada pelo usuário, tais como consultar produtos, incluir produtos na cesta de compras e confirmar compra.

Nota-se por este exemplo que muitas vezes não há uma correspondência entre as URLs e eventos no domínio de aplicações, comprometendo assim, a interpretação dos padrões de navegação. Em outras palavras, não há interesse em padrões formados por URLs, mas sim

em padrões que expressem os eventos de domínio que estimulam a navegação. Stumme *et al.* [STU02] e Berendt *et al.* [BER02a] definem eventos de domínio pelos conteúdos e serviços oferecidos pela aplicação. De acordo com a Figura 9, um evento de serviço oferecidos nas páginas seria "adicionar produtos na cesta de compras" e eventos de conteúdo poderiam ser "Hamlet" e "Romeu e Julieta".

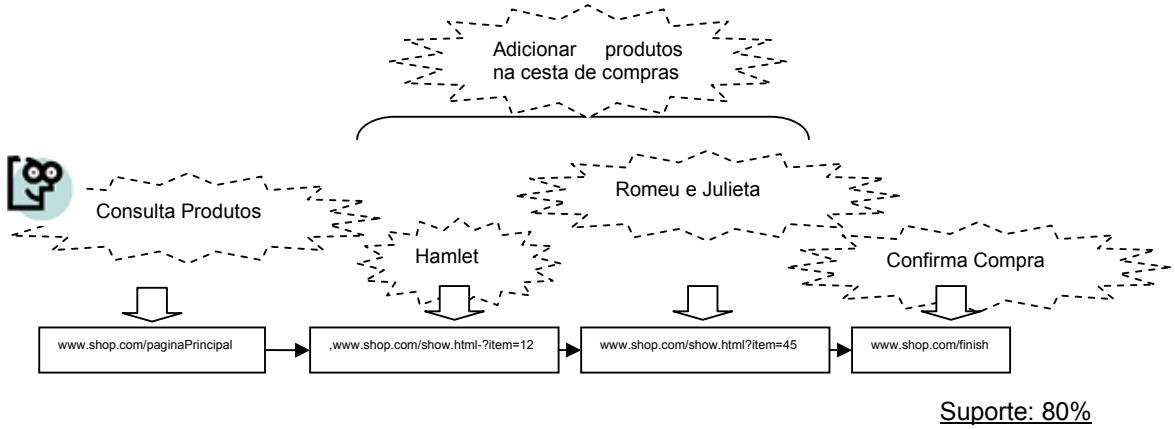


Figura 9: Interpretação de um padrão seqüencial pelo especialista

Diversas são as abordagens propostas para amenizar os diferentes problemas apresentados, as quais podem ser classificadas em abordagens de filtragem, semânticas e de representação. As abordagens de filtragem têm por objetivo a redução do número de padrões com base na definição de filtros, que por sua vez se diferenciam em estatísticos e estruturais. Filtros estatísticos (e.g. [SIL96, AGR93, AGR94a, COO03, POH03]) são utilizados para reduzir o número de padrões retornados pelos algoritmos de mineração com base em medidas objetivas e subjetivas. Filtros estruturais (e.g. [KLE94, SPI98]) determinam as características estruturais que os padrões devem possuir de acordo com o interesse do analista, limitando assim o conjunto de padrões. Já abordagens semânticas (e.g. [BER00, BER02c, SRI95, SRI97, KLE94, BER02a, STU02, BER, DAI02, OBE03]) preocupam-se com a representação do conhecimento expresso pelos padrões, ou seja, propõem-se a associar padrões com os eventos do domínio da aplicação, facilitando assim a atividade de interpretação. Abordagens de representação (e.g. [KLE94, SPI98, BLA03]) exploram técnicas de representação gráfica para facilitar a visualização de padrões. Estas abordagens são apresentadas em detalhes no próximo capítulo.

2.3 Considerações

Este capítulo apresentou as diferentes áreas da Mineração da *Web,* focando-se na Mineração do Uso da *Web.* O processo iterativo e interativo de MUW propõe fases que guiam a descoberta de conhecimento a partir de dados extraídos da *Web.* São elas: Preparação de Dados, Mineração de Dados e Análise de Padrões.

Perante a grande diversidade de fontes de dados e da inconsistência destes, a fase de Preparação de Dados constitui a execução de atividades como filtragem e a limpeza dos dados brutos; identificação de usuários, sessões e transações; complemento do caminho de acesso às páginas *Web.*

Esta fase também pode incluir o enriquecimento semântico dos dados contidos nos *logs* de acordo com o conhecimento extraído do domínio. O objetivo principal é obter padrões mais representativos e facilitar a interpretação dos resultados da fase de Mineração de Dados. Porém, existem limitações apresentadas pelo enriquecimento semântico do *log* realizado na fase de Preparação de Dados. Se o enriquecimento semântico não for o adequado para se atingir os objetivos, é necessário retornar à fase de Preparação de Dados, revisar como este enriquecimento semântico foi definido e executar a fase de Mineração de Dados novamente.

Mesmo que o enriquecimento semântico tenha sido o adequado à interpretação de padrões, existe a limitação devido ao fato de os dados contidos no *log* serem estáticos, ou seja, o analista somente pode explorar a dimensão de interesse representada no *log* semântico. A necessidade da utilização de outras dimensões de interesse implica a re-execução da fase de Preparação e Mineração de Dados.

A fase de Mineração de Dados compreende a aplicação de técnicas de mineração sobre os dados pré-processados resultantes da fase de Preparação de dados. A técnica que enfoca a descoberta de padrões seqüenciais foi a selecionada para esta pesquisa por revelar padrões de acesso às páginas *Web* obedecendo a uma determinada seqüência temporal. O algoritmo utilizado para a descoberta de padrões seqüenciais é o *AprioriAll* por ser uma técnica tradicional aplicada no contexto de descoberta de conhecimento. Porém, outros algoritmos de geração de padrões seqüenciais e regras associativas poderiam considerados para a proposta deste trabalho.

A fase de Análise de Padrões, foco desta pesquisa compreende a identificação de padrões relevantes ao domínio. Porém, é uma fase trabalhosa para o analista devido ao elevado número de padrões que geralmente resultam da aplicação de técnicas de Mineração de Dados durante a fase de Mineração de Dados, em particular regras associativas e padrões seqüenciais. Outro problema enfrentado no contexto da MUW é a dificuldade de compreensão dos padrões, geralmente formados por URLs, uma vez que as URLs nem sempre expressam claramente os serviços e conteúdos que impulsionam os usuários durante a navegação no *site Web.*

3 TRABALHOS RELACIONADOS

Este capítulo descreve as principais abordagens propostas a auxiliar a interpretação e recuperação de padrões relevantes. Para finalizar, é descrito um comparativo entre as abordagens.

A fase de Análise de Padrões compreende a identificação de padrões relevantes ao domínio, porém descobrir padrões interessantes não é uma tarefa trivial uma vez que a própria definição de *interessante* é subjetiva, ou seja, o que é interessante para um usuário pode não ser para outros.

Como descrito no Capítulo 2, a fase de Análise de Padrões torna-se extremamente extenuante quando o analista se depara com um grande volume de padrões, muitos deles irrelevantes e redundantes, e de difícil interpretação. Neste contexto, pesquisas vêm sendo realizadas visando facilitar a recuperação e interpretação de padrões. Neste trabalho, entende-se por interpretação de padrões as atividades executadas para o entendimento das informações expressadas por um padrão. Já recuperação de padrões refere-se às atividades realizadas para encontrar padrões com determinadas características em meio a um conjunto destes.

Para facilitar o entendimento dos diferentes trabalhos relacionados, é proposta uma classificação que diferencia as abordagens de acordo com o objetivo principal: de reduzir o escopo da pesquisa por padrões relevantes; de facilitar o entendimento dos eventos de domínio suportado por eles; e finalmente de representação gráfica dos padrões descobertos. Com base nestes objetivos, os trabalhos relacionados foram classificados por esta pesquisa como abordagens de filtragem; abordagens semânticas e abordagens de representação gráfica. Estas abordagens são detalhadas no restante deste capítulo.

3.1 Abordagens de Filtragem

Abordagens de filtragem referem-se à definição de filtros que recuperam padrões potencialmente relevantes de acordo com as características especificadas. Filtros podem ser utilizados após a fase de Mineração de Dados (Figura 10-a), restringindo o volume de padrões na etapa de Análise de Padrões. Neste caso, os padrões resultantes da fase de Mineração de Dados são filtrados, selecionando apenas os padrões potencialmente relevantes ao domínio de

acordo com o interesse do analista definido através dos filtros. Outra alternativa, mostrada pela Figura 10-b, é aplicar filtros acoplados a técnicas de Mineração de Dados, buscando gerar apenas regras potencialmente importantes para o domínio, antecipando assim a aplicação de alguns critérios de validação que seriam utilizados na fase de Análise de Padrões.

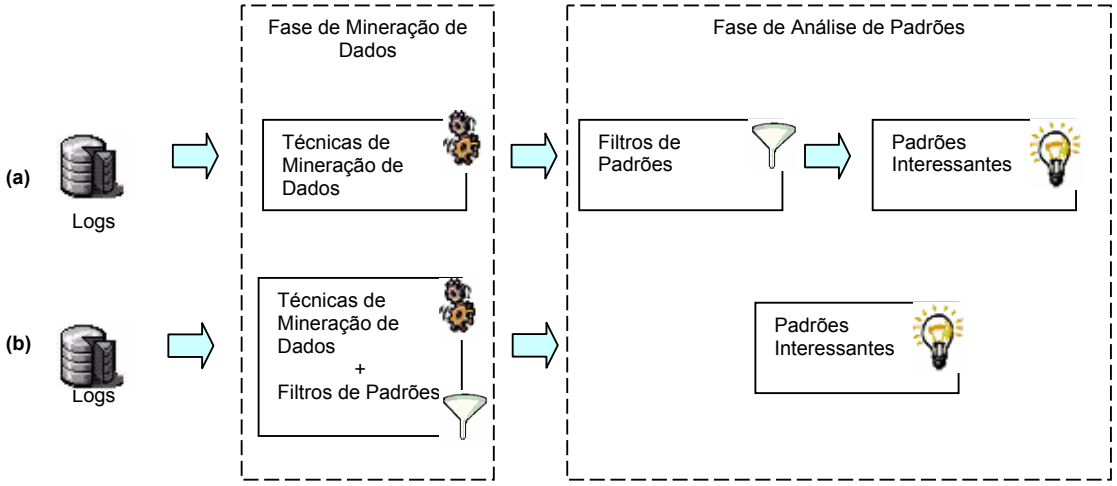


Figura 10: Filtros e as fases do processo de KDD

Os filtros de padrões são diferenciados em filtros estatísticos e filtros estruturais, detalhados nas seções subseqüentes.

3.1.1 Filtros Estatísticos

Filtros estatísticos têm como objetivo a recuperação de padrões através do uso de medidas estatísticas. Silberschatz *et al.* [SIL96] diferenciam entre medida de interesse objetiva e medida de interesse subjetiva. A medida objetiva depende somente da estrutura das regras e dos dados usados no processo de mineração. Um exemplo deste tipo de medida é o suporte utilizado para a geração de padrões seqüênciais e regras de associação (e.g. [AGR93, AGR94a]) durante a fase de Mineração de Dados. Neste caso, o objetivo é definir um valor mínimo de tal forma que o algoritmo de mineração gere apenas os padrões que possuam um valor acima do especificado pela medida objetiva.

Filtros estatísticos demonstram eficácia na redução do número de padrões retornados, mas a efetividade em se encontrar padrões relevantes pode estar comprometida [KLE94, HIP02]. Por exemplo, ao definir um valor alto para a medida do suporte, apenas os padrões com suporte superior ao definido serão recuperados. Existe um risco nesta operação pois

muitos padrões com suporte inferior e potencialmente interessante serão desconsiderados. Outro fato a considerar é que nem sempre padrões que possuem suporte alto são interessantes por representar um conhecimento prévio e comum. Por exemplo o padrão seqüencial "*home-page - login-page*" deve possuir um suporte alto, mas expressa um conhecimento trivial refletindo que usuários que acessam a página principal do *site* posteriormente realizam a autenticação do usuário.

Além da dificuldade na definição de um valor adequado para as medidas objetivas, existe outro problema quando estas são associadas à fase de Mineração de Dados. Elas implicam a re-execução desta fase tantas vezes quantas for necessário, parametrizando os algoritmos com diferentes valores, até que padrões relevantes sejam identificados (e.g. [AGR93, AGR94a]). Em função disto, Hipp *et al.* [HIP02] propõem, no contexto das regras associativas, a utilização de valores irrisórios para o suporte mínimo (*minsup*) visando gerar todas as regras possíveis durante a fase de Mineração de Dados. Um mecanismo de filtragem permite então selecionar aqueles padrões dentre os limiares interessantes.

Filtros estatísticos quando aplicados na fase de Análise de Padrões, são utilizados apenas para restringir o foco de busca pelos padrões relevantes considerando o conjunto total de padrões resultantes da aplicação das técnicas de mineração (e.g. [HIP02]).

Como visto, medidas objetivas não são suficientes para determinar se um padrão é interessante. Para isso é proposta a medida de interesse subjetiva, que depende não somente da estrutura das regras e dos dados usados no processo de mineração, mas também de um conhecimento prévio especificado pelo usuário que determina se uma regra é interessante ou não.

Silberschatz *et al.* [SIL96] discutem como estas medidas podem ser utilizadas na descoberta de padrões relevantes. A proposta é descobrir padrões inesperados com base num conjunto de crenças (*beliefs*) previamente especificadas. Um padrão é inesperado quando ele discorda dos padrões definidos como crença. Porém, para definir se um padrão é inesperado, é necessário primeiramente especificar as crenças. Os autores distinguem dois tipos de crenças: as invariantes (*hard beliefs*) e as variantes (*soft beliefs)*. Ambos possuem uma medida de confiança associada que determina o grau de sua veracidade em relação ao domínio, valor este especificado a partir de cálculos de probabilidade. A diferença é que as crenças invariantes, uma vez definidas, não têm sua medida de confiança alterada, diferentemente das crenças

variantes onde a medida de confiança é atualizada sempre que uma nova evidência extraída do domínio da aplicação contradisser o padrão representado pela crença variante. Assim, uma vez definido o que é interessante através das crenças, é possível identificar se os padrões descobertos são interessantes, ou seja, inesperados neste contexto.

Cooley [COO03] propõe a utilização de filtros objetivos e subjetivos baseado nas medidas apresentadas anteriormente para auxiliar a fase de Análise de Padrões. Cooley *et al.* [COO99a] também propõem uma forma alternativa de derivar crenças a partir do conteúdo e estrutura do *site*.

Pohle [POH03] utiliza as medidas subjetivas para determinar quanto um padrão seqüencial é semelhante aos armazenados numa base de conhecimento. Como complemento às medidas defendidas pela abordagem de Silberschatz *et al.* [SIL96], ele propõe o uso de conjuntos *fuzzy* para classificar o grau de semelhança do padrão seqüencial descoberto com as crenças, ao invés da definição de um limiar fixo para as medidas subjetivas. Segundo Phole, a vantagem da utilização dos conjuntos *fuzzy* é que eles se assemelham à linguagem utilizada pelos analistas, como por exemplo os conjuntos definidos por "baixo" e "alto", e garantem uma melhor classificação dos níveis de interesse em relação aos limiares pré-definidos. Por exemplo, não é intuitiva a razão pela qual um padrão que possui uma medida subjetiva de 20% é menos significativo que um padrão que possua 19,99%. A utilização de conjuntos *fuzzy* trata estas incertezas.

No geral, as medidas subjetivas são úteis para determinar o quanto um padrão é inesperado para o domínio. Porém a desvantagem é que a efetividade destas medidas está diretamente relacionada com a habilidade de expressar o conhecimento do domínio na forma de crenças, suas probabilidades de ocorrência e nos métodos de comparação entre os padrões descobertos e as crenças armazenadas em bases de conhecimento.

3.1.1.1 *Considerações*

Filtros estatísticos objetivam reduzir o número de padrões baseando-se em medidas objetivas e subjetivas. Estes filtros são aplicados tanto na fase de Mineração de Dados quanto na Análise de Padrões.

Na fase de Mineração de Dados, a medida estatística é utilizada para guiar a descoberta de padrões. A desvantagem é que, além do risco de limitar os padrões descobertos,

esta fase deve ser re-executada com novos valores para as medidas consideradas sempre que os padrões não atenderem aos objetivos propostos. Este problema já não ocorre quando filtros estatísticos são utilizados na fase de Análise, reduzindo apenas o escopo da busca pelos padrões.

Medidas objetivas nem sempre são suficientes para determinar se um padrão é relevante ou não para o domínio da aplicação, afinal um padrão com suporte alto não é necessariamente relevante, assim como um padrão com o suporte baixo não é necessariamente irrelevante.

O julgamento quanto a um padrão ser relevante ou não depende muito das informações expressas por ele. Neste contexto, medidas subjetivas podem ser aplicadas baseadas no conhecimento prévio especificado na forma de crenças. Elas são úteis para determinar o quanto um padrão é inesperado para o domínio. Porém, sua efetividade está diretamente relacionada com a habilidade de expressar o conhecimento do domínio na forma de crenças, suas probabilidades de ocorrência e nos métodos de comparação entre os padrões descobertos e as crenças previamente armazenadas em bases de conhecimento.

3.1.2 Filtros Estruturais

Filtros estruturais têm como objetivo a recuperação de padrões que estão de acordo com as restrições estruturais definidas. Estas restrições podem definir o conteúdo do padrão, assim como a disposição dos itens que o formam.

Um exemplo da aplicação de filtros estruturais na fase de Análise de Padrões é o trabalho de Klemettinen *et al.* [KLE94] que propõem a utilização de filtros (restritivos e inclusivos) visando recuperar regras associativas que obedecem a restrições definidas pelo analista em filtros. Taxonomias, que definem uma hierarquia de conceitos, podem ser utilizadas para simplificar a especificação dos filtros.

Os autores apresentam um cenário de uso dos filtros considerando regras associativas descobertas no domínio de um curso de Ciência da Computação. A taxonomia classifica as disciplinas do curso em três classes: Básica, Intermediária e Avançada, como representado na Figura 11. Todas as classes são, por sua vez, especializações da classe “Qualquer disciplina”.

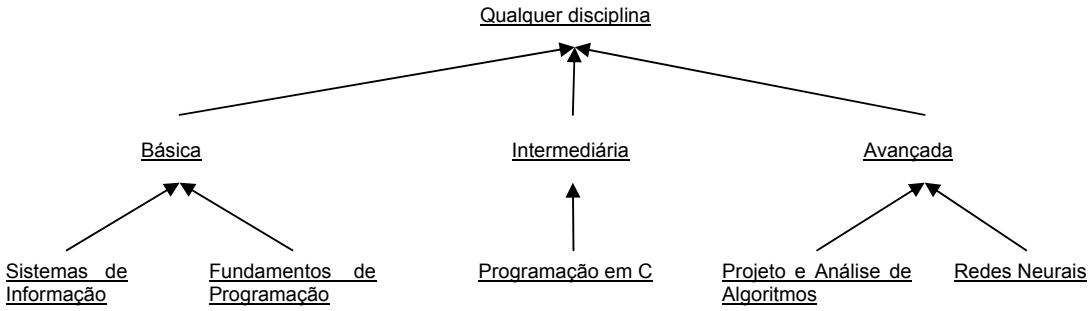


Figura 11: Taxonomia das disciplinas de um curso de Ciência da Computação

Um exemplo de filtro poderia ser:

```
Intermediária, Qualquer disciplina* → Projeto e Análise de Algoritmos
```

visando selecionar todas as regras que possuem disciplinas relacionadas com a classe Intermediária como primeiro item, e os itens seguintes correspondem a qualquer disciplina. Ainda, de acordo com este filtro, o conseqüente das regras associativas obrigatoriamente deve ser a disciplina Projeto e Análise de Algoritmos. Dois exemplos de regras associativas recuperadas de acordo com este filtro definido, seriam:

```
Programação em C → Projeto e Análise de Algoritmos
Programação em C, Redes Neurais → Projeto e Análise de Algoritmos
```

Um exemplo da aplicação de filtros estruturais na fase de Mineração de Dados é o trabalho de Spiliopoulou *et al.* [SPI98], que propõe a especificação de filtros que alimentam um algoritmo de geração de padrões seqüenciais. Esta proposta é incorporada no ambiente WUM (*Web Utilization Miner*), um sistema que descobre padrões seqüenciais de navegação satisfazendo os critérios de filtragem definidos pelo usuário através da linguagem de MINT. Estes critérios referem-se à estrutura, conteúdo e estatísticas dos padrões a serem descobertos. A seguir apresenta-se um exemplo de filtro definido pela linguagem MINT onde o analista declara interesse pelos caminhos de navegação que iniciam na primeira ocorrência da página A.html ou C.html e que convergem para página B.html com no mínimo 5% dos acessos à página inicial.

```
SELECT t
FROM NODE AS x y, TEMPLATE x * y as t
WHERE ((x.URL = "A.html")  or (x.URL = "C.html"))
AND y.URL = "B.html"
(b.support / a.support) >= 0.05
```

A Figura 12 ilustra alguns padrões de navegação que poderiam ser retornados, dois neste caso. O ambiente WUM ainda propõe uma representação gráfica para facilitar a visualização destes padrões retornados. Na representação gráfica, cada nodo representa uma página *Web* acessada. O primeiro número entre colchetes representa a ocorrência da página, e o segundo, o número de acessos.

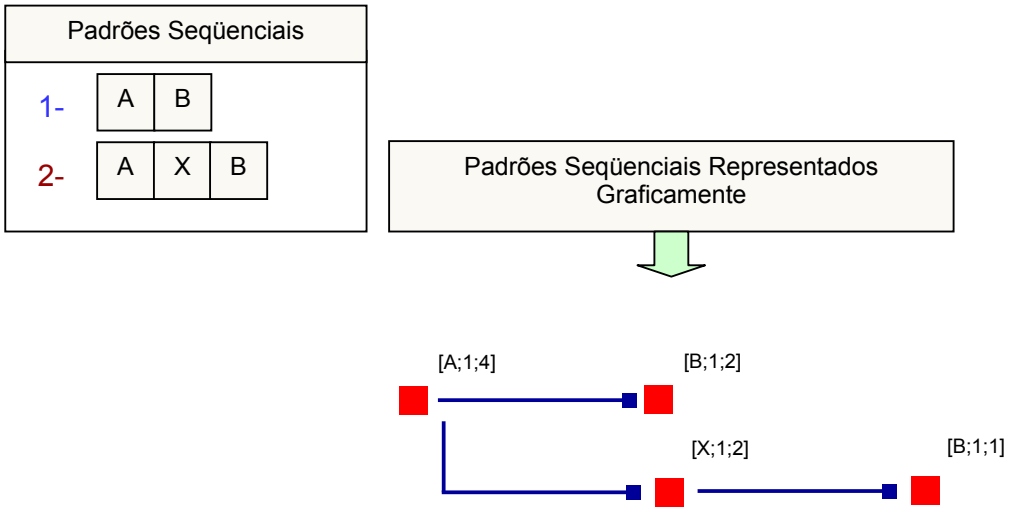


Figura 12: Padrões de navegação retornados pela ferramenta WUM

Nota-se que os padrões representados na Figura 12 respeitam as restrições definidas pelo filtro, ou seja, todos iniciam pela página A, finalizam na página B e com no mínimo 5% dos acessos à página inicial do caminho de navegação. Compreende-se que das quatro sessões que iniciaram o caminho de navegação pela página A, três concluíram com a página B, sendo que uma destas sessões passa pelo acesso à página X antes de chegar a B.

Como descrito, filtros estruturais visam reduzir o escopo na busca por padrões relevantes. Estes podem ser aplicadas tanto na fase de Descoberta quanto de Análise de Padrões. Porém, a integração de filtros à fase de Mineração de Dados (e.g. [SPI98]) implica a re-execução desta fase até que padrões relevantes sejam identificados. Diferentemente, se associados à fase de Análise de Padrões, os filtros são utilizados apenas para restringir o foco

de busca pelos padrões relevantes considerando o conjunto total de padrões resultantes da aplicação das técnicas de mineração.

A desvantagem destas abordagens é que o analista deve possuir: a) domínio sobre a sintaxe da linguagem de especificação dos filtros; b) clareza quanto às características dos padrões que deseja recuperar e c) conhecimento do domínio para especificar estas características utilizando a sintaxe de uma linguagem de filtragem.

Numa proposta semelhante às apresentadas, Shah *et al.* [SAH99], no contexto de regras associativas, visam restringir o escopo na busca por regras relevantes através da formação e eliminação de grupos ou famílias de regras. A eliminação de um grupo de regras é determinada pela classificação de regras realizadas pelo usuário. Fazem parte de uma mesma família de regras, todas as regras que possuem associação entre uma determinada hipótese e um conseqüente específico.

Quando o usuário interage com as regras, ele deve classificar as regras de acordo com a sua veracidade e relevância. Especificar quando uma regra é verdadeira é importante para a construção da base de conhecimento. Já a sua relevância implica a exclusão (regra não relevante) ou permanência (regra relevante) das regras que pertencem à mesma família. Por exemplo, muitas regras de senso comum podem ser excluídas por serem irrelevantes, como a regra "se grávida então mulher" ($\langle$grávida$\rangle \rightarrow \langle$mulher$\rangle$). Assim, todas as regras que pertencem à mesma família da regra classificada, serão automaticamente excluídas.

Estas duas dimensões definem quatro possibilidades de classificação: regra verdadeira e não interessante (RVNI), regra falsa e não interessante (RFNI), regra falsa e interessante (RFI) e regra verdadeira e interessante (RVI). Os autores admitem a dificuldade em classificar regras nestes dois últimos conjuntos.

Os benefícios desta abordagem referem-se à redução do tempo consumido pela identificação exaustiva do que é não interessante para o usuário; ao baixo nível de interação com o usuário, o qual classifica algumas regras representativas; ao processo de classificação simples e a eliminação de uma grande quantidade de regras durante uma única interação. Porém isto não garante que todas as retras que restem sejam realmente interessantes ao domínio.

3.1.2.1 *Considerações*

Filtros estruturais objetivam reduzir o número de padrões com base em variadas caracteristicas dos padrões, tais como estrutura, conteúdo e mesmo estatística. Da mesma forma que os filtros estatísticos, eles são aplicados tanto na fase de Mineração de Dados quanto na Análise de Padrões.

A desvantagem de aplicar filtros estruturais na fase de Mineração de Dados, além do risco de limitar o número de padrões descobertos, está na re-execução desta fase com diferentes propriedades até que padrões que satisfaçam os objetivos sejam recuperados. Este problema já não ocorre quando filtros estruturais são utilizados na fase de Análise, reduzindo apenas o escopo da busca pelos padrões.

Filtros estruturais são úteis pois permitem a definição das propriedades que os padrões devem possuir de acordo com o interesse do analista. A desvantagem é que o analista deve possuir: a) domínio sobre a sintaxe da linguagem de especificação dos filtros; b) clareza quanto às características dos padrões que deseja recuperar e c) conhecimento do domínio para especificar as características utilizando a sintaxe de uma linguagem de filtragem.

Visando amenizar estas desvantagens, alguns autores propõem apenas a classificação de regras com o objetivo de desconsiderar a família de regras daquelas que são irrelevantes e falsas para o domínio. A vantagem é que com poucas interações uma grande quantidade de regras da mesma família pode ser excluída e o analista não precisa se preocupar em definir filtros utilizando uma sintaxe. A desvantagem é que nem sempre é trivial classificar uma regra de acordo com sua veracidade e relevância e ao final do processo não existe garantia de que as regras que restaram são todas relevantes ao domínio.

3.2 Abordagens Semânticas

Para melhor entender os resultados da mineração, o analista deve ter conhecimento sobre a semântica dos acessos às páginas em termos de eventos e conteúdo disponíveis no domínio. Neste contexto, surgem as abordagens semânticas visando fornecer suporte a interpretação de padrões através da representação do conhecimento do domínio.

As abordagens semânticas foram distinguidas quanto à representação do conhecimento do domínio utilizada na integração com o processo de MUW, a saber, taxonomias e ontologias de domínio.

3.2.1 Taxonomia

Taxonomias correspondem a uma forma primitiva de representação de conhecimento, as quais definem uma hierarquia conceitual formada por classes e sub-classes de objetos, relacionadas através de relacionamentos de generalização/especialização (*é-um*). Os trabalhos pesquisados associam o uso de taxonomias a todas as fases do processo de MUW: Preparação de Dados (e.g. [BER00]), Mineração de Dados (e.g. [SRI95, SRI97]) e Análise de Padrões (e.g. [KLE94]).

Relativamente à aplicação de taxonomias na fase de Preparação de Dados, Berendt *et al.* [BER00] propõem a classificação de URLs em hierarquias conceituais que refletem as estratégias de navegação através das múltiplas dimensões do espaço de características das páginas *Web*. Esta classificação é feita baseando-se nos diferentes serviços oferecidos pelas páginas resultantes de consultas geradas dinamicamente num determinado *site Web*. Cabe ao especialista especificar os conceitos que descrevem a aplicação nas diferentes dimensões do espaço de características. A Figura 13, adaptada de Berendt *et al.* [BER00], exemplifica esta abordagem através de uma hierarquia definida para um *site* de consulta de escolas. Nela, são observadas as diferentes dimensões considerando os tipos de estratégias de pesquisas disponíveis aos usuários deste *site*.

Para análise de padrões, Berendt *et al.* utilizam o ambiente WUM [SPI98] integrado com as hierarquias conceituais definidas, permitindo assim analisar o perfil de navegação dos usuários através das diferentes dimensões de interesse. Trabalhos subseqüentes (e.g. [POH02, SPI02]) exploram metodologias para a definição de hierarquias conceituais para posterior utilização da ferramenta WUM. Em um outro trabalho relacionado, Berendt [BER02c] propõe um módulo que estende as funcionalidades propostas pela ferramenta WUM, visando a atualização das hierarquias conceituais definidas a partir da descoberta de padrões que determinam estratégias de navegação diferentes daquelas especificadas na hierarquia.

O uso de taxonomias na fase de Mineração de Dados é exemplificado pelos trabalhos de Srikant e Agrawal [SRI95, SRI97], os quais propõem a extensão de algoritmos de geração

de padrões seqüenciais e regras associativas respectivamente. Estes algoritmos estendidos visam a geração de padrões generalizados, os quais conseqüentemente apresentam maior representatividade, relacionando-se diretamente com a medida de suporte. Como as taxonomias são integradas aos algoritmos de mineração, a necessidade de preparação dos dados é reduzida. Um exemplo desta abordagem foi representado pela Figura 8 no Capítulo 2. Como já mencionado, as limitações desta abordagem incluem o aumento do número de padrões retornados devido à geração dos padrões incluindo os conceitos generalizados definidos pela taxonomia e à dificuldade de relacionar os padrões detalhados aos respectivos padrões generalizados.

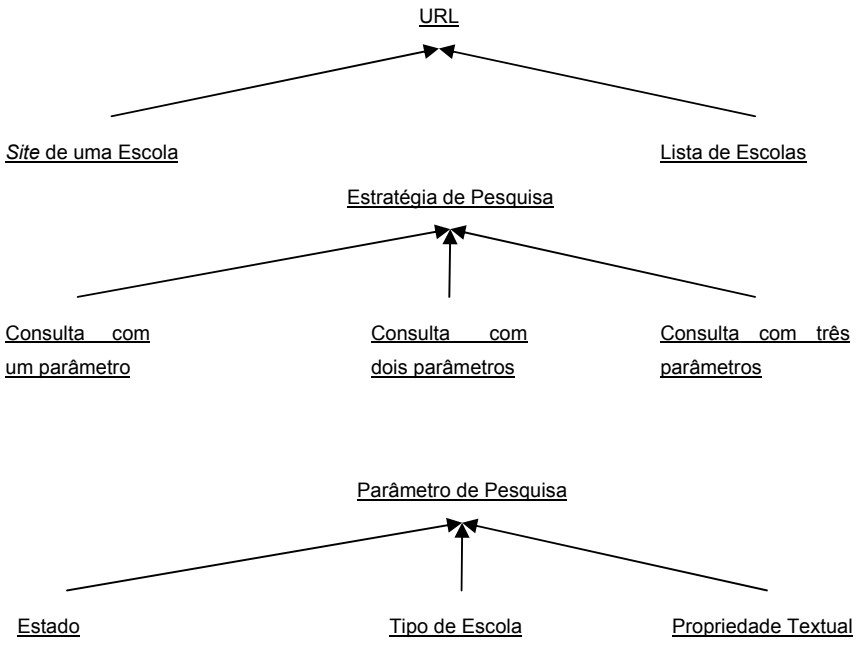


Figura 13: Dimensões da hierarquia Conceitual do *site SchulWeb*

OLAP (*On-line Analytical Processing*) [HAN97] é outro exemplo de aplicação de taxonomias à fase de Mineração de Dados. Em um modelo de dados OLAP é possível visualizar os dados de uma maneira analítica e multidimensional, realizando operações de sumarização, comparação, taxas multidimensionais e variações percentuais.

Em um modelo de dados OLAP, a informação é conceitualmente organizada em cubos que armazenam valores quantitativos (fatos) organizados em dimensões que formam a estrutura de um cubo. Uma dimensão pode ser qualquer visão do negócio de interesse para

análise dos dados, como por exemplo produto, departamento ou tempo. Dentro de cada dimensão de um modelo OLAP, os dados podem ser organizados em uma taxonomia que define diferentes níveis de detalhe (hierarquias, como são comumente denominadas). Por exemplo, dentro da dimensão tempo, pode existir uma taxonomia representando os níveis ano, mês, e dia. Da mesma forma, a dimensão local poderá ter os níveis país, região, estado e cidade. Assim, um usuário visualizando dados em um modelo OLAP poderá sumarizar (*roll-up*) ou detalhar (*drill-down*) a informação explorando diferentes níveis hierárquicos das dimensões.

OLAP é um dos recursos analíticos que podem estar associados a um *data warehouse* (DW), como exemplificado no ambiente DBMiner [HAN96]. Neste os dados são representados através de um DW multidimensional, onde cada dimensão representa uma taxonomia, os quais são analisados através de OLAP (geração e visualização de sumários) ou OLAM (*On-line Analytical Mining*).

Taxonomias também podem ser utilizadas para facilitar a fase de Análise de Padrões, como abordado por Klemettinen *et al.* [KLE94], como já descrito na Seção 3.1.2. Os autores utilizam taxonomias para flexibilizar a definição de filtros na fase de Análise de Padrões, selecionando padrões de acordo com as classes especificadas na taxonomia.

3.2.1.1 *Considerações*

Taxonomias limitam-se a uma hierarquia de conceitos conectados por relações de generalização/especialização. Para muitas aplicações este tipo de relação é suficiente para obtenção de resultados eficazes.

Vários trabalhos aplicam taxonomias às mais distintas fases do processo de MUW. A utilização de taxonomias na fase de Preparação de Dados tem como objetivo o enriquecimento semântico do *log* para facilitar a interpretação dos padrões na fase de Análise de Padrões. No entanto, somente na fase de análise será possível avaliar se o enriquecimento semântico foi adequado. Caso contrário, é necessário retornar à fase de Preparação de Dados e rever o enriquecimento semântico do *log*.

Quando taxonomias são utilizadas na fase de Mineração de Dados, permitem a descoberta de padrões generalizados e mais representativos ao domínio, ou seja, com suporte mais elevado [SRI95, SRI97]. Porém, a desvantagem deve-se ao aumento do número de

padrões retornados, considerando os diferentes níveis de abstração definidos pela taxonomia, e à redundância semântica entre eles, a qual dificulta ainda mais a atividade de análise.

Outra desvantagem é que não existe conexão entre os padrões generalizados e os padrões que os detalham. Para este tipo de interpretação, o analista tem que utilizar outros recursos para recuperar os padrões detalhados. Em OLAP, é possível explorar um padrão considerando as diferentes dimensões definidas pelos cubos, facilitando a interpretação e evitando o retorno à fase de Preparação de Dados.

Finalmente, taxonomias são aplicadas à fase de Análise de Padrões para facilitar a definição de filtros utilizados na recuperação de padrões. Limitações desta abordagem referem-se ao domínio de uma sintaxe para a definição do filtro.

3.2.2 Ontologia de domínio

O uso de taxonomias demonstra ser eficiente para alguns propósitos, porém limita a representação do conhecimento do domínio a relações do tipo *é-um*. Com o desenvolvimento da *Web* Semântica [LEE01], há um grande incentivo para a formalização e exploração do conhecimento embutido nas páginas *Web* através de ontologias, definidas como "uma especificação formal e explícita de uma conceitualização compartilhada" [GRU93]. Formal pois é processada por computador. Explicita pois é composta por um conjunto de conceitos, propriedades, relações, funções, axiomas e restrições. Conceitualização pois é um modelo abstrato e simples dos objetos da realidade. Finalmente compartilhada, pois as informações que ela representa são definidas e aceitas por um grupo de especialistas.

Segundo Berners-Lee *et al.* [LEE01], a *Web* Semântica é uma extensão da *Web* atual, que se preocupa em disponibilizar informações também para as máquinas (*Machine-understandable Information*). A *Web* Semântica propõe à rede global a construção de uma estrutura que permita a evolução de uma rede de documentos para uma rede de dados na qual toda informação tenha um significado bem definido, podendo ser interpretada e processada por humanos ou computadores (agentes computacionais).

Neste contexto, novas perspectivas de pesquisas (e.g. [BER02a, STU02]) buscam explorar quais os possíveis benefícios resultantes da combinação da área de Mineração de Dados e da *Web* Semântica, constituindo assim, uma nova área de pesquisa denominada Mineração da *Web* Semântica (*Semantic Web Mining*). Os objetivos desta frente de pesquisa

são complementares: investigar como a definição de ontologias pode auxiliar na geração de resultados mais interessantes no processo de mineração da *Web,* e por outro lado, como a Mineração da *Web* pode auxiliar na criação e refinamento da semântica para *Web.* A Figura 14, extraída de um trabalho de Berendt *et al*. [BER02a], ilustra a relação entre estas áreas.

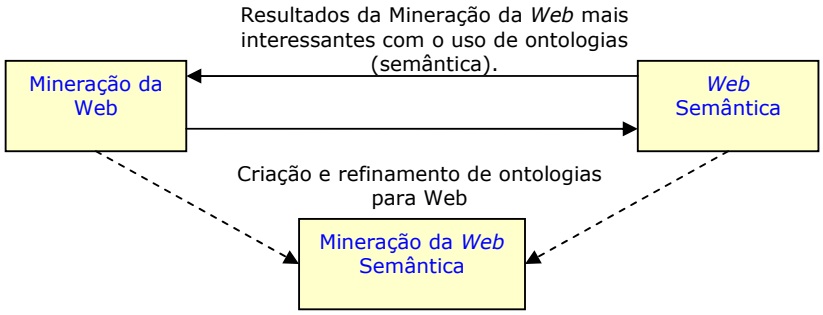


Figura 14: Relação entre *Web* Semântica e Mineração da *Web*

Stumme *et al.* [STU02] e Berendt *et al.* [BER02a] discutem a possibilidade de agregar semântica aos registros de *log Web,* atualmente pobres em informação. Eles classificam os eventos do domínio da aplicação em eventos atômicos e complexos. Eventos atômicos distinguem-se entre serviço ou conteúdo de acordo com os propósitos das páginas *Web.* Evento atômico de conteúdo refere-se ao assunto (e.g. restaurantes, hotéis) descrito por uma página *Web.* Evento atômico de serviço refere-se ao serviço disponibilizado em uma página *Web* (e.g. comprar, pesquisar, comunicar). Já um evento complexo é constituído de um conjunto de eventos atômicos representando um caminho de navegação que pode, por exemplo, corresponder a uma estratégia de resolução de problema conscientemente exercida pelo usuário (e.g. refinamento de termos num processo de busca), uma seqüência de atividades padronizadas relativas ao tipo de *site* (e.g. operações padrões em *site* de comércio eletrônico) ou a descrição de comportamentos identificados pelos especialistas na análise dos padrões descobertos.

Partindo para uma abordagem prática, Dai *et al.* [DAI02] propõem um conjunto de atividades para a fase de Preparação de Dados. O objetivo principal é caracterizar perfis de usuários a partir de um conjunto de objetos semânticos provenientes da Ontologia de Domínio que descreve o conteúdo da aplicação. Neste contexto, os autores classificam a representação de objetos do domínio em dois níveis:

- *Nível físico*, representado pelas URLs que correspondem às páginas *Web.*
- *Nível lógico*, representado por uma Ontologia de Domínio que descreve o conteúdo disponibilizado pelo *site Web.*

Nesta abordagem, os autores consideram as URLs como um conjunto de objetos que referenciam entidades (conceitos e atributos) definidas na Ontologia de Domínio. Assim, cada sessão de usuário é representada como um conjunto de objetos, isto é, instâncias de conceitos e atributos da ontologia, constituindo assim, uma base de conhecimento. Posteriormente, estas sessões, definidas por estes objetos são agrupadas de acordo com características similares aplicando a técnica de agrupamento (*clustering*) [HAN96]. Desta forma, cada grupo gerado é formado por diversas sessões similares. O uso de ontologias permite a geração de agrupamentos que refletem a semântica das requisições às páginas, ou seja, o conteúdo de interesse dos usuários que navegam no *site*. Desta forma a interpretação dos agrupamentos é facilitada, uma vez que URLs são representadas por um conjunto de instâncias de conceitos e atributos da ontologia.

A abordagem de Oberle *et al.* [OBE03] é baseada nas abordagens de Berendt *et al.* [BER00] e Dai *et al.* [DAI02] e explora o conhecimento estruturado e representado por uma ontologia. Esta ontologia refere-se à camada *Ontology Vocabulary* da *Web* Semântica, expressa em RDFS (*Resource Description Framework Schema*) [BRI02]. As URLs presentes nos *logs* são mapeadas para um conjunto de entidades da Ontologia de Domínio e representadas por um vetor de características com pesos (e.g. pesos binários podem determinar se uma característica está ou não presente na página acessada). O assim chamado *log* semântico, contém para cada requisição de página, o horário de acesso, a URL e o vetor de características extraído a partir da Ontologia de Domínio. A identificação do usuário também pode estar presente. A Figura 15 representa um registro extraído do *log* semântico, onde "Pessoa" e "Publicação" representam conceitos da ontologia que descrevem o conteúdo do *site*.

Usuário	**Horário**	**URL**	**Vetor de características**	
			Pessoa	**Publicação**
4711	12:45		1	1

Figura 15: *Log* Semântico

Visando analisar os conteúdos de interesse dos usuários que acessaram o *site Web,* os autores aplicaram a técnica de agrupamento sobre o *log* semântico. O resultado foi um conjunto de grupos que definem os conceitos de interesse. Cada grupo representa componentes similares referentes a conceitos e relações da ontologia, e não necessariamente das páginas acessadas. Desta forma, o conhecimento extraído a partir destes grupos de interesse pode ser utilizado para melhorar a estrutura navegacional do *site*.

3.2.2.1 *Considerações*

Trabalhos no contexto da MUW buscam explorar ontologias de domínio para o enriquecimento semântico dos dados armazenados no *log* (*log* semântico), e conseqüentemente obter resultados mais compreensíveis pelos analistas. Porém, ainda são poucos os trabalhos desenvolvidos integrando ontologias ao processo de MUW e os que existem, focam-se na fase de Preparação de Dados.

A desvantagem do enriquecimento semântico do *log* na fase de Preparação de Dados acontece devido à falta de flexibilidade no suporte à fase de Análise de Padrões, ou seja, se o enriquecimento semântico não foi adequado para suportar a interpretação de padrões, é necessário retornar à fase de Preparação de Dados, revisar a semântica representada e minerar novamente o *log* para obter novos padrões.

3.3 Abordagem de Representação

Abordagens de representação enfocam a visualização gráfica de padrões visando auxiliar tanto na interpretação quanto na recuperação de padrões.

Um exemplo da aplicação da abordagem de representação para a interpretação de padrões é o trabalho de Spiliopoulou *et al.* [SPI98], o qual propõe a visualização dos padrões na forma de uma árvore, complementando o propósito principal que é recuperar padrões através da definição de filtros estruturais pela linguagem MINT.

A árvore é composta por um conjunto de nodos. O início da árvore é caracterizado por um nodo fictício, que representa a chegada de todo visitante ao *site*. Os demais nodos na árvore correspondem à ocorrência de uma página nos percursos feitos pelos usuários, com o respectivo número de acessos. A Figura 16 adaptada de Spiliopoulou *et al.* [SPI98], ilustra um exemplo dessa árvore representando que dos 34 usuários que visitaram o *site*, 21 iniciaram

pela página “a” e 13 tomaram a página “b” como ponto de partida. Dos que iniciaram pela página “a”, 11 dirigiram-se diretamente à página “b”, e 10 usuários atingiram esta página após acesso à página “d”.

Nota-se neste exemplo, que os padrões são interpretados mais facilmente através da representação gráfica.

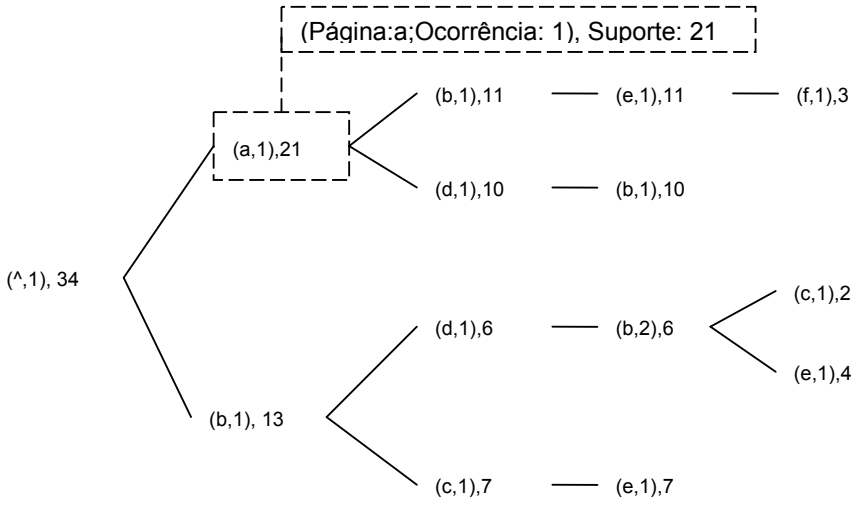


Figura 16: Visualização do Padrão de navegação pela ferramenta WUM

Considerando representação gráfica com o propósito de facilitar a recuperação de padrões, pode-se citar o trabalho de Klemettinen *et al.* [KLE94], o qual propõe a representação gráfica de todas regras associativas na forma de um grafo. Cada atributo é composto por um nodo no grafo e a relação é representada pelos arcos (Figura 17). A espessura dos arcos é diretamente proporcional ao suporte e à confiança da regra correspondente, assim o analista pode identificar aquelas regras com maior relevância estatística. A desvantagem desta proposta é que, dependendo do volume de regras representado, o grafo pode se tornar denso e difícil de ser compreendido. Recursos semelhantes estão presentes em muitas ferramentas de descoberta de conhecimento, tais como DBMiner [HAN96] e Clementine [ISL98].

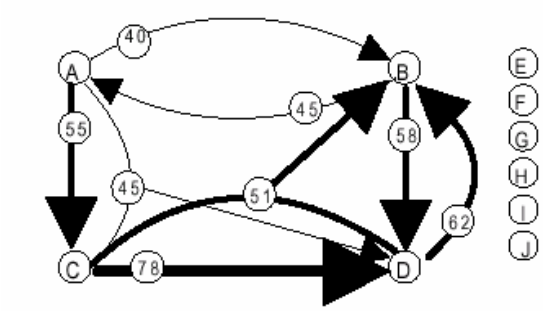


Figura 17: Representação gráfica de regras associativas

Com o mesmo propósito de facilitar a recuperação de padrões, abordagens alternativas de visualização propõem a representação gráfica considerando características específicas. Blanchard *et al.* [BLA03], por exemplo, propõem a representação gráfica de regras associativas disponibilizadas em uma arena, onde cada uma é constituída por uma esfera e um cone. A altura do cone é proporcional à confiança da regra e o diâmetro da esfera representa o suporte. A posição das regras na arena determina o quanto elas são novas para o domínio da aplicação. A Figura 18 demonstra como as regras associativas são visualizadas e disponibilizadas na arena.

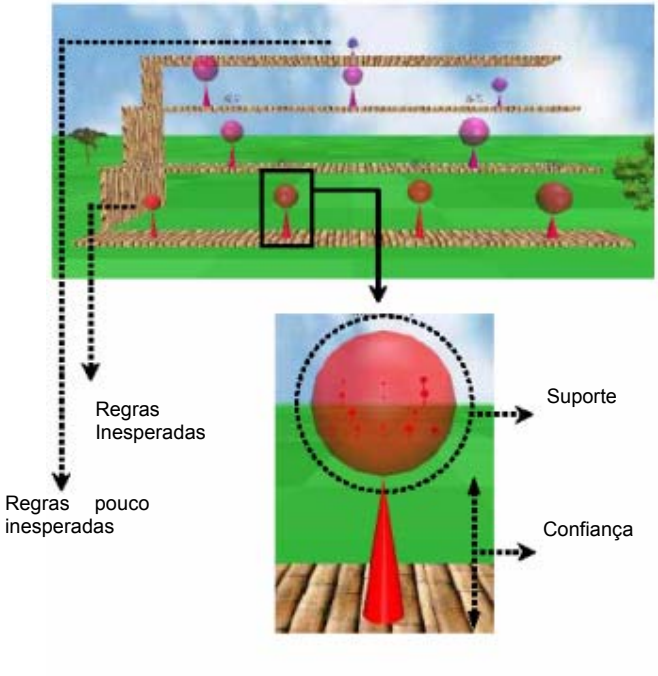


Figura 18: Regras associativas na arena

3.4 Considerações

Como apresentado nas seções anteriores, diversas são as abordagens propostas para dar apoio à fase de Análise de Padrões, cada qual com suas contribuições e limitações. Desta forma, se torna conveniente a combinação das diferentes abordagens apresentadas, como já propõem alguns dos trabalhos analisados (e.g. [BER02c, BER00, KLE94]). A Tabela 1 apresenta um comparativo entre as principais características destas abordagens, visando destacar suas limitações, as quais serviram como motivação para a definição dos objetivos propostos para este trabalho.

Como descrito, existe uma carência muito grande em trabalhos que explorem uma Ontologia de Domínio para o processo de MUW. Os que existem focam-se (e.g. [DAI02, OBE03]) apenas na fase de Preparação de Dados, oferecendo ainda limitações quanto à criação do *log* semântico e interpretação dos padrões, já discutidas na Seção 3.2.2.

Em propostas mais simples, o uso de taxonomia também apresenta algumas limitações quanto à interpretação de padrões, principalmente devido à falta de flexibilidade na interpretação e relacionamento dos padrões generalizados e padrões específicos correspondentes.

Quanto à recuperação de padrões, abordagens de filtragem através de filtros estatísticos limitam o interesse do analista em medidas estatísticas. Como complemento, filtros estruturais oferecem outros critérios, como conteúdo e estrutura. Porém, estes requerem domínio de uma sintaxe para a especificação dos filtros e também profundo conhecimento do domínio requerido para expressar o que é relevante.

Com base nas deficiências apresentadas pelas abordagens pesquisadas, o presente trabalho propõe um conjunto de mecanismos que visa assistir os analistas durante a fase de Análise de Padrões no processo de MUW, suportando as atividades de interpretação e recuperação dos padrões seqüenciais de navegação através do uso de Ontologia de Domínio previamente disponível.

Cabe ressaltar que a fase de Análise de Padrões foi a escolhida para este trabalho para permitir um enriquecimento dinâmico dos *logs*, à medida que a interpretação suscita diferentes tipos de interesse.

Tabela 1. Comparação das abordagens

	Abordagens				
	Filtragem		**Semântica**		**Representação**
	Filtros estatísticos	Filtros Estruturais	Taxonomia	Ontologia	
Objetivo	Recuperar padrões		Interpretar padrões		Recuperar e Interpretar
Contribuição	Reduzir o número de padrões e restringir o foco de pesquisa em determinado conjunto de padrões		Facilitar a interpretação de Padrões		Facilitar a recuperação e interpretação de padrões relevantes através da representação gráfica
Fase do Processo	Mineração de Dados Análise de Padrões		Preparação de dados Mineração de Dados Análise de Padrões	Preparação de dados	Análise de Padrões
Desvantagens	- Medidas objetivas não necessariamente determinam um padrão interessante; - Medidas Subjetivas: Dificuldade em expressar o conhecimento do domínio e limitações dos algoritmos de comparação.	- Domínio de uma linguagem para especificação dos filtros; - Necessidade de objetivos claros para a definição dos filtros; - Necessidade de profundo conhecimento do domínio; - Re-execução da fase de Mineração de Dados para cada novo objetivo.	- Limitação à utilização de relações do tipo *é-um*; - Geração de muitos padrões quando associado à fase de Mineração de Dados e falta de suporte a interpretação dos padrões resultantes e seus relacionamentos; - Na fase de Análise de Padrões, está vinculada a abordagens estruturais.	- Poucos trabalhos; - Representam o conteúdo dos *sites Web;* - Trabalhos limitam-se a apresentar perspectivas da exploração da *Web* Semântica. - Enriquecimento semântico na Preparação limita a análise de padrões por impossibilitar a exploração de diferentes dimensões de interesse.	- Complementar às demais abordagens.

4 REPRESENTAÇÃO DA ONTOLOGIA DE DOMÍNIO PARA A INTERPRETAÇÃO E RECUPERAÇÃO DE PADRÕES SEQÜENCIAIS

Este capítulo apresenta os principais objetivos da abordagem proposta assim como os requisitos para a representação da Ontologia de Domínio para a interpretação e recuperação de padrões seqüenciais. Também são apresentadas algumas particularidades quanto às fases do processo de MUW.

O objetivo principal deste trabalho é propor mecanismos que facilitem a interpretação e recuperação de padrões seqüenciais de navegação através da utilização de Ontologia de Domínio disponibilizada previamente. Estes mecanismos referem-se a duas dificuldades principais encontradas na fase de Análise de Padrões: a grande quantidade de padrões resultantes da aplicação de algoritmos para a busca de padrões seqüenciais e a falta de semântica neles representada.

Os objetivos específicos são:

- propor mecanismos que facilitem a interpretação de padrões através da representação de padrões seqüenciais de URLs em padrões conceituais;
- propor mecanismos que facilitem a interpretação dos padrões conceituais através da análise exploratória da semântica destes padrões conceituais;
- propor mecanismos que auxiliem a recuperação de padrões conceituais através da definição de filtros com o uso de Ontologia de Domínio;
- definir um ambiente de apoio à fase de Análise de Padrões que incorpore estes mecanismos, permitindo uma avaliação sobre a utilidade dos mesmos.

As próximas seções deste capítulo descrevem os requisitos para aplicação da abordagem proposta. Estes se referem ao conhecimento do domínio e às particularidades das etapas do processo de MUW.

4.1 Ontologia de Domínio

Visando agregar semântica às URLs pobres em informação, propõe-se a exploração de Ontologia de Domínio previamente definida, a qual especifica os eventos do domínio, suas propriedades e relacionamentos com outros eventos. Neste contexto, eventos de domínio são considerados segundo duas dimensões, a saber, serviço e conteúdo. Este trabalho adota a classificação de Stumme *et al.* [STU02] e Berendt *et al.* [BER02a] para eventos, restrita a eventos atômicos. Assim, a ontologia especifica eventos atômicos de serviço e de conteúdo.

Como na abordagem de Oberle *et al.* [OBE03], no presente trabalho os eventos de domínio são representados em dois níveis: Nível Conceitual e Nível Físico. O primeiro é representado pela Ontologia de Domínio, e o segundo pelas URLs que compõem o *site Web*. Existe uma conexão entre estes níveis, feita através do mapeamento das URLs para os conceitos da ontologia. A Figura 19, ilustra os eventos do domínio segundo estes dois níveis, para um *site* turístico no qual os usuários podem visualizar detalhes sobre a descrição de restaurantes e hotéis, fazer reservas e assim por diante.

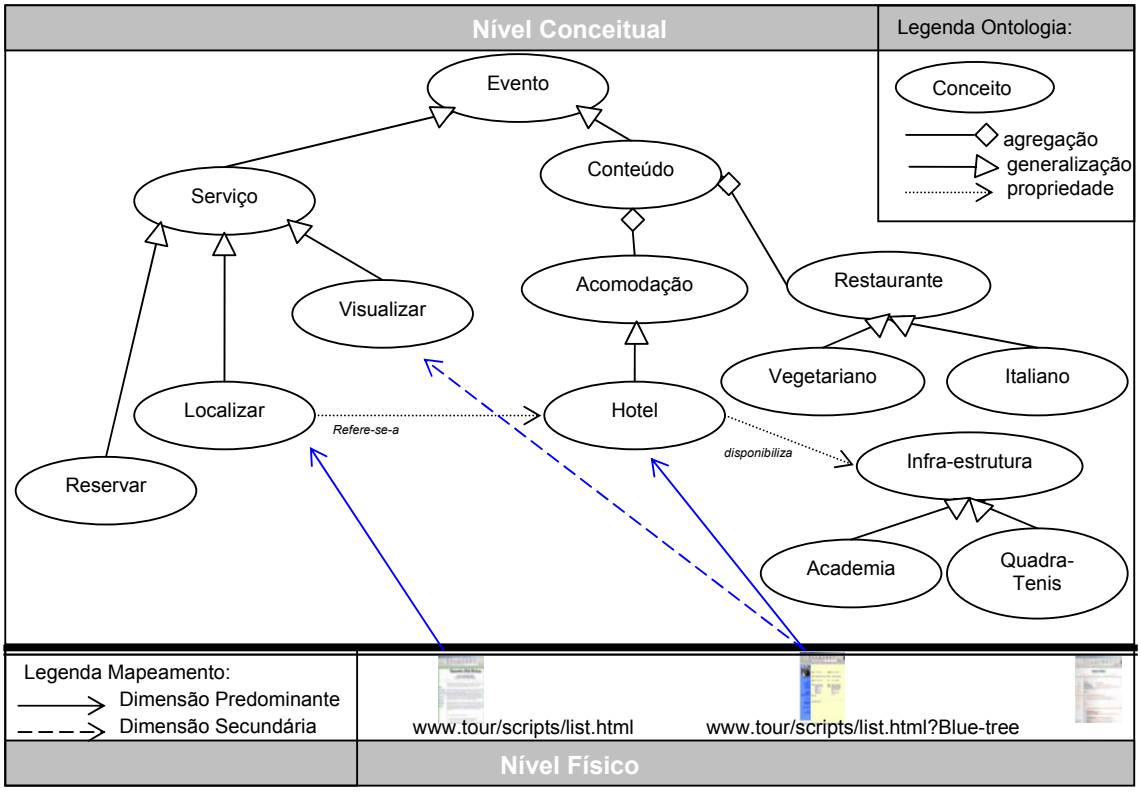


Figura 19: Níveis de representação dos eventos de domínio

4.1.1 Nível Conceitual

Neste trabalho, a Ontologia de Domínio representa e suporta o relacionamento entre os conceitos do domínio, provendo assim a semântica da aplicação. Um conceito da ontologia representa um evento de domínio. O conceito que representa um evento de conteúdo é chamado de conceito de conteúdo. O conceito que representa um evento de serviço é chamado de conceito de serviço.

Os conceitos de conteúdo e serviço relacionam-se entre si considerando dois tipos de relações: relação de hierarquia e relação de propriedade. A estrutura da ontologia está representada no diagrama de classes de UML apresentado na Figura 20.

Uma relação de hierarquia define diferentes níveis de abstração entre dois conceitos, ou seja, uma relação de hierarquia relaciona um conceito denominado ascendente a um conceito denominado descendente. O conceito ascendente é aquele que está representado num nível de abstração superior ao conceito descendente. Da mesma forma, um conceito descendente representa um nível de abstração mais detalhado que o conceito ascendente.

Um conceito descendente relaciona-se com um conceito ascendente através de apenas uma relação de hierarquia. Já um conceito ascendente pode relacionar-se com diversos conceitos descendentes.

Dois tipos de relações hierárquicas são considerados:

- generalização (relações do tipo *é-um*): corresponde à abstração de conceitos que compartilham similaridades. Por exemplo, cachorro é um mamífero. Neste exemplo, o conceito ascendente "mamífero" é uma generalização, e o conceito descendente "cachorro" é uma especialização.
- agregação (relações do tipo *parte-de*): representa associação de componentes para compor uma classe. Por exemplo, porta é parte de carro. Neste exemplo, o conceito ascendente "carro" é uma agregação, e o conceito descendente "porta" é o componente.

Os relacionamentos de propriedade definem um conceito dito sujeito, através de uma propriedade que possui um nome, para a qual um outro conceito representa o objeto. Um relacionamento de tipo propriedade não é simétrico. A tripla formada pelo sujeito,

propriedade e objeto é chamada de sentença em RDFS [BRI02]. Por exemplo, Hotel atende cliente. Neste exemplo, “Hotel” representa o conceito sujeito, “atende” corresponde ao nome da propriedade e “cliente” ao objeto.

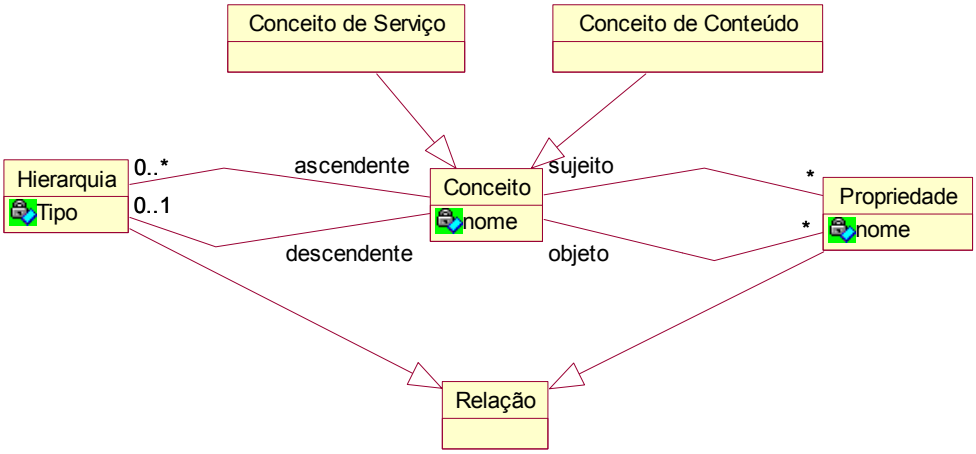


Figura 20: Estrutura da Ontologia de Domínio

As restrições que definem a estrutura da Ontologia de Domínio são necessárias para simplificar os mecanismos de interpretação e recuperação propostos neste trabalho. Estas restrições podem ser estendidas em trabalhos subseqüentes.

A ontologia descrita em Nível Conceitual da Figura 19 ilustra os diferentes tipos de conceitos e seus relacionamentos. A ontologia é representada graficamente sem seguir uma convenção específica, e pode ser encontrada no Anexo I deste volume, descrita utilizando a linguagem OWL (*Ontology Web Language*). Neste exemplo, os conceitos *Reservar*, *Localizar*, *Visualizar* são descendentes do conceito *Serviço*. *Reservar* tem como conceito ascendente *Serviço*, e *Serviço* é ascendente de *Reservar*, *Localizar* e *Visualizar*. Da mesma forma, *Acomodação* e *Restaurante* são descententes do conceito *Conteúdo*. O serviço *Localizar* é sujeito propriedade do nome *refere-se-a*, cujo objeto é o conteúdo *Hotel*.

A Ontologia de Domínio pode ser representada utilizando qualquer linguagem que permita a definição de conceitos e a relação entre eles. Estas linguagens vão desde a simplicidade proposta por RDFS [BRI02] até as que oferecem maior representação semântica como a OWL [SMI04].

4.1.2 Nível Físico e Mapeamento

URLs representam os conceitos atômicos de conteúdo e serviço em Nível Físico. As URLs disponíveis em Nível Físico são mapeadas para conceitos da ontologia (Nível Conceitual) de acordo com o evento de domínio que estas URLs representam (Figura 21). Uma URL pode ser mapeada para um conceito de serviço, um conceito de conteúdo ou ambos. Neste último caso é definida a dimensão predominante, a qual representa o principal evento de domínio simbolizado pela URL. A outra dimensão, se existente, passa a ser denominada dimensão secundária. Um único conceito da ontologia pode ser mapeado para diferentes URLs, sendo que nem todas as URLs necessitam ser mapeadas.

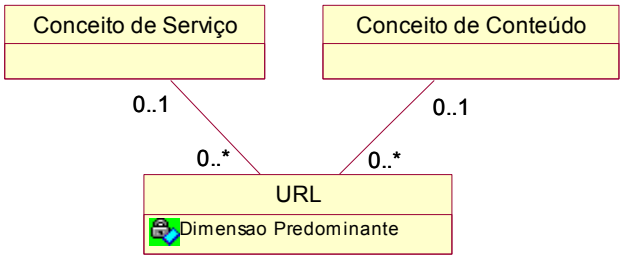


Figura 21: Mapeamento entre Nível Físico e Nível Conceitual

O Nível Físico e o seu mapeamento para o Nível Conceitual também estão exemplificados na Figura 19. O Nível Físico é constituído pelas URLs que compõem o *site* turístico. O Nível Físico relaciona-se com o Conceitual através do mapeamento das URLs para conceitos da Ontologia de Domínio. Estes conceitos utilizados no mapeamento indicam o nível de abstração mais detalhado da URL em Nível Conceitual.

Na Figura 19, a URL *www.tour/scripts/list.html* é mapeada para o conceito de serviço *Localizar*. Isso significa que ela disponibiliza o recurso de localizar informações. A URL *www.tour/scripts/list.html?Blue-tree* é mapeada para o conceito de conteúdo *Hotel*, significando que disponibiliza informações sobre um hotel em específico. Esta mesma URL é mapeada para um conceito de serviço, embora simples, que é o de visualizar informações (conceito *Visualizar*). Como uma única página está sendo mapeada para dois conceitos de tipos diferentes, é necessário definir a dimensão predominante. No caso da página *www.tour/scripts/list.html?Blue-tree*, a dimensão predominante seria a dimensão de conteúdo uma vez que o objetivo principal é disponibilizar conteúdo aos usuários.

A utilização de ontologia visa enriquecer as informações representadas pelas URLs. A partir da URL *www.tour/scripts/list.html*, utilizada como exemplo e mapeada para um conceito de serviço na Ontologia de Domínio, é possível inferir que esta URL diz respeito a localizar informações referentes a hotéis e que estes por sua vez disponibilizam infra-estrutura, como por exemplo quadra de tênis e academia.

Nota-se que a ontologia utilizada neste exemplo não se limita a uma hierarquia de conceitos (somente relações do tipo *é-um*) como apresentada em muitos trabalhos (e.g. [SRI95, BER00]), mas sim, declara diversas relações entre os objetos que enriquecem a semântica do domínio representada. Conseqüentemente, os *logs* mapeados para estas ontologias também poderão fazer uso das vantagens que ela pode proporcionar para a atividade de interpretação de padrões.

4.2 O Processo de MUW

No processo de MUW, a abordagem proposta enfoca a fase de Análise de Padrões. Desta forma, algumas premissas são consideradas para a execução do processo de MUW.

4.2.1 Criação da Ontologia de Domínio e Mapeamento

Ontologias podem ser criadas manualmente (*ad-hoc*) ou utilizando mecanismos semi-automáticos. Nesta última abordagem, técnicas de aprendizado de máquina e extração de informações vêm sendo utilizadas para melhorar o processo de construção de ontologias (e.g. [SUR02]). Com o advento da *Web* Semântica, diferentes grupos de pesquisa vêm incentivando a formalização do conhecimento representado em *sites* por ontologias de domínio.

Este trabalho não se preocupa com o processo de aquisição e validação de ontologias. Parte-se do pressuposto que ontologias já estejam disponíveis para compartilhamento de conhecimento do domínio da aplicação e as URLs estejam devidamente mapeadas. A estrutura da ontologia e os mapeamentos devem respeitar as restrições estabelecidas.

4.2.2 Preparação de Dados

Na fase de Preparação de dados, atividades típicas são executadas sobre os dados coletados do *log* do servidor *Web,* como limpeza de dados, identificação do usuário e sessões, atividades estas descritas na Seção 2.2.1. Devido à existência da Ontologia de Domínio e do mapeamento das URLs para os conceitos desta, não se assume qualquer atividade relativa ao enriquecimento semântico dos *logs* no tocante a URLs a partir da Ontologia de Domínio.

4.2.3 Mineração de Dados

Nesta fase é considerada a técnica de descoberta de padrões seqüenciais definida pelo algoritmo *AprioriAll* [AGR94a], descrito na Seção 2.2.2.1. Assim, tem-se como resultado desta fase um conjunto de padrões seqüenciais formados por URLs, com o respectivo suporte.

4.2.4 Análise de Padrões

Os mecanismos propostos para este trabalho são aplicados nesta fase. Desta forma, algumas entradas se fazem necessárias. Devem ser considerados os dados pré-processados resultantes da fase de Preparação de Dados, a lista de padrões seqüenciais formados por URLs resultante da fase de Mineração de Dados, a Ontologia de Domínio, assim como o mapeamento das URLs para os conceitos da ontologia. A Figura 22 representa as entradas para a fase de Análise de Padrões. Mais detalhes sobre estas nos próximos capítulos.

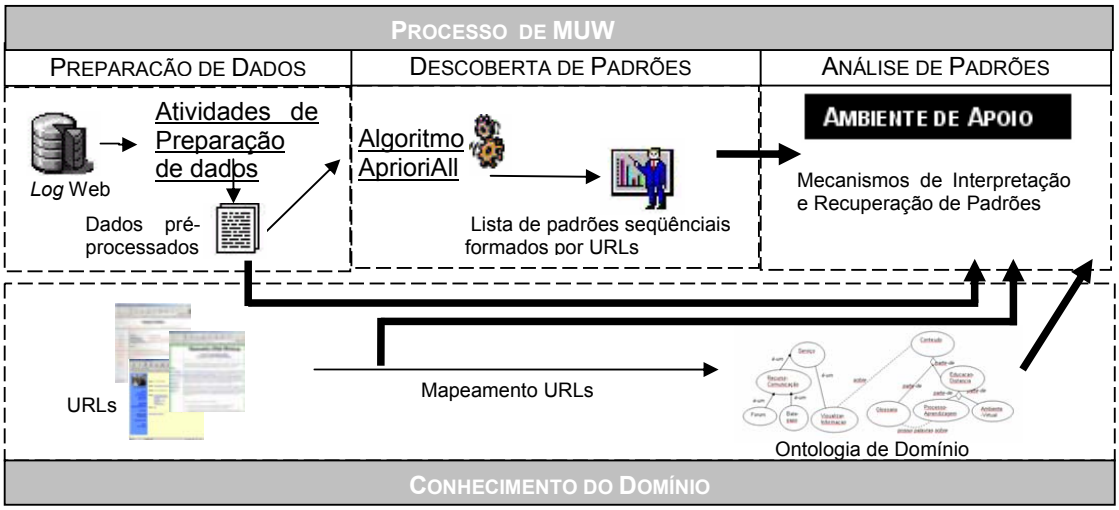


Figura 22: Entradas para a fase de Análise de Padrões

5 MECANISMOS DE INTERPRETAÇÃO DE PADRÕES DO USO DA *WEB*

Este capítulo descreve os mecanismos propostos para apoiar a atividade de interpretação de padrões, através da representação dos padrões seqüenciais conceituais e da análise exploratória destes.

O processo de MUW é eficaz quando padrões relevantes ao domínio da aplicação são identificados. Porém, anteriormente ao julgamento referente à importância de um padrão para um domínio, é fundamental entender o conhecimento expresso por ele.

Visando facilitar a atividade de interpretação de padrões, esta abordagem propõe alguns mecanismos que fazem uso da Ontologia de Domínio e do mapeamento previamente definidos. Estes mecanismos referem-se a: representação de padrões seqüenciais de URLs na forma de padrões seqüenciais conceituais de acordo com uma dimensão de interesse; e a análise exploratória dos padrões conceituais, que permite um aprofundamento da compreensão do significado destes.

Estes mecanismos foram definidos visando complementar as deficiências apresentadas pelos trabalhos relatados no Capítulo 3. Os mecanismos propostos para interpretação de padrões são detalhados no restante deste capítulo.

5.1 Representação de Padrão Seqüencial Conceitual

Uma das características dos padrões interessantes é a sua simplicidade de modo a possibilitar a compreensão do conhecimento pelos analistas. Sem o enriquecimento semântico provido na fase de Análise de Padrões, os resultados da fase de Mineração de Dados são padrões seqüenciais compostos por URLs ordenadas, denominados neste trabalho de padrões seqüenciais físicos.

Explorando a Ontologia de Domínio disponível e o mapeamento das URLs para os conceitos da ontologia, este trabalho propõe representar os padrões seqüenciais físicos na forma de padrões seqüenciais conceituais visando facilitar o entendimento do conhecimento expresso por um padrão seqüencial.

Um padrão seqüencial conceitual é um padrão formado por uma seqüência ordenada de conceitos definidos pela Ontologia de Domínio. A denominação de padrão conceitual foi inspirada na abordagem de Oberle *et al.* [OBE03], o qual refere-se a um agrupamento de conceitos da ontologia que representa os conteúdos do *site* acessados pelos usuários como "*Conceptual User Tracking*" (Capítulo 3, Seção 3.2.2).

Um padrão conceitual é uma representação conceitual de um padrão seqüencial físico de acordo com uma dimensão de interesse especificada pelo analista. A dimensão de interesse pode ser de serviço, conteúdo ou serviço e conteúdo. Esta dimensão é definida de acordo com o interesse do analista nos eventos de domínio envolvidos no padrão seqüencial físico. Desta forma, é possível interpretar um mesmo padrão seqüencial físico considerando 3 dimensões de interesse.

A dimensão de serviço permite visualizar um padrão seqüencial físico como um conjunto de conceitos de serviço, que constituem o padrão seqüencial conceitual. A dimensão de conteúdo representa um padrão seqüencial físico como um padrão seqüencial conceitual formado por conjunto de conceitos de conteúdo. Já a dimensão de serviço e conteúdo representa um padrão seqüencial físico como um padrão seqüencial conceitual definido por um conjunto de conceitos de serviço ou conteúdo, representados de acordo com a dimensão predominante da URL.

Para exemplificar a utilidade destas dimensões, considera-se o *site* de busca por hotéis e restaurantes descrito anteriormente, para o qual a Ontologia de Domínio da Figura 19 foi desenvolvida. As URLs foram mapeadas para os conceitos da ontologia, como descrito pela Tabela 2.

Tabela 2. Mapeamento das URLs para os conceitos da ontologia

URL	Conceito de Serviço	Conceito de Conteúdo
`www.tour/scripts/list.html`	Localizar (DP)	Hotel
`www.tour/scripts/list.html?Blue-tree`	Visualizar	Hotel (DP)
`www.tour/scripts/commited`	Reservar (DP)	Hotel
`www.tour/scripts/list.html?NewLife`	Visualizar	Restaurante (DP)

Legenda:
DP –Dimensão Predominante;

```
www.tour/scripts/serve_find.html  ⟶  www.tour/scripts/list.html?Blue-tree  ⟶
www.tour/scripts/commited ⟶ www.tour/scripts/list.html?NewLife.
```

Figura 23: Exemplo do padrão seqüencial físico

A Tabela 3 apresenta os três padrões seqüenciais conceituais, correspondentes ao padrão seqüencial físico da Figura 23, de acordo com as diferentes dimensões de interesse previamente definidas pelo usuário.

Tabela 3. Exemplo de padrões seqüenciais conceituais

Dimensão de Interesse	Padrão Seqüencial Conceitual
Serviço e Conteúdo	Localizar → Hotel → Reservar → Vegetariano
Serviço	Localizar → Visualizar → Reservar → Visualizar
Conteúdo	Hotel → Hotel → Hotel → Vegetariano

Nota-se que interpretar o conhecimento representado por um padrão seqüencial conceitual é mais fácil de interpretar do que por um padrão seqüencial físico. Por exemplo, o padrão conceitual associado à dimensão de serviço e conteúdo da Tabela 3 claramente expressa que um grupo de usuários localizou informações, acessou uma página sobre um hotel específico, requisitou uma reserva e posteriormente acessou uma página sobre o conteúdo vegetariano. A interpretação do mesmo padrão seqüencial físico através de um padrão seqüencial conceitual de acordo com a dimensão de serviço é similar à representada anteriormente. A única diferença é que as páginas de conteúdo passam a ser interpretadas pelo serviço que oferecem. Neste caso, a segunda e a última URLs são interpretadas usando o serviço de visualizar informações. Já considerando a dimensão de conteúdo, o mesmo padrão seqüencial é facilmente interpretado, expressando que os usuários do *site* estão interessados em hotéis e vegetarianos.

Cabe salientar que neste exemplo, todas as URLs foram mapeadas para as dimensões de serviço e conteúdo. Caso uma URL não tenha sido mapeada para a dimensão de interesse, o analista é informado que o serviço ou conteúdo está indisponível.

O diagrama de classes UML da Figura 24 ilustra as classes envolvidas e os relacionamentos entre estas.

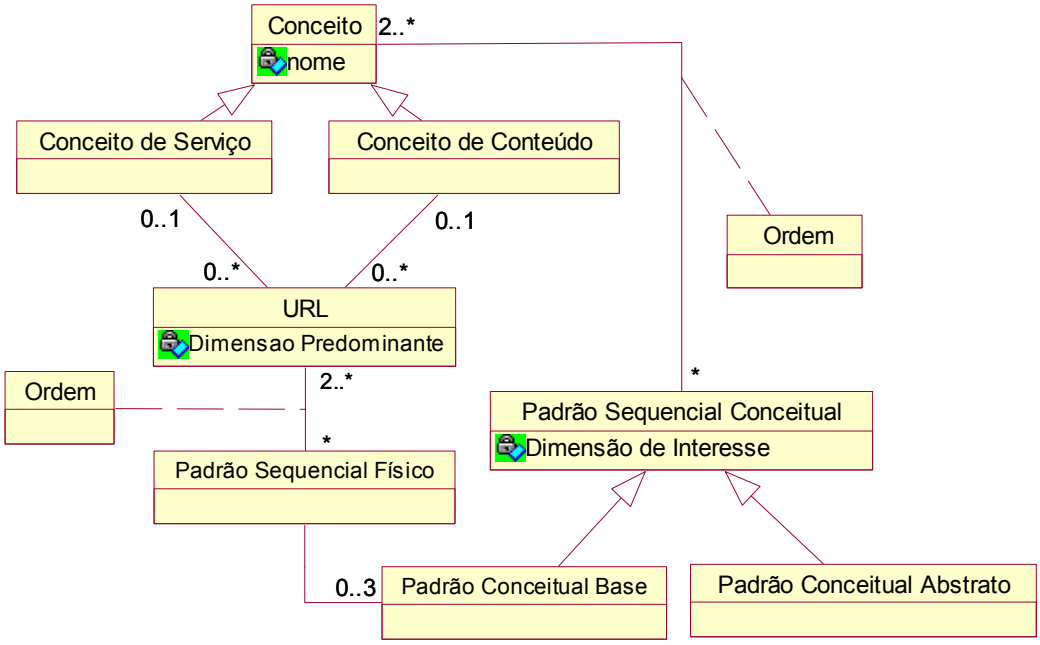


Figura 24: Padrão Seqüencial Conceitual

Um padrão seqüencial conceitual é dito padrão conceitual base quando ele é representado pelos conceitos que foram utilizados no mapeamento entre os níveis Físico e Conceitual do respectivo padrão seqüencial físico. Desta forma, os padrões seqüenciais conceituais representados na Tabela 3 são padrões conceituais base.

Um padrão conceitual base corresponde assim a no máximo um padrão seqüencial físico. Já um padrão conceitual abstrato corresponde a vários. Um padrão conceitual abstrato é derivado originalmente de um padrão seqüencial base. A diferença é que ele é formado por pelo menos um conceito ascendente obtido a partir de um conceito do padrão conceitual base. Um padrão conceitual abstrato é criado a partir da operação de *roll-up* que compõe a análise exploratória, descrita na próxima seção.

Um padrão seqüencial conceitual pode ser representado gráfica ou textualmente.

5.2 Análise Exploratória

Este trabalho também propõe um conjunto de mecanismos que constituem a análise exploratória, permitindo ao analista investigar as relações dos conceitos que formam um padrão seqüencial conceitual com os demais definidos pela Ontologia de Domínio.

A análise exploratória é composta pelas operações de:

- detalhamento de relacionamentos;
- *roll-up*;
- *drill-down*.

Estes mecanismos são requisitados pelo analista e suportam a interação com os padrões seqüenciais conceituais. Estas operações são aplicáveis sobre a representação gráfica do padrão conceitual base. As operações de *roll-up* e *drill-down* foram definidas em analogia às operações de mesmo nome propostas pela tecnologia OLAP.

Estes mecanismos são abordados em detalhes nas próximas seções.

5.2.1 Detalhamento de Relacionamentos

A operação de detalhamento de relacionamentos explora a semântica das relações e dos conceitos que estão associados aos conceitos que compõem um padrão seqüencial conceitual (base ou abstrato). Neste trabalho, esta operação distingue-se de acordo com o tipo de relação utilizada para conectar dois conceitos, podendo ser operação de detalhamento de hierarquia e de propriedade.

A operação definida como detalhamento de hierarquia permite investigar os conceitos ascendentes aos conceitos que compõem um padrão seqüencial conceitual, isto é, que estão conectados através de uma relação de hierarquia. Portanto, ela pode ser executada sobre um conceito de um padrão seqüencial conceitual sempre que existir um conceito ascendente. A representação gráfica do padrão seqüencial conceitual indica quando a operação de detalhamento de hierarquia está habilitada.

A Figura 25 ilustra graficamente um padrão conceitual base formado por quatro conceitos definidos conforme mapeamento da Tabela 2 e seu detalhamento de hierarquia com

base na Ontologia de Domínio definida pela Figura 19. Neste exemplo, a possibilidade de utilização desta operação de detalhamento de hierarquia é representada por uma seta apontando para cima em cada conceito que compõem o padrão conceitual base. Assim, percebe-se que a operação de detalhamento de hierarquias pode ser executada sobre todos os conceitos do padrão conceitual base.

Requisitando esta operação sobre o conceito *Localizar*, descobre-se que este é um tipo de serviço oferecido pelo *site* turístico. Solicitando a mesma operação sobre o conceito *Hotel*, verifica-se que hotel é um tipo de acomodação. Finalmente, realizando esta operação sobre o conceito *Acomodação* é possível interpretar que acomodação faz parte do conteúdo disponibilizado pelo *site*.

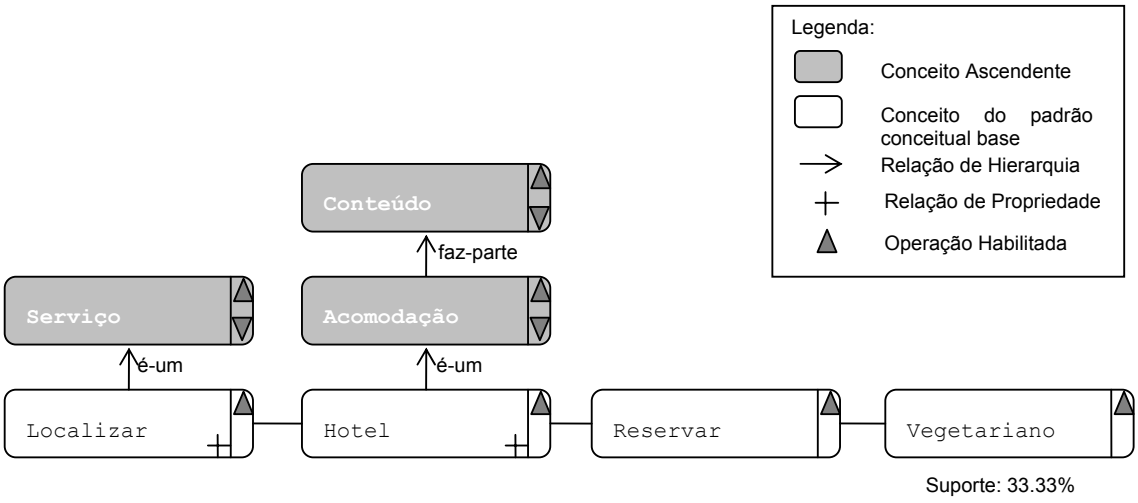


Figura 25: Detalhamento de hierarquias

Outra operação complementar ao detalhamento de hierarquias, é o detalhamento de propriedade entre os conceitos que compõe o padrão seqüencial conceitual e os existentes na Ontologia de Domínio. O conceito que possui uma relação de propriedade deve apresentar um símbolo que indique a existência da relação, significando que esta operação está habilitada. Ao requisitar a operação de detalhamento da relação de propriedade sobre um conceito, o significado desta é mostrado através de uma sentença.

Por exemplo, na ontologia representada pela Figura 19 existe uma relação de propriedade entre os conceitos *Localizar* e *Hotel*, chamada *refere-se-a*. Desta forma, qualquer

padrão seqüencial conceitual que apresentar o conceito *Localizar* poderá utilizar o significado desta propriedade para auxiliar na interpretação do padrão. Como representado na Figura 26, os conceitos *Localizar* e *Hotel* estão relacionados com outros conceitos através de relações de propriedade sinalizadas por uma cruz. Explorando-as, é possível interpretar que o evento de localizar informações no *site* refere-se a informações sobre hotéis, que por sua vez disponibilizam infra-estrutura.

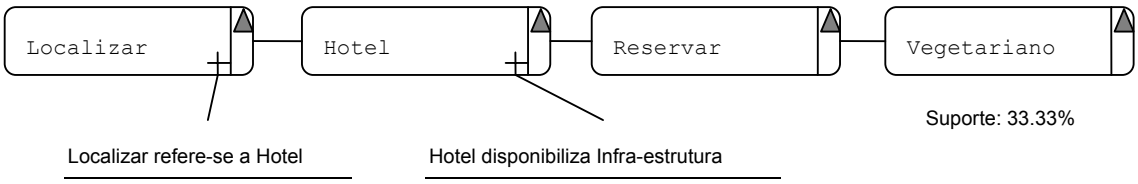


Figura 26: Detalhamento de relacionamentos

5.2.2 *Roll-up*

A operação de *roll-up* refere-se à sumarização de padrões seqüenciais conceituais através da substituição de um conceito que compõe o padrão por seu conceito ascendente. Quando aplicada a um conceito que compõe um padrão conceitual base, a operação de *roll-up* tem como resultado um padrão conceitual abstrato. Quando aplicada a um conceito que compõe um padrão conceitual abstrato, a operação de *roll-up* gera outro padrão conceitual mais genérico. Portanto, um padrão conceitual abstrato é uma abstração de um ou mais padrões conceituais base, e por conseguinte, de um ou mais padrões seqüenciais físicos.

A Figura 27 representa três padrões seqüenciais conceituais visualizados de acordo com a dimensão de interesse serviço e conteúdo. Dois destes padrões são padrões conceituais abstratos resultantes de sucessivas operações de *roll-up*. Observa-se um padrão conceitual base, a partir do qual foi obtido o padrão conceitual abstrato 1, aplicando-se a operação de *roll-up* sobre o conceito *Hotel* que compõe o padrão conceitual base. Como resultado, foi gerado um padrão conceitual base onde o conceito hotel foi substituído pelo seu ascendente acomodação, de acordo com a ontologia da Figura 19.

Da mesma forma, o padrão conceitual abstrato 2 foi obtido através de uma operação de ro*ll-up* sobre o conceito *Acomodação* do padrão conceitual abstrato 1, que o substituiu pelo conceito ascendente *Conteúdo*.

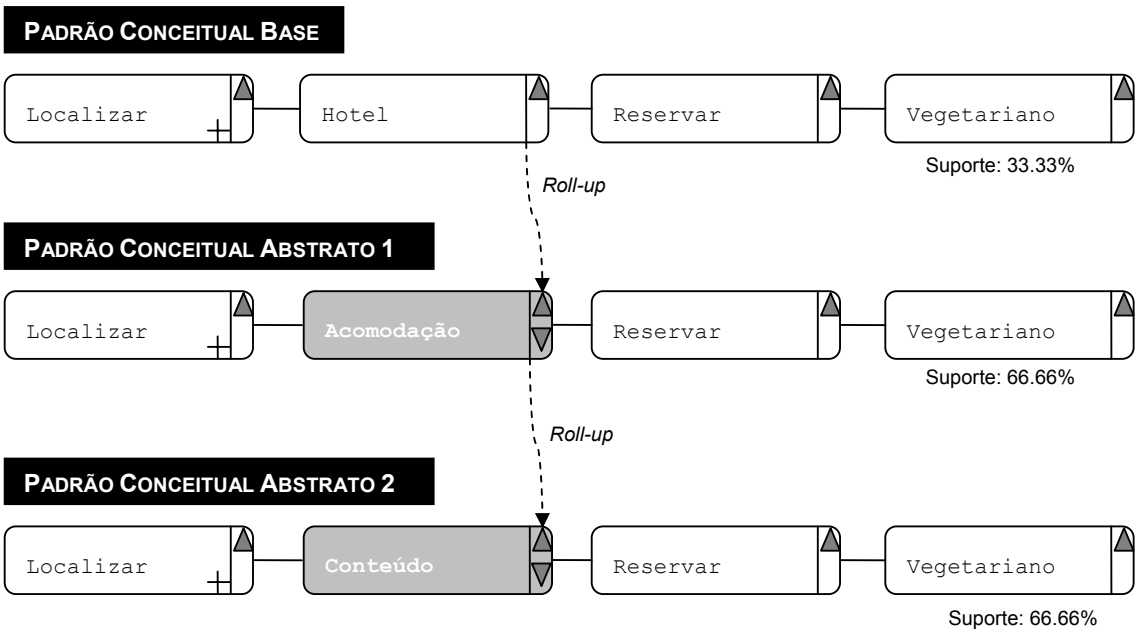


Figura 27: Padrão Conceitual Base e Padrões Conceituais Abstratos

Um padrão conceitual abstrato sumariza todos os padrões conceituais base que estiverem de acordo com as restrições expressas pelo padrão. Por exemplo, de acordo com a Ontologia de Domínio representada na Figura 19, o padrão conceitual abstrato 1 da Figura 27 suporta todos os padrões seqüenciais conceituais que possuem a estrutura "*Localizar - (Hotel ou Pensionato) - Reservar - Restaurante*".

O suporte dos padrões seqüenciais conceituais varia conforme o número de sessões que estão de acordo com as restrições definidas pelo padrão seqüencial conceitual. Nota-se que o suporte do padrão conceitual abstrato 1 (66,66%) é maior do que o padrão conceitual base que originou o padrão conceitual abstrato (33,33%). Isto significa que existem outros padrões conceituais base suportados por ele além daquele originalmente utilizado para gerá-lo ("*Localizar - Hotel - Reservar – Vegetariano"*).

Sempre que um padrão conceitual abstrato é criado, o valor do suporte para aquele padrão deve ser calculado. Para um padrão conceitual base isso não é preciso pois o valor do suporte é o mesmo do padrão seqüencial físico ao qual ele está associado. A seção seguinte descreve como este cálculo é realizado.

5.2.2.1 *Suporte de um Padrão Conceitual Abstrato*

A criação de um padrão conceitual abstrato requer cálculo do suporte. O suporte de um padrão conceitual abstrato corresponde ao percentual de sessões de usuários que suportam aquele padrão conceitual abstrato. Para que uma sessão suporte um padrão conceitual abstrato é necessário que a seqüência de URLs da sessão esteja mapeada para a seqüência de conceitos do padrão conceitual abstrato, considerando a dimensão de interesse. Lembrando que a seqüência de conceitos, imediatos ou não, considera os conceitos descendentes.

O cálculo do suporte é realizado por uma função que recebe como parâmetro de entrada: um padrão conceitual abstrato; o *log* pré-processado resultante da fase de Preparação de Dados; a Ontologia de Domínio; mapeamento que define a associação das URLs para os conceitos da ontologia e a dimensão de interesse.

O padrão conceitual abstrato, a Ontologia de Domínio, a dimensão de interesse e o mapeamento são utilizados para definir quais sessões suportam o padrão conceitual abstrato. Já o *log* pré-processado é utilizado na contagem das sessões de usuários que estão de acordo com os possíveis padrões conceituais base candidatos. Desta forma, o *log* pré-processado é fundamental para o cálculo do suporte do padrão conceitual abstrato, uma vez que este não corresponde à soma do valor do suporte dos padrões seqüenciais físicos que são sumarizados por ele.

A função que determina o cálculo do suporte é composta pelos seguintes passos:

1. identificar os possíveis padrões conceituais base candidatos que são sumarizados pelo padrão conceitual abstrato.
    a. Para cada conceito do padrão conceitual abstrato, identificar quais são os conceitos descendentes na ontologia. Para cada conceito descendente, por sua vez, são verificados seus descendentes, e assim recursivamente. Do conjunto de descendentes assim extraído, eliminam-se todos os que não possuem um mapeamento para URL na dimensão de interesse, uma vez que estes conceitos não poderiam compor os padrões conceituais base sumarizados pelo padrão conceitual abstrato.

b. identificar a seqüência na qual os conceitos devem aparecer nos possíveis padrões conceituais base candidatos. Esta seqüência tem que estar de acordo com a ordem dos conceitos no padrão conceitual abstrato, ou seja, uma determinada posição ocupada por um conceito ascendente no padrão abstrato pode ser ocupada por qualquer conceito descendente deste no padrão conceitual base.

2. encontrar os padrões físicos candidatos que referenciam os padrões conceituais base candidatos. Para identificar os padrões seqüenciais físicos candidatos é necessário verificar para quais URLs os conceitos dos padrões conceituais base candidatos estão mapeados de acordo com a dimensão de interesse.

3. identificar quais as sessões de usuários que contêm a seqüência definida pelos padrões seqüenciais físicos candidatos.

4. somar as sessões resultantes do passo 3 e dividir o valor obtido da soma pelo total de sessões presentes no *log* pré-processado. Este valor corresponde ao valor do suporte do padrão conceitual abstrato.

Para ilustrar o cálculo do suporte, considera-se o padrão conceitual abstrato 1 representado na Figura 27, obtido a partir do padrão conceitual base também representado na Figura 27, a Ontologia de Domínio ilustrada na Figura 19, o *log* pré-processado da Tabela 4 e o mapeamento para os conceitos da ontologia na Tabela 5.

Tabela 4. Dados preparados resultantes da fase de Preparação de Dados.

Sessão	Seqüência de acesso as URLs
1	URL1; URL2; URL3, URL4; URL5; URL6
2	URL1; URL3; URL4; URL5; URL6
3	URL1; URL2; URL4; URL5; URL6
4	URL1; URL2; URL3; URL5; URL6
5	URL1; URL2; URL4; URL5; URL6
6	URL2; URL5; URL3; URL4; URL6

Tabela 5. Mapeamento

URL	Conceito de Serviço	Conceito de Conteúdo
URL1		Turismo
URL2	Localizar (DP)	Hotel
URL3	Visualizar	Hotel (DP)
URL4	Visualizar	Pensionato (DP)
URL5	Reservar (DP)	Hotel
URL6	Visualizar	Vegetariano (DP)

Legenda:
DP –Dimensão Predominante;

Primeiramente, o passo 1-a é executado retornando o conjunto de conceitos descendentes incluídos no padrão conceitual abstrato. Este conjunto é composto pelos conceitos *Localizar*, *Hotel*, *Pensionato*, *Reservar* e *Vegetariano*.

Em seguida, o passo 1-b define a seqüência na qual estes conceitos devem aparecer nos possíveis padrões conceituais base, sendo esta: *Localizar*, (*Hotel ou Pensionato)*, *Reservar* e *Vegetariano*. Nota-se que a segunda posição deve ser ocupada por um dos conceitos (*Hotel* ou *Pensionato*) descendentes do conceito *Acomodação*.

De acordo com o passo 2, os padrões seqüenciais físicos candidatos são identificados. São eles "URL2 - URL3 - URL5 - URL6" e "URL2 – URL4 - URL5 - URL6", considerando a dimensão de interesse em serviço e conteúdo. Posteriormente, as seqüências de URLs das sessões do *log* preparado são comparadas com os padrões seqüenciais físicos (passo 3).

Somente as sessões 1, 3, 4 e 5 são utilizadas para o cálculo do suporte (passo 4) por serem compostas por uma seqüência de URLs definida pelos padrões seqüenciais físicos candidatos. O valor final do suporte é calculado pela divisão do número de sessões resultantes do passo 3 pelo número total de sessões, ou seja, 4/6 corresponde a um suporte de 66,66%.

O padrão conceitual abstrato sumariza dois padrões seqüenciais físicos. Eles estão representados pela Figura 28. O primeiro padrão seqüencial físico é suportado pelas sessões 1 e 4. Já o segundo, pelas sessões 1, 3 e 5. É importante ressaltar que o suporte de um padrão conceitual abstrato não é a soma dos padrões seqüenciais físicos sumarizados, pois, como neste caso, uma mesma sessão do *log* pré-processado pode estar suportando diferentes padrões seqüenciais físicos. Devido a este fato, é fundamental que o *log* pré-processado seja utilizado no cálculo do suporte dos padrões conceituais abstratos.

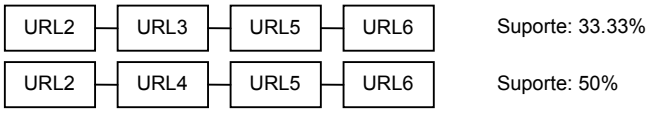


Figura 28: Padrões Seqüenciais Físicos

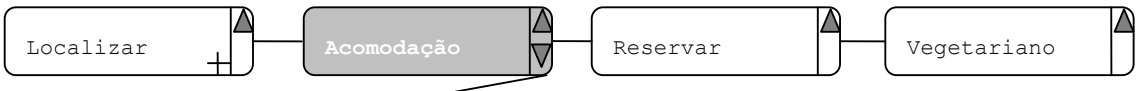
Uma alternativa para o cálculo do suporte implica a geração de todos os padrões generalizados possíveis utilizando a Ontologia de Domínio durante a fase de Mineração de Dados (e.g. [SRI95, SRI97]). Desta forma, o valor do suporte de um padrão abstrato estaria

previamente calculado, bastando apenas consultá-lo de acordo com o padrão conceitual abstrato criado. A vantagem estaria no ganho com o tempo de processamento para o cálculo do suporte. A desvantagem é que muitos padrões generalizados seriam descobertos e mantidos desnecessariamente, assim como um aumento no tempo de processamento consumido pela fase de Mineração de Dados.

5.2.3 *Drill-down*

A operação de *drill-down* é utilizada para encontrar os padrões conceituais base sumarizados por um padrão conceitual abstrato, isto é, padrões conceituais base que estão de acordo com as restrições definidas pelo padrão conceitual abstrato. Estes padrões são denominados padrões conceituais detalhe. Ao contrário das demais operações propostas para a análise exploratória, a operação de *drill-down* somente pode ser aplicada sobre os padrões conceituais abstratos. A representação gráfica do padrão conceitual abstrato indica sobre quais conceitos podem ser aplicadas as operações *drill-down*.

Por exemplo, ao requisitar a operação de *drill-down* sobre o conceito *Acomodação* do padrão conceitual abstrato 1 representado na Figura 27, todos os padrões conceituais base que suportam aquele padrão abstrato devem ser retornados, como representado pela Figura 29. Desta forma, é possível explorar os padrões conceituais detalhe com seu respectivo suporte num nível de abstração mais detalhado que o representado pelo padrão conceitual abstrato. Neste exemplo, sabe-se que existem grupos de usuários que estão interessados em hotéis e outro grupo em pensionatos. Note que a operação de *drill-down* aplicado ao padrão conceitual abstrato 2 da Figura 27 teria o mesmo resultado.



Padrão Conceitual Detalhe	Suporte
Localizar - Hotel - Reservar – Vegetariano	33.33%
Localizar - Pensionato - Reservar – Vegetariano	50%

Figura 29: Operação de *drill-down*

Os padrões conceituais base são encontrados a partir da lista de padrões seqüenciais físicos retornados da fase de Mineração de Dados. Os dois primeiros passos são semelhantes aos utilizados para o cálculo do valor do suporte, uma vez que definem as restrições que devem ser respeitadas pelos padrões seqüenciais físicos candidatos a partir de um padrão conceitual abstrato. Os passos seguintes são:

1. identificar quais os padrões seqüenciais físicos candidatos respeitam as restrições, ou seja, suportam os padrões conceituais base candidatos identificados no passo 1-a e de acordo com a seqüência do passo 1-b.
2. transformar cada padrão seqüencial físico candidato em um padrão conceitual base através do mapeamento das URLs para os conceitos da ontologia e da dimensão de interesse.
3. mostrar o conjunto de padrões conceituais detalhe e seu respectivo suporte.

5.3 Considerações

Os mecanismos propostos para facilitar a interpretação de padrões foram definidos com base nas deficiências apresentadas pelas abordagens semânticas pesquisadas. Primeiramente, optou-se pelo uso de Ontologia de Domínio pelo fato de fornecer maior suporte à representação semântica do que uma taxonomia, e devido à motivação impulsionada pela *Web* Semântica na formalização do conhecimento na forma de Ontologia de Domínio. Desta forma, é possível explorar as vantagens proporcionadas pelo conhecimento do domínio especificado para outros propósitos, integrando-o à fase de Análise de Padrões no processo de MUW.

As abordagens semânticas (e.g. [DAI02, OBE03]) pesquisadas utilizam Ontologia de Domínio na fase de Preparação de Dados, preocupando-se com o enriquecimento semântico dos *logs* (*log* semântico) para posterior descoberta e análise. A diferença para a abordagem proposta está na fase em que o conhecimento do domínio é explorado no processo de MUW, a saber, fase de Análise de Padrões. A vantagem é que nesta fase existe uma flexibilidade na interpretação de padrões considerando as diferentes dimensões de interesse. Esta flexibilidade não seria possível a partir do *log* semântico por este ser estático, representando uma única dimensão de interesse e limitando a atividade de análise. A análise de outras dimensões de

interesse implica a re-execução do processo de MUW, desde a Preparação de Dados à Análise de Padrões, como discutido na Seção 3.2.2.1.

Ontologias são úteis quando aplicadas ao processo de MUW, mas as abordagens semânticas estudadas (e.g. [DAI02, OBE03]) exploram as ontologias apenas para formalização do conteúdo dos *sites*, não considerando os serviços oferecidos por estes. A abordagem proposta por este trabalho considera que a Ontologia de Domínio especifica tanto o conteúdo como os serviços suportados pelo *site Web.* Afinal, os serviços disponíveis no *site Web* também são responsáveis por motivar a navegação dos usuários pelas páginas *Web*.

Um dos diferenciais da abordagem proposta para o processo tradicional de descoberta de padrões seqüenciais está nos mecanismos de interpretação, que facilitam o entendimento dos padrões seqüenciais na fase de análise, que até então eram compostos por um conjunto de URLs de difícil entendimento. Desta forma, a representação dos padrões seqüenciais físicos em padrões seqüenciais conceituais ameniza o esforço do analista para interpretar o significado dos padrões seqüenciais. Ainda, o analista não necessita ter profundo conhecimento do domínio, uma vez que a ontologia representa parte deste conhecimento.

Além do mais, através das operações de análise exploratória é possível aprofundar a compreensão do conhecimento suportado pelos padrões seqüenciais conceituais de forma interativa, descobrindo conceitos e outros padrões relacionados, principalmente no que diz respeito à relação dos padrões conceituais abstratos com os padrões conceituais detalhe. Abordagens apresentadas por Srikant *e* Agrawal [SRI95, SRI97] propõem a extensão de algoritmos de geração de padrões visando a geração de padrões generalizados. Porém limitam-se gerar todos os padrões possíveis de acordo com a taxonomia associada, onde nem todos os padrões resultantes são de interesse dos usuários, dificultando a fase de análise devido ao grande número de padrões retornados; redundância entre padrões; e inexistência de um relacionamento explícito entre padrões especializados e generalizados.

A Tabela 6 apresenta um comparativo da abordagem atual no que diz respeito aos mecanismos de interpretação de padrões com as abordagens semânticas pesquisadas na literatura e detalhadas no Capítulo 3, Seção 3.2.

Tabela 6. Comparação da abordagem proposta X abordagens semânticas pesquisadas.

	Abordagens Semânticas		
	Taxonomia	Ontologia - Abordagens pesquisadas	Ontologia - Abordagem de Vanzin
Objetivo	Interpretar padrões		
Contribuição	- classificação dos serviços oferecidos pelas páginas Web resultantes de consultas geradas dinamicamente; - geração de padrões generalizados na fase de Mineração de Dados; - definição de filtros através de conceitos da taxonomia;	- Definição do *log* semântico;	- representação dos padrões seqüenciais físicos na forma de padrões seqüenciais conceituais; - visualizaçã dos padrões seqüenciais físicos de acordo com diferentes dimensões de interesse; - análise exploratória dos padrões conceituais, através da operação de detalhamento de relacionamento, *roll-up* e *drill-down*.
Fase do Processo de MUW	Preparação de Dados Mineração de Dados Análise de Padrões	Preparação de Dados	Análise de Padrões
Desvantagens	- Limitação à utilização de relações do tipo *é-um*; - Geração de muitos padrões quando associado à fase de Mineração de Dados e falta de suporte a interpretação e relacionamento dos padrões resultantes e seus relacionamentos; - Na fase de Análise de Padrões, está vinculada a abordagens estruturais.	- Poucos trabalhos; - Representam o conteúdo disponibilizado pelos *sites Web;* - Trabalhos limitam-se a apresentar perspectivas da exploração da *Web* Semântica. - Enriquecimento semântico na Preparação limita a análise de padrões.	- Limitação quanto às restrições que definem a Ontologia de Domínio; e ao mapeamento das URLs para os conceitos da ontologia.

As limitações da abordagem proposta referem-se ao mapeamento das URLs para os conceitos da ontologia de domínio e às restrições quanto à definição da Ontologia de Domínio. As restrições que definem a estrutura da Ontologia de Domínio são necessárias para simplificar os mecanismos de interpretação e recuperação propostos neste trabalho, porém estes podem ser estendidos em trabalhos subseqüentes para abranger ontologias de domínio sem restrições quanto à estrutura.

Embora a representação gráfica seja utilizada para suportar a atividade de análise exploratória, não foram feitas comparações com abordagens de representação por não ser o foco desta pesquisa, uma vez que não temos o objetivo de propor técnicas de representação e estas são vistas como abordagens complementares à interpretação e recuperação de padrões.

6 MECANISMOS DE RECUPERAÇÃO DE PADRÕES DO USO DA *WEB*

Este capítulo apresenta os mecanismos voltados à recuperação de padrões. Estes possibilitam a geração de agrupamentos de padrões focando o escopo da busca; definição de filtros de interesse, utilizando a Ontologia de Domínio como apoio; e finalmente a definição de mecanismos de busca por padrões, envolvendo ou não medidas de similaridade.

A atividade de recuperação de padrões relevantes é facilitada quando o analista tem clareza dos objetivos que deseja atingir, por exemplo na verificação de hipóteses. Neste caso, abordagens de filtragem estruturais são úteis por reduzirem o foco da busca por padrões potencialmente relevantes.

A aplicação do processo de MUW freqüentemente inclui a descoberta exploratória por padrões interessantes, ou seja, o analista não tem idéia sobre o conhecimento que os padrões podem revelar, passando a analisá-los aleatoriamente. Muitas vezes deseja-se descobrir padrões inesperados, por exemplo, que contradizem as crenças de domínio. Neste caso, as abordagens de filtragem estatísticas relacionadas às medidas subjetivas (e.g [SIL96, COO03]) podem auxiliar os analistas neste propósito.

O que comumente acontece na fase de Análise de Padrões é a inspeção *ad hoc*, caracterizando a busca exaustiva e demorada por padrões interessantes em meio a tantos retornados pelas técnicas de Mineração de Dados. Assim, o analista interpreta cada padrão retornado sem seguir um critério de ordenação, buscando identificar os que potencialmente agregariam valor ao domínio da aplicação.

A abordagem proposta neste trabalho sugere alguns mecanismos para facilitar a atividade de recuperação de padrões. São eles:

- gerar agrupamentos de padrões conceituais base direcionando o foco da inspeção a conjunto de padrões relacionados de acordo com critérios previamente definidos;

- definir filtros de interesse com base na manipulação interativa dos conceitos da Ontologia de Domínio e na dimensão de interesse;
- selecionar mecanismo de busca que recupere padrões equivalentes ou aproximados ao interesse dos analistas especificados nos filtros de interesse.

As seções seguintes detalham os mecanismos propostos relacionados à recuperação de padrões.

6.1 Agrupamento de Padrões

Os padrões seqüenciais físicos retornados da fase de Mineração de Dados geralmente são ordenados pela medida de suporte, não considerando sua estrutura e nem o conteúdo. O agrupamento de padrões possibilita a geração de grupos de padrões, cada qual formado por um conjunto de padrões conceituais base agrupados de acordo com um critério previamente especificado.

Este mecanismo é adequado para situações em que o analista não possui clareza quanto aos objetivos que pretende atingir com o processo de MUW, e nem idéia do conhecimento que os padrões possam revelar. Desta forma, os agrupamentos facilitam o processo de inspeção *ad hoc* restringindo o foco da busca por padrões relevantes através dos grupos de padrões conceituais base.

Este mecanismo não reduz o número de padrões, apenas reorganiza-os visando facilitar a busca por padrões relevantes. Ao considerar alguns padrões irrelevantes para o domínio, possivelmente os padrões que pertencem ao mesmo grupo também serão irrelevantes por possuírem características em comum. Assim, a atividade de inspeção é otimizada uma vez que um grupo de padrões é desconsiderado. Da mesma forma, ao encontrar padrões relevantes para o domínio, é possível explorar os que pertencem ao mesmo grupo, sendo potencialmente interessantes.

Diferentes técnicas podem ser utilizadas para geração de agrupamentos baseadas em critérios como maximal [AGR94b], segmentação baseada em medida de distância [HAN00], entre outros. O critério maximal foi o escolhido para demonstração desta funcionalidade.

6.1.1 Critério Maximal

O critério maximal baseia-se nos padrões seqüenciais físicos que são maximais. Um padrão maximal [SRI95] é um padrão seqüencial físico que não é subseqüência de nenhum outro padrão, como definido na Seção 2.2.2.1. No conjunto de padrões seqüenciais físicos representado pela Figura 30, os padrões em negrito (1 e 2) são padrões maximais em relação aos demais padrões. Nota-se que o padrão maximal 1 contém todos os elementos do padrão maximal 2, porém as seqüências de URLs que os compõe difere.

```
1.  URL1 - URL3 - URL4 - URL5
2.  URL1 - URL4 - URL3 - URL5
3.  URL3 - URL3 - URL4
4.  URL4 - URL3
5.  URL4 - URL5
6.  URL1 - URL3 - URL4
```

Figura 30: Exemplo de Padrões Maximais

Um padrão seqüencial físico é denominado padrão contido em um padrão maximal quando este é uma subseqüência do padrão maximal. Por exemplo, o padrão seqüencial físico 5 ("URL4 - URL5") expressa que a URL4 deve ser seguida pela URL5. Este padrão é uma subseqüência do padrão maximal 1, pois a URL4 e URL5 fazem parte do padrão maximal, sendo que a URL4 é seguida pela URL5. Assim, o padrão seqüencial físico "URL4 - URL5" é um padrão contido no padrão maximal 1.

Na geração de agrupamentos de acordo com o critério maximal, cada agrupamento é composto por todos os padrões contidos em um padrão maximal, e o próprio padrão maximal. O número de agrupamentos resultantes é igual ao número de padrões maximais identificados. Por exemplo, com base nos padrões seqüenciais representados na Figura 30, apenas dois agrupamentos são gerados pois existem somente dois padrões maximais (Figura 31).

Cabe ressaltar que um padrão seqüencial físico pode estar contido em mais de um padrão maximal. Por exemplo, o padrão seqüencial físico 5 é um padrão contido em ambos os padrões maximais representados na Figura 30.

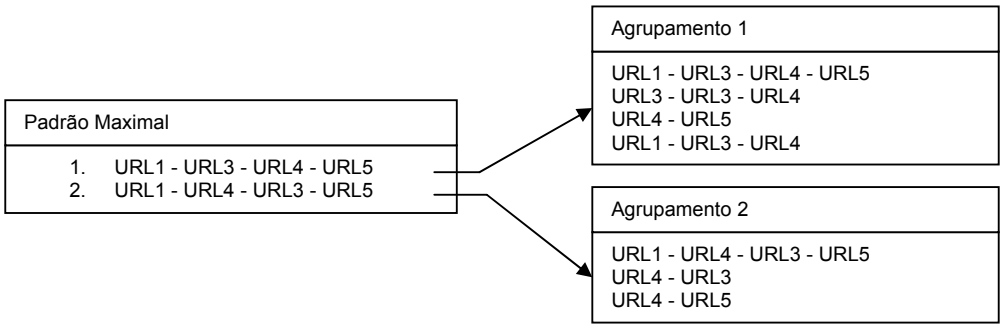


Figura 31: Agrupamentos de acordo com o critério maximal

A função responsável pela geração dos agrupamentos de acordo com o critério maximal recebe como entrada um conjunto de padrões seqüenciais físicos resultantes da fase de Mineração de Dados e retorna os agrupamentos. A função é constituída pelos seguintes passos:

1. identificar os padrões seqüenciais físicos que são maximais. Este passo é sugerido pela última fase do algoritmo *Aprioriall* [AGR94a] como descrito na Seção 2.2.2.1. Cada padrão seqüencial físico do conjunto de entrada é comparado com o restante dos padrões seqüenciais físicos e verificado se este está contido em algum outro padrão do conjunto. Se o padrão estiver contido em pelo menos um padrão seqüencial físico, este padrão não é um maximal. Caso contrário, ele é um padrão maximal.
2. criar os agrupamentos. Para cada padrão maximal identificado no passo anterior, são verificados quais padrões seqüenciais físicos são subseqüências deste. Desta forma, padrão maximal e os padrões nele contidos formam um agrupamento.

6.2 Filtros de Interesse baseados na Ontologia de Domínio

Filtros de interesse representam um conjunto de restrições que especificam as características que devem existir nos possíveis padrões conceituais base recuperados. Assim, a definição destes filtros permite restringir o conjunto de padrões na busca por aqueles que são relevantes.

Um filtro de interesse é definido de acordo com uma dimensão de interesse e é formado por um conjunto de elementos que definem as restrições. Cada elemento por sua vez

está ligado a um conceito da ontologia, bem como qualquer seqüência. A estrutura de um filtro de interesse é ilustrada no diagrama de classes UML apresentada na Figura 32.

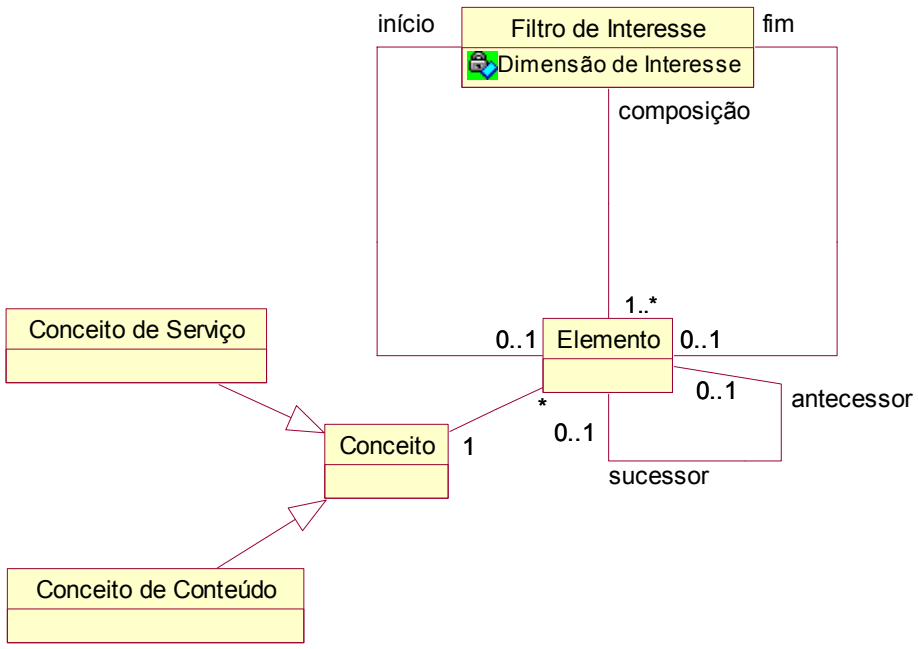


Figura 32: Estrutura de um filtro de interesse

Um filtro de interesse pode definir até três tipos de restrições: conceitual, estrutural e estatística.

Restrição conceitual refere-se aos conceitos de conteúdo ou serviço que compõem o filtro. Esta restrição é definida através da interação com a Ontologia de Domínio previamente disponível e visualizada graficamente. Padrões conceituais base que não violam a restrição conceitual devem ser formados por todos os conceitos especificados na restrição conceitual ou pelos seus conceitos descendentes.

Restrição estrutural define relações de ordem entre os elementos que compõem o filtro de interesse. Três tipos de restrições estruturais são consideradas:

- Restrição de Início: É definida entre um elemento início e um conceito da ontologia. O elemento início associado a um conceito significa que os padrões

conceituais base devem iniciar por aquele conceito ou um de seus descendentes (diretos ou por recursão).

- Restrição de Fim: É definida entre um elemento fim e um conceito da ontologia. O elemento fim associado a um conceito significa que os padrões conceituais base devem finalizar por aquele conceito ou um de seus descendentes (diretos ou por recursão).
- Restrição de Ordem: É definida entre um elemento antecessor e um elemento sucessor, que definem uma subseqüência. A associação entre um elemento antecessor e um elemento sucessor define que os padrões conceituais base devem respeitar a ordem (imediata ou não) entre os dois conceitos especificados ou seus descendentes (diretos ou por recursão).

Uma restrição estrutural pode ser formada pela combinação das restrições de início, fim e de ordem.

Restrição estatística refere-se a um limiar mínimo estabelecido para uma determinada medida estatística disponível e este limiar deve ser respeitado pelos padrões conceituais base recuperados.

Para facilitar a compreensão, são apresentados alguns exemplos de filtros de interesse e o resultado retornado da aplicação de cada um deles. Para isso, considera-se novamente o domínio do *site* turístico apresentado nos capítulos anteriores. Supõe-se que o analista deseja inspecionar os caminhos de navegação dos usuários que estão interessados no serviço de localizar informações, e no conteúdo sobre hotéis e restaurantes. A dimensão de interesse especificada pelo analista é a de serviço e conteúdo.

Primeiramente, é definida uma restrição conceitual uma vez que o analista tem interesse nos conceitos *Localizar*, *Hotel* e *Restaurante*. Para definir um filtro de interesse com esta restrição basta selecionar estes conceitos na Ontologia de Domínio representada graficamente e adicioná-los no filtro de interesse. A Figura 33 ilustra graficamente a Ontologia de Domínio, um filtro de interesse formado apenas por uma restrição conceitual e o conjunto de padrões conceituais base que poderiam ser retornados por este filtro. Nota-se que

todos os padrões conceituais base possuem o conceito *Localizar*, *Hotel,* e os conceitos descendentes de *Restaurante*.

Cabe ressaltar que a ordem com que os conceitos estão disponibilizados no filtro não corresponde à ordem que eles devem assumir nos possíveis padrões conceituais base recuperados. A ordem entre os conceitos é definida pelas restrições estruturais, não utilizadas neste exemplo.

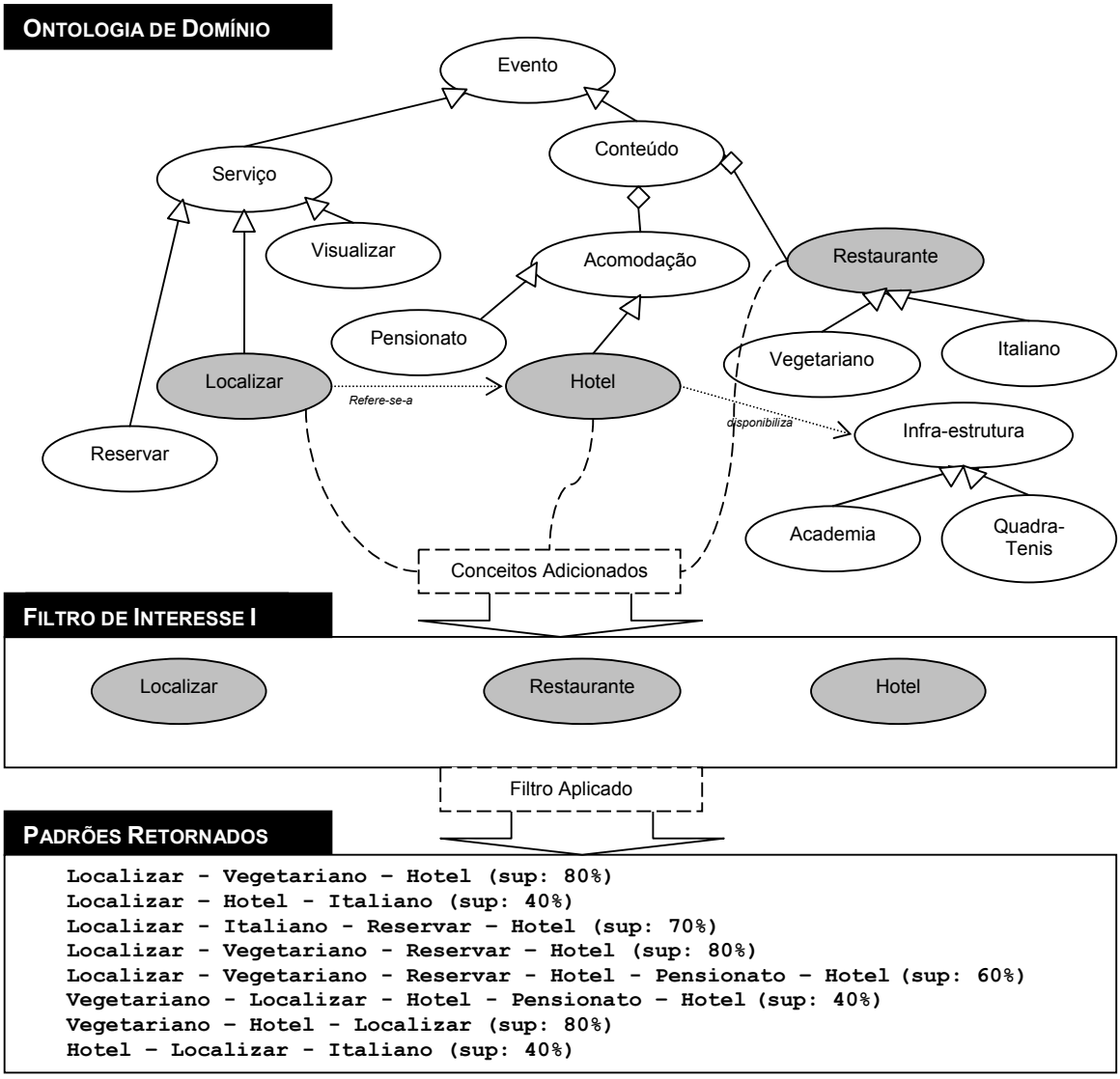


Figura 33: Filtro de Interesse composto por uma restrição conceitual

Posteriormente, suponha-se que o analista tenha interesse nos padrões que possuam os conceitos *Localizar*, *Hotel* e *Restaurante*, mas que iniciem com o conceito *Localizar*. Para isso, uma restrição estrutural de início é inserida no filtro de interesse da Figura 33, ou seja, um elemento de início é associado ao conceito *Localizar*, como representado pela Figura 34. Desta forma, filtro de interesse passa a ser formado por uma restrição conceitual e uma estrutural. Aplicando o filtro, nota-se que os padrões conceituais base retornados respeitam as restrições definidas.

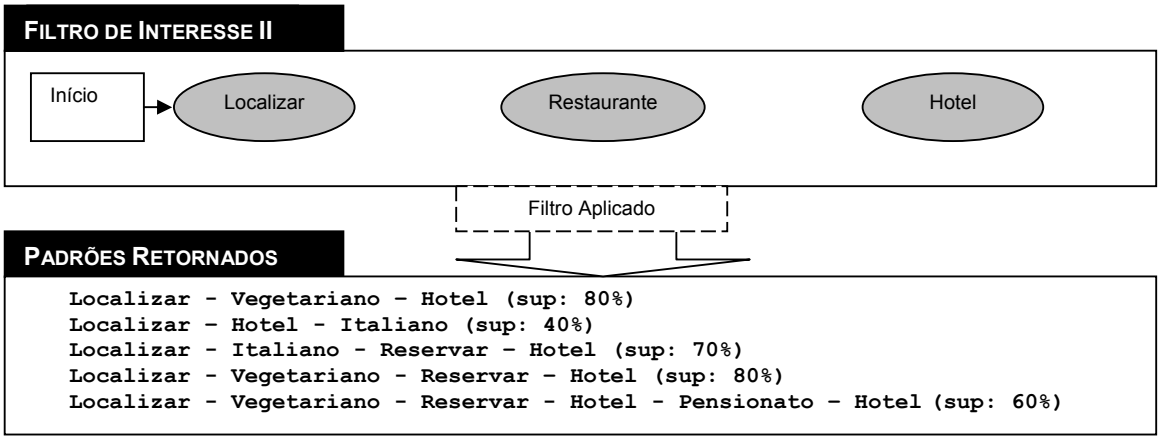


Figura 34: Filtro de Interesse composto por uma restrição conceitual e uma estrutural

Ainda, imagina-se que o analista deseje recuperar apenas os padrões conceituais base iniciados por *Localizar*, e nos quais o conceito *Restaurante* seja seguido pelo conceito *Hotel*. Desta forma, mais uma restrição estrutural é adicionada ao filtro de interesse, correspondendo a uma restrição de ordem, que associa o conceito antecedente *Restaurante* ao conceito sucessor *Hotel*, como representado na Figura 35.

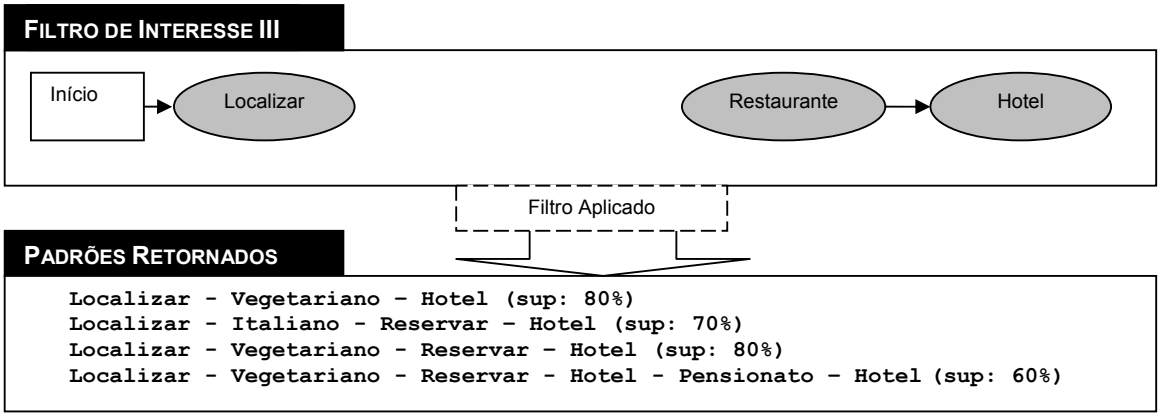


Figura 35: Filtro de Interesse composto por uma restrição conceitual e duas estruturais

Para complementar o filtro, o analista define o interesse por padrões que possuam suporte mínimo de 80% através de uma restrição estatística. A Figura 36 representa o filtro correspondente, composto por uma restrição conceitual, duas estruturais e uma estatística. Exemplos de padrões conceituais base retornados por ele também são ilustrados.

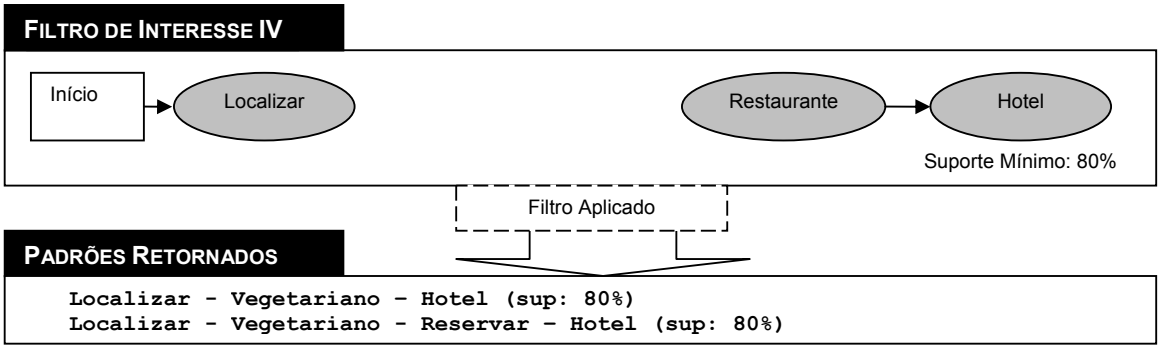


Figura 36: Filtro de Interesse composto por uma restrição conceitual, duas estruturais e uma estatística

Nota-se nestes sucessivos exemplos, que o conjunto de padrões recuperados diminui de acordo com as restrições adicionadas ao filtro de interesse.

A vantagem da utilização de filtros de interesse é restringir o foco da busca por padrões de acordo com as restrições definidas pelo analista. Este mecanismo torna-se extremamente útil quando o analista tem clareza das características que os padrões devem possuir para atingir os objetivos do processo de MUW. Outras vantagens propostas por esta

abordagem referem-se à forma com que os filtros de interesse são definidos, através da interatividade com a Ontologia de Domínio representada graficamente e com os demais elementos que compõem o filtro.

A aplicação de um filtro de interesse recupera padrões conceituais base de acordo com um mecanismo de busca. Neste trabalho, propõem-se dois mecanismos de busca denominados equivalente e aproximado. O mecanismo de busca equivalente recupera padrões que respeitam exatamente as restrições especificadas pelo filtro. O mecanismo de busca aproximado estende o mecanismo de busca equivalente, recuperando também os padrões que são similares ao filtro. Neste caso, para cada padrão recuperado é atribuído um valor de similaridade, que expressa o quão similar o padrão é do filtro de interesse definido. Estes mecanismos de busca são detalhados nas próximas seções.

6.2.1 Mecanismo de Busca Equivalente

Este mecanismo é representado por uma função que recebe como parâmetros de entrada: o filtro de interesse; a dimensão de interesse; a Ontologia de Domínio; o mapeamento; o valor de suporte mínimo (não obrigatório); e o conjunto de padrões seqüenciais físicos resultantes da fase de Mineração de Dados. O retorno desta função é um conjunto de padrões conceituais base que respeitam as restrições definidas pelo filtro de interesse.

Os principais passos que constituem esta função são:

1. recuperar os padrões conceituais base de acordo com a dimensão de interesse.
2. identificar os conceitos descendentes que implicitamente fazem parte do filtro de interesse. Para cada conceito que compõe o filtro, devem ser identificados quais são os conceitos descendentes na ontologia. Para cada conceito descendente, por sua vez, são verificados seus descendentes, e assim recursivamente. Do conjunto de descendentes extraído, eliminam-se todos os que não possuem um mapeamento para as URLs na dimensão de interesse especificada, uma vez que estes conceitos não poderiam compor os possíveis padrões conceituais base candidatos a serem retornados pelo filtro.

3. se existirem restrições estatísticas especificadas pelo filtro, aplicá-las sobre os padrões conceituais base. Considerando os padrões conceituais base candidatos, apenas os que possuem o valor a medida estatística dentro do limiar estabelecido são considerados.
4. aplicar a restrição conceitual especificada no filtro de interesse. Dos padrões conceituais base candidatos recuperados do passo anterior, apenas os que são formados por todos conceitos de serviço e conteúdo especificados no filtro, ou seus descendentes, são considerados.
5. se existirem restrições estruturais especificadas no filtro, aplicá-las. Dos padrões conceituais base candidatos resultantes do passo 4, apenas os que respeitam as restrições estruturais definidas são considerados, a saber restrições de início, de fim e/ou de ordem.
6. retornar os padrões conceituais base restantes.

Os exemplos apresentados na seção anterior utilizaram este mecanismo de busca para recuperar os padrões conceituais base. A dimensão de interesse considerada foi a de serviço e conteúdo.

6.2.2 Mecanismo de Busca Aproximada

O mecanismo de busca aproximada visa a recuperação de padrões semelhantes ao filtro de interesse. Para isso propõe-se a combinação de filtros e medidas de similaridade. As seções seguintes apresentam a medida de similaridade utilizada para exemplificar este mecanismo, e o algoritmo proposto para este mecanismo.

6.2.2.1 *Medidas de Similaridade*

A noção de similaridade é utilizada em muitos contextos para identificar objetos que possuem características semelhantes [GAN03]. Por exemplo, uma máquina de busca encontra documentos que são similares a uma consulta ou a outros documentos; algoritmos de segmentação agrupam seqüências de elementos que possuem características em comum [HAN00]. Já filtros colaborativos analisam usuários que compartilham interesses em comum [GOL92].

A medida de similaridade utilizada neste trabalho para exemplificar este mecanismo é baseada no modelo espaço vetorial generalizado (GVSM – *Generalized Vector Space Model*) proposto por Ganesan *et al*. [GAN03]. Neste modelo, a medida de similaridade entre dois conceitos é definida pela distância entre os conceitos numa hierarquia previamente definida, de acordo com a função $Sim(l_1 , l_2)$ cuja fórmula é representada na Figura 37. Os conceitos são representados por l_1 e l_2. A função LCA (*Lowest Common Ancestor*) representa o antecedente comum mais próximo de ambos os conceitos. A função *depth()* representa a distância do conceito até o nodo raiz da hierarquia de conceitos.

$$Sim(l_1 , l_2) \quad = \quad \frac{2*depth\ (\mathrm{LCA}(l_1, l_2))}{depth\ (l_1) + depth\ (l_2)}$$

Figura 37: Similaridade entre dois conceitos definida pela função $Sim(l_1 , l_2)$

A função $Sim(l_1 , l_2)$ retorna um valor entre 0 e 1. Quanto mais próximo do valor de 0, menor é o grau de similaridade entre os objetos. O valor será 1 quando os objetos forem iguais. A Tabela 7 ilustra as medidas de similaridade calculadas entre diferentes conceitos da ontologia representada na Figura 33, de acordo com o GVSM. Observa-se que o conceito *Italiano* possui o valor de similaridade maior que o conceito *Hotel* em relação ao conceito *Vegetariano*. Cabe ressaltar que a aplicação do cálculo de similaridade considera somente as relações hierárquicas entre os conceitos.

Tabela 7. Medidas de similaridade entre conceitos

l_1	l_2	LCA(l_1, l_2)	Deph (LCA(l_1, l_2))	Valor de Similaridade
Hotel	*Vegetariano*	*Conteúdo*	1	0,33
Italiano	*Vegetariano*	*Restaurante*	2	0,66

6.2.2.2 *Similaridade de um padrão conceitual base em relação ao filtro*

Com base na medida de similaridade apresentada, a similaridade de um padrão conceitual base em relação ao filtro de interesse muitas vezes requer vários cálculos de similaridade realizados sobre conceitos específicos, denominado neste trabalho de cálculo de

similaridade pontual. A média aritmética dos valores de similaridade obtidos pelos cálculos de similaridade pontual determina o grau de similaridade em relação ao filtro de interesse.

Restrições estruturais definidas no filtro de interesse são consideradas para selecionar os conceitos que serão utilizados no cálculo de similaridade. Por exemplo, se uma restrição estrutural do filtro define que o padrão conceitual base deve iniciar pelo conceito *Hotel*, o cálculo de similaridade é aplicado entre o conceito *Hotel* do filtro, com o primeiro conceito do padrão conceitual base. Cabe ressaltar que para um único filtro, o número de cálculos de similaridade realizados por padrão conceitual base é dependente do número de restrições estruturais definidas no filtro de interesse.

A Figura 38 apresenta um exemplo que enfoca o cálculo da similaridade de um padrão conceitual em relação ao filtro de interesse. Para o cálculo de similaridade, considera-se a Ontologia de Domínio representada pela Figura 19. De acordo com as restrições estruturais do filtro, são realizados dois cálculos de similaridade pontual: um aplicado sobre o primeiro conceito do padrão conceitual base abstrato e o outro sobre o último conceito do padrão. Os valores obtidos por estes cálculos estão representados na Figura 38. A média aritmética destes valores é 0,75, constituindo o valor de similaridade do padrão em relação ao filtro de interesse.

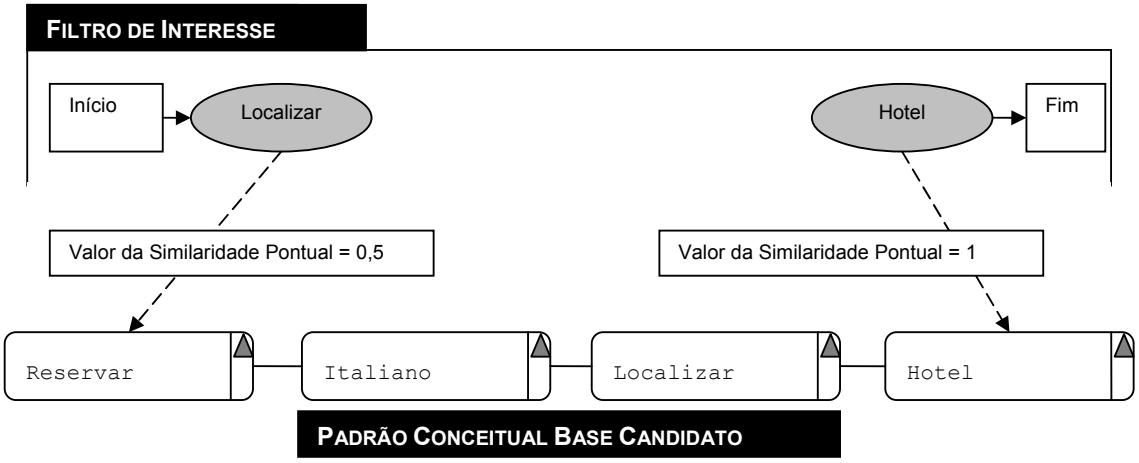


Figura 38: Medida de similaridade pontual – Restrição estrutural de início e fim

Restrições estruturais de ordem possuem algumas particularidades quanto à aplicação do cálculo de similaridade pontual uma vez que deve ser considerado o valor de similaridade

da seqüência. O cálculo de similaridade de uma seqüência é definido pela média aritmética dos valores de similaridade pontual entre os conceitos da seqüência e os do filtro. A Figura 39 ilustra o cálculo de similaridade de uma seqüência em relação ao filtro de interesse.

Cabe ressaltar que uma seqüência de conceitos no filtro de interesse pode corresponder a mais de uma seqüência em um padrão conceitual base, como ilustrado na Figura 40. Nesta situação, o cálculo de similaridade é aplicado para todas as seqüências possíveis, porém somente a seqüência que possui maior valor de similaridade é considerada para o cálculo de similaridade de um padrão.

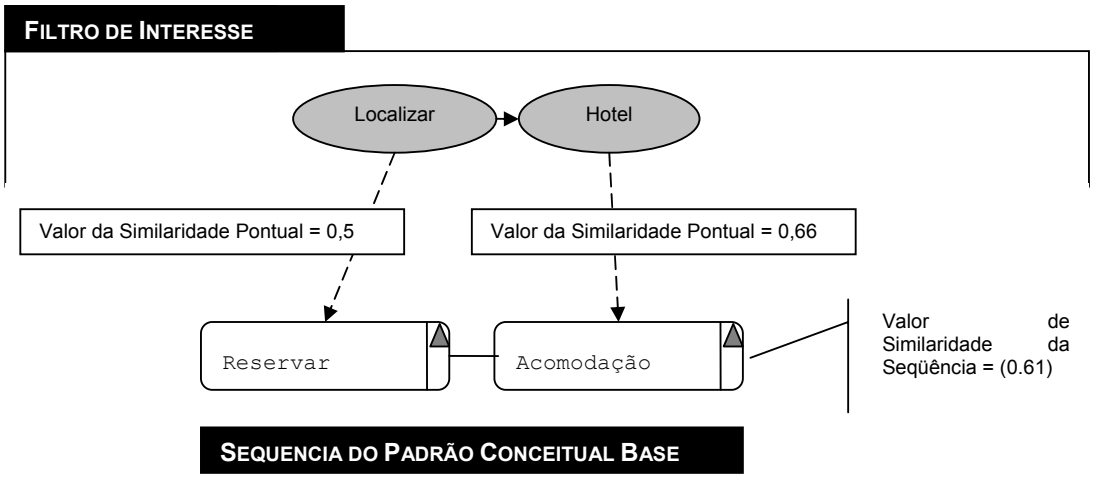


Figura 39: Valor de Similaridade de uma seqüência do padrão conceitual base

Nota-se na Figura 40 que a seqüência definida pelo filtro corresponde a duas seqüências (S1 e S2) no padrão conceitual base candidato. Desta forma, o cálculo de similaridade é realizado para as duas seqüências. Os valores de similaridade obtidos estão na Tabela 8. Apenas o maior valor de similaridade é considerado para o cálculo de similaridade do padrão conceitual base em relação ao filtro, ou seja, o valor de similaridade da seqüência S2. Como neste caso não há outras restrições estruturais definidas pelo filtro, o valor de similaridade do padrão é igual ao valor de similaridade da seqüência S2 .

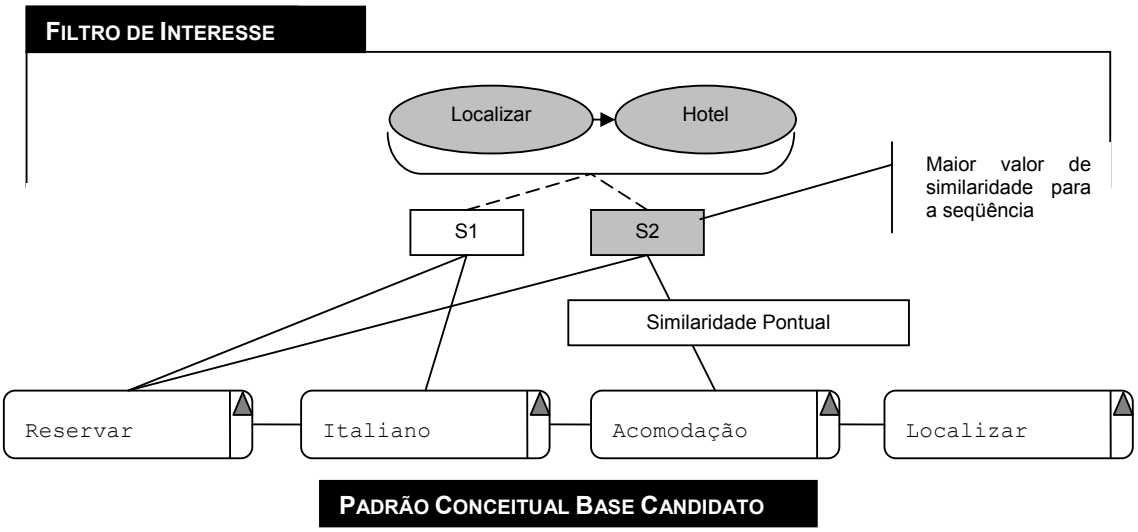


Figura 40: Medida de similaridade pontual – Restrição estrutural de ordem

Tabela 8. Medidas de similaridade nas seqüências

Seq	Primeiro Elemento Seqüência	Primeiro Elemento Filtro	VSP	Segundo Elemento Seqüência	Segundo Elemento Filtro	VSP	VS Seqüência
S1	*Reservar*	*Localizar*	0,5	*Italiano*	*Hotel*	0,33	0,42
S2	*Reservar*	*Localizar*	0,5	*Acomodação*	*Hotel*	0,66	**0,61**

Legenda:
VSP – Valor de Similaridade Pontual;
VS Seqüência – Valor de similaridade da seqüência;

A função que determina o grau de similaridade entre um padrão conceitual base e um filtro de interesse recebe como parâmetro de entrada o filtro de interesse e padrão conceitual base. Os passos principais desta função são:

- Verificar sobre quais conceitos o cálculo de similaridade deve ser aplicado. Para isso, é necessário verificar as restrições existentes:
    - Restrição de Início: O cálculo de similaridade pontual é aplicado entre o primeiro conceito do Padrão e o primeiro conceito do filtro.
    - Restrição de Fim: O cálculo de similaridade pontual é aplicado entre o último conceito do Padrão e o último conceito do filtro.

- Restrição de Ordem: O cálculo de similaridade pontual é aplicado nos conceitos da seqüência. Ou seja, no primeiro conceito da seqüência, com o primeiro conceito da restrição de ordem, e assim por diante.

- Aplicar o cálculo de similaridade entre dois conceitos (Seção 6.2.2.1)
- Se existir restrição de ordem, calcular o valor de similaridade das seqüências possíveis e selecionar a seqüência com maior valor de similaridade.
- Calcular o valor de similaridade do padrão pela média aritmética dos valores de similaridade pontual (restrição de início e fim) e dos valores de similaridade de seqüência.

6.3 Combinação de Filtros e Medidas de Similaridade

O mecanismo de busca aproximada é uma extensão do mecanismo de busca equivalente. Para representar como este é utilizado e definido, considera-se a medida de similaridade GVSM, descrita na Seção 6.2.2.1. Outras medidas de similaridade também poderiam ser aplicadas.

A função responsável por este mecanismo recebe como parâmetro de entrada: os parâmetros necessários para o mecanismo equivalente especificados anteiormente; a medida de similaridade selecionada; o valor de similaridade mínimo (não obrigatório); o nível de abrangência. Este último é utilizado para a criação de um filtro generalizado definido a partir do filtro de interesse e do nível de abrangência.

A busca aproximada é baseada no conceito filtro generalizado, que é uma extensão de um filtro de interesse, mas considera também descendentes dos conceitos ascendentes aos que formam o filtro de interesse, de acordo com um nível de abrangência especificado pelo usuário. O nível de abrangência define o quão distante um conceito ascendente pode estar, na hierarquia, dos conceitos que compõem o filtro de interesse.

A Figura 41 representa um filtro de interesse definido por um especialista, a ser aplicado utilizando o método de busca aproximada. O nível de abrangência definido foi 1. O filtro generalizado obtido mantém todas as restrições definidas pelo filtro de interesse, com exceção da restrição de conteúdo, onde os conceitos são considerados a partir dos seus

ascendentes de acordo com o nível de abrangência. Por exemplo, o conceito *Hotel* passa a ser interpretado pelo conceito ascendente *Restaurante* por este estar distante de 1 nível na relação de hierarquia.

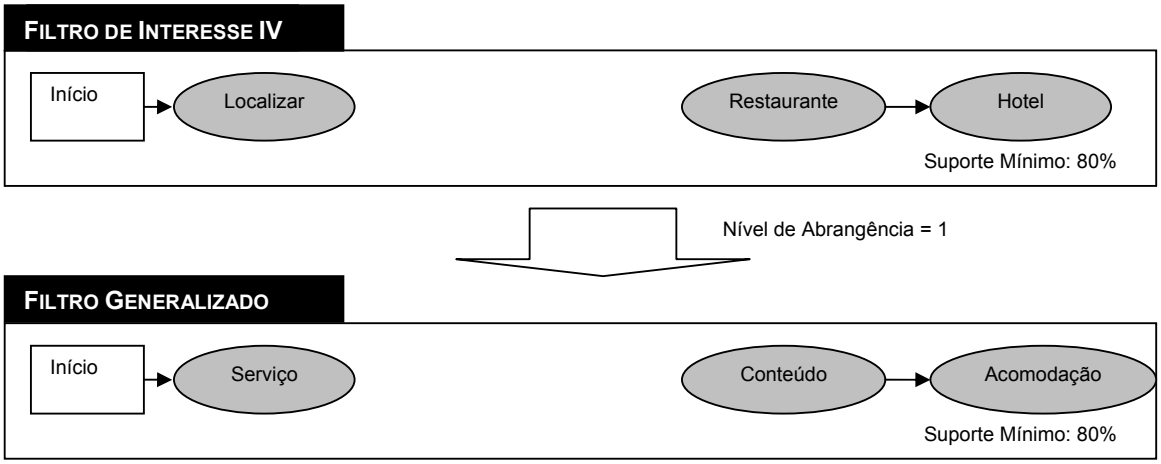


Figura 41: Filtro de Interesse e Filtro Generalizado

O retorno da função responsável pelo mecanismo de busca aproximada constituiu um conjunto de padrões conceituais base, juntamente com o seu respectivo suporte e o valor de similaridade. Neste contexto, o valor de similaridade representa o quão similar é um padrão conceitual base do filtro de interesse definido pelo analista.

Os principais passos que constituem esta função são:

1. criar o filtro generalizado com base no filtro de interesse. São identificados os conceitos ascendentes que fazem parte do filtro generalizado. Para cada conceito de serviço e conteúdo que compõe o filtro de interesse, são identificados todos os conceitos ascendentes que estão de acordo com o nível de abrangência especificado pelo parâmetro de entrada. Estes conceitos ascendentes passam a compor o filtro generalizado. O restante das restrições especificadas pelo filtro de interesse se mantém no filtro generalizado.
2. encontrar os padrões conceituais base candidatos que estão de acordo com o filtro generalizado. Executar todos os passos especificados pelo mecanismo de busca equivalente, considerando o filtro generalizado como entrada ao invés do filtro de interesse definido pelo analista.

3. calcular o valor de similaridade dos padrões conceituais base retornados pelo passo 2 em relação ao filtro de interesse especificado pelo analista. Para cada padrão conceitual base, é calculado o valor de similaridade do padrão em relação ao filtro de interesse.
4. se existir um valor mínimo de similaridade especificado pelo analista, apenas os padrões conceituais base que possuem o valor de similaridade maior ou igual ao limiar estabelecido são considerados.
5. retornar os padrões conceituais base, juntamente com o valor do suporte e de similaridade.

6.4 Considerações

Neste capítulo foram apresentados os mecanismos de recuperação de padrões. O mecanismo de recuperação de padrões através da utilização de agrupamentos facilita e otimiza o processo de inspeção *ad hoc*, restringindo o foco da busca por padrões relevantes nos grupos de padrões conceituais base.

O mecanismo de recuperação através da utilização de filtros de interesse apresenta diversas vantagens em relação aos trabalhos propostos no Capítulo 3, desde a forma como eles são definidos até os mecanismos de busca. As vantagens referem-se:

- à definição visual do filtro de interesse, minimizando a necessidade de aprendizado de uma sintaxe;
- à definição do filtro baseada nos conceitos da ontologia. Desta forma, o analista não necessita ser um especialista no domínio ou no *site*, podendo utilizar o conhecimento do domínio para a formulação de hipóteses, ou definição de áreas de interesse de forma facilitada;
- à riqueza dos filtros de interesse, por permitirem a definição das restrições conceitual, estrutural e estatística;
- à definição de filtros considerando diferentes dimensões de interesse, com base na representação conceitual dos padrões seqüenciais físicos. Desta forma, é

possível explorar dinamicamente diferentes dimensões sem retorno à fase de Preparação de Dados;

- à definição de filtros considerando diferentes níveis de abstração. O analista pode utilizar conceitos em diferentes níveis de abstração na definição de um filtro de interesse;
- ao poder dos mecanismos de busca, permitindo a recuperação de padrões equivalentes ou aproximados ao interesse do analista.

A Tabela 9 apresenta um comparativo entre os mecanismos de recuperação propostos por este trabalho e as abordagens de filtragem pesquisadas na literatura e detalhadas no Capítulo 3, Seção 3.1.

Tabela 9. Comparação da abordagem proposta X abordagens de filtragem pesquisadas

	Abordagens de Filtragem		
	Filtros Estatísticos	**Filtros Estruturais**	**Abordagem de Vanzin**
Contribuição	Recuperar Padrões		Recuperar Padrões interativamente
Objetivo	- Definição de filtros que envolvam restrições estatísticas. - Mecanismo de busca recupera padrões de acordo com o limiar estabelecido para as medidas objetivas e subjetivas;	- Definição de filtros que envolvam restrições estruturais e de conteúdo. - Especificação de filtros através do uso de taxonomias e sintaxes; - Mecanismo de busca recupera padrões equivalentes ao interesse do analista;	- Geração de agrupamentos de padrões conceituais base de acordo com critérios previamente definidos; - Definição de filtros de interesse com base na manipulação interativa dos conceitos da Ontologia de Domínio; - Mecanismos de busca recuperam padrões equivalentes ou aproximados ao interesse do analista;
Fase Mineração	Mineração de Dados Análise de Padrões	Mineração de Dados Análise de Padrões	Análise de Padrões.
Desvantagens	- Medidas objetivas não necessariamente determinam um padrão interessante; - Medidas Subjetivas: Dificuldade em expressar o conhecimento do domínio e limitações dos algoritmos de comparação.	- Domínio de uma linguagem para especificação dos filtros; - Necessidade de objetivos claros para a definição dos filtros; - Necessidade de profundo conhecimento do domínio; - Não representam qualquer tipo de expressão regular; - Re-execução da fase de Mineração de Dados para cada novo objetivo.	- Filtros não representam qualquer tipo de expressão regular. - Limitação nas medidas de similaridade e critérios de agrupamento considerados.

7 AMBIENTE DE APOIO À INTERPRETAÇÃO E RECUPERAÇÃO DE PADRÕES DO USO DA *WEB*

Este capítulo descreve o ambiente de apoio proposto para avaliar a utilidade dos mecanismos de interpretação e recuperação de padrões durante a fase de Análise de Padrões.

A principal contribuição deste trabalho refere-se aos mecanismos de interpretação e recuperação de padrões descritos nos capítulos anteriores para apoio à fase de Análise de Padrões. Neste capítulo é proposto um ambiente de apoio que disponibiliza estes mecanismos através das funcionalidades de um protótipo. O protótipo foi desenvolvido utilizando a linguagem de programação Java e o banco de dados *Microsoft Access*.

As funcionalidades do protótipo estão representadas no diagrama de casos de uso UML da Figura 42. O ator principal corresponde ao analista que fará uso do ambiente de apoio durante a análise dos padrões. A Tabela 10 e Tabela 11 descrevem brevemente as funcionalidades oferecidas pelo protótipo.

Tabela 10. Funcionalidades para definições

Nome do Caso de Uso	Descrição
Importar Padrões Seqüenciais	O analista importa os padrões do arquivo de padrões. Para isso o analista define os parâmetros para a importação do arquivo de padrões. São eles: arquivo de padrões, parâmetros de formatação.
Definir Dimensão de Interesse	O analista define a dimensão de interesse na qual os padrões importados serão analisados.
Definir Critério de Agrupamento	O analista define o critério pelo qual os padrões serão agrupados.
Preparar Padrões para Interpretação e Recuperação	O analista prepara os padrões seqüenciais físicos para a interpretação e recuperação. Isso inclui a geração dos agrupamentos de acordo com o critério especificado e a criação dos padrões seqüenciais conceituais de acordo com a dimensão de interesse.

Tabela 11. Funcionalidades para recuperação e interpretação de padrões

Nome do Caso de Uso	Descrição
Inspecionar Agrupamentos	O analista inspeciona os agrupamentos, verificando os padrões que os compõem.
Visualizar Padrão Conceitual	O analista visualiza um padrão conceitual.
Visualizar Padrão Conceitual Textual	O analista visualiza um padrão seqüencial na forma textual.
Visualizar Padrão Conceitual Gráfico	O analista visualiza um padrão seqüencial em uma representação gráfica.
Selecionar Padrão Conceitual Textual	O analista seleciona um padrão conceitual na forma textual para que ele possa ser visualizado graficamente para análise exploratória.
Configurar Análise Exploratória	O analista configura alguns parâmetros para realizar a análise exploratória sobre um padrão conceitual em especifico. São eles: a dimensão de interesse; e detalhamentos das relações hierárquicas para todos os conceitos ou somente para o conceito selecionado.
Verificar Detalhamento de Relacionamentos	O analista visualiza detalhes sobre o relacionamento dos conceitos que compõem um padrão conceitual em relação a outros conceitos da Ontologia de Domínio. Este detalhamento refere-se às relações hierárquicas e de propriedade.
Executar Operação *Roll-up*	O analista executa a operação de *roll-up* gerando um padrão conceitual abstrato.
Executar Operação *Drill-down*	O analista requisita os padrões detalhe sumarizados pelo padrão conceitual abstrato.
Interagir com a Ontologia de Domínio.	O analista visualiza e interage com a Ontologia de Domínio.
Definir Filtro de Interesse	O analista define filtros de interesse: - selecionando conceitos da ontologia e adicionando-os na área de definição de filtros; - adicionando elementos de início e fim; - conectando conceitos; e ainda definido um suporte mínimo.
Aplicar Mecanismo de Busca	O analista define o mecanismo de busca que é utilizado para recuperar os padrões.
Buscar por Equivalência	O analista recupera os padrões pelo mecanismo de busca equivalente.
Buscar por Aproximação	O analista recupera os padrões pelo mecanismo de busca aproximada. Parâmetros são informados: Medida de similaridade; Nível de Abrangência; Valor de similaridade mínimo.
Verificar Padrões Conceituais Base Recuperados	O analista verifica os padrões conceituais base retornados de diferentes operações. São eles: Padrões Contidos; Padrões Detalhe; Padrões Filtrados.
Verificar Padrões Contidos	O analista verifica os padrões contidos nos agrupamentos.
Verificar Padrões Detalhe	O analista verifica os padrões detalhe que sumarizam um padrão abstrato
Verificar Padrões filtrados	O analista verifica os padrões recuperados através dos filtros de interesse e mecanismos de busca definidos.

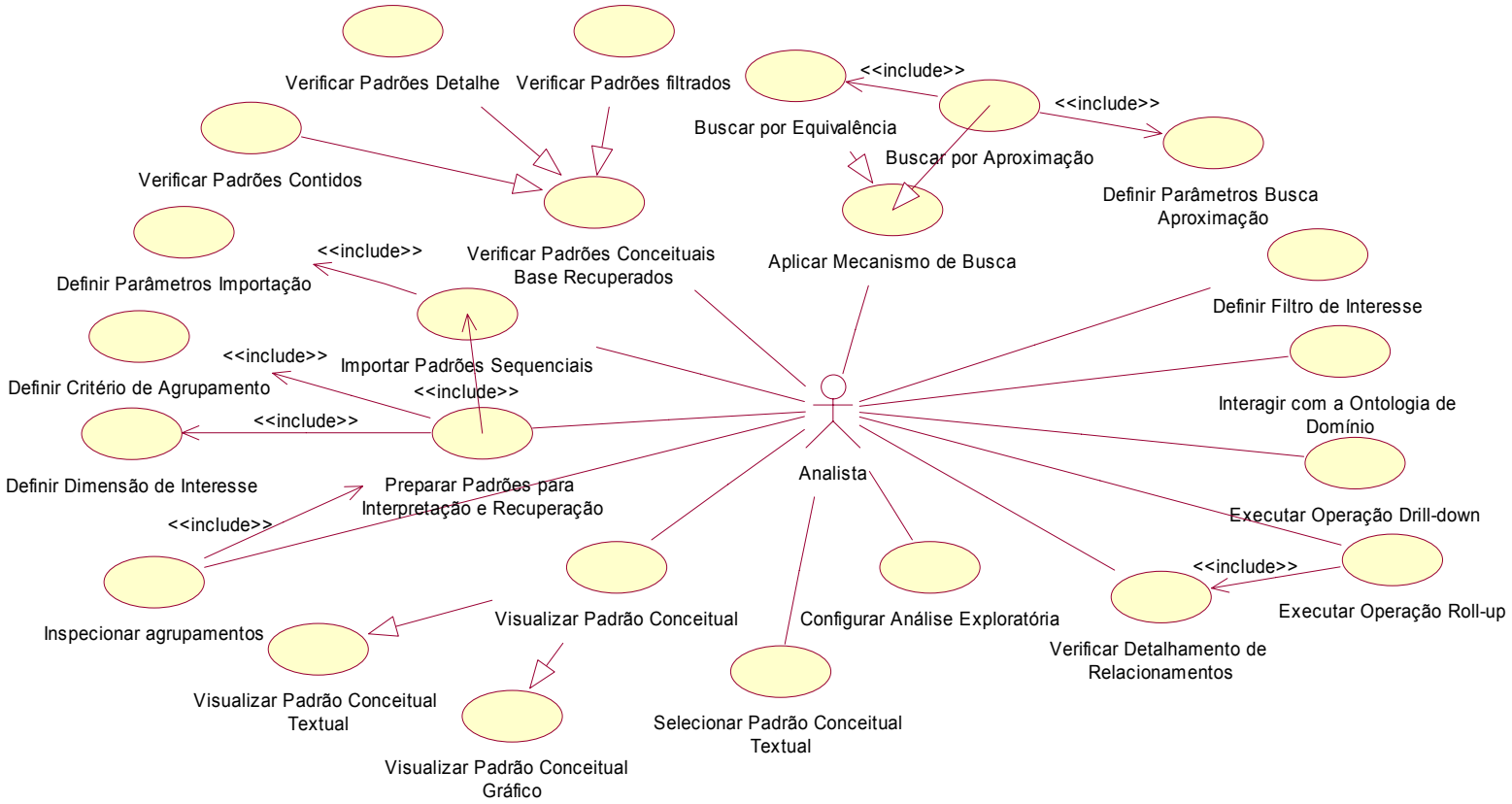


Figura 42: Diagrama de Casos de Uso do Protótipo

As funcionalidades descritas são disponibilizadas em diferentes áreas no protótipo, e suportadas por diferentes elementos da arquitetura.

7.1 Arquitetura do Protótipo

A Figura 43 contextualiza o protótipo no processo de MUW, caracterizando as entradas necessárias para sua utilização e as saídas geradas. As entradas são:

- *Log* pré-processado resultante da fase de Preparação de Dados;
- Conjunto de padrões seqüenciais físicos resultantes da aplicação do algoritmo *AprioriAll* durante a fase de Mineração de Dados;
- Ontologia de Domínio que descreve os principais eventos do domínio em termos de conteúdo e serviços disponibilizados pelo *site*;
- Mapeamento das URLs para os conceitos de serviço e conteúdo definidos pela Ontologia de Domínio.

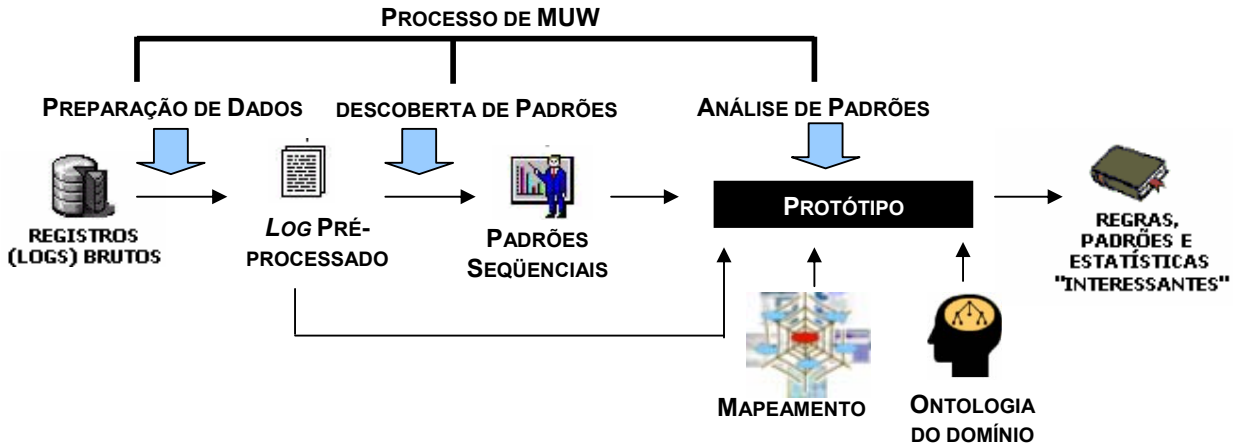


Figura 43: Ambiente de Apoio e suas entradas e saída

A arquitetura do protótipo, ilustrada na Figura 44, representa como estas entradas e os demais elementos estão estruturados para atender as funcionalidades propostas. A arquitetura é composta por uma base de dados e por conjuntos de módulos. Estes elementos suportam as funcionalidades disponibilizadas em diferentes áreas da interface do protótipo. A seguir, estas áreas são descritas em detalhe, assim como a base de dados que compõe a arquitetura.

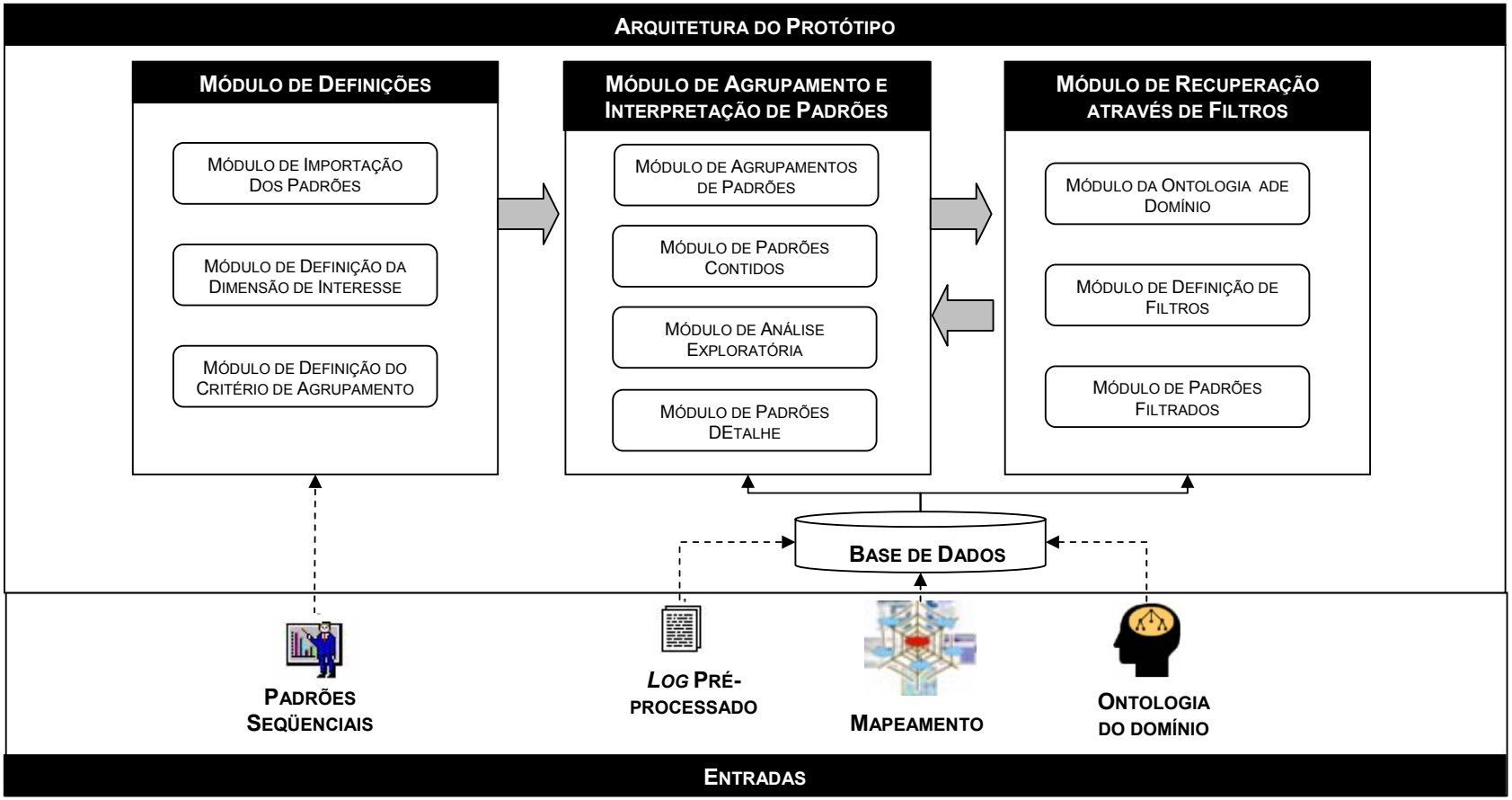


Figura 44: Arquitetura do Protótipo e suas entradas

7.1.1 Base de Dados

A base de dados armazena informações do *log* pré-processado, da Ontologia de Domínio e do mapeamento. O esquema da base de dados está representado na Figura 45.

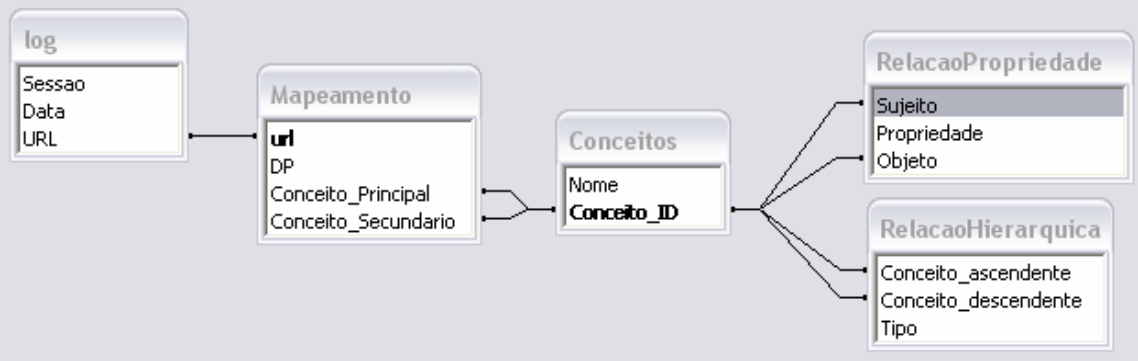


Figura 45: Esquema da base de dados

7.1.1.1 Log *pré-processado*

O *log* pré-processado resulta da fase de Preparação de Dados. Os dados pré-processados devem incluir pelo menos informações sobre o identificador da sessão do usuário, *time stamp* de acesso à página e finalmente a URL da página. Cada registro corresponde a um acesso, sendo que os registros devem ser ordenados considerando o identificador de sessão e *time stamp*. No ambiente proposto, os dados que compõem o *log* preparado são armazenados na tabela *Log* da base de dados. A Figura 46 representa um exemplo de um conjunto de dados extraído de um *log* pré-processado.

sessao	data	URL
Sessao1	26/10/2004 18:30:00	/webct/homearea/homearea
Sessao1	26/10/2004 18:32:30	/SCRIPT/Curso/scripts/serve_home
Sessao1	26/10/2004 18:40:12	/SCRIPT/Curso/scripts/student/serve_home?_homepage+START
Sessao1	26/10/2004 18:41:07	/SCRIPT/Curso/scripts/student/serve_home?1010412547+view
Sessao1	26/10/2004 18:41:53	/SCRIPT/Curso/scripts/student/serve_calendar?START+homepage+1010412547
Sessao2	27/10/2004 17:59:23	/webct/homearea/homearea
Sessao2	27/10/2004 18:00:23	/SCRIPT/Curso/scripts/serve_home
Sessao2	27/10/2004 18:02:15	/SCRIPT/Curso/scripts/student/serve_page.pl?1010418606+Material_Apoio/redes.htm
Sessao2	27/10/2004 18:04:15	/SCRIPT/Curso/scripts/student/serve_resume_session?__TOP__

Figura 46: Exemplo de dados extraídos de um *log* pré-processado

7.1.1.2 *Ontologia de Domínio*

Os conceitos e relações que compõem a Ontologia de Domínio são armazenados nas tabelas da base de dados: *Conceitos*, *RelacaoPropriedade*, *RelacaoHierarquica*. A tabela

Conceitos armazena todos os conceitos que compõem a ontologia, atribuindo a eles um nome e um identificador. A tabela *RelacaoPropriedade* contém as informações das relações de propriedade existentes entre os conceitos da ontologia especificados na tabela *Conceitos*. Já a tabela *RelacaoHierarquica* armazena as relações de hierarquia entre os conceitos. Nesta última, o atributo *Tipo* identifica a relação entre o conceito descendente e ascendente, podendo ser de generalização ou agregação.

7.1.1.3 *Mapeamento*

O mapeamento das URLs para os conceitos da ontologia é especificado na tabela *Mapeamento* (Figura 45). Nela, cada URL é mapeada para conceitos de serviço e/ou conteúdo. O atributo *DP* indica qual é a dimensão predominante para a qual URL está sendo mapeada, em termos de serviço e conteúdo. O atributo *Conceito_Principal* armazena o conceito de serviço ou conteúdo associado a dimensão predominante (atributo *DP*) representado pela URL, e o *Conceito_Secundário*, o conceito de serviço ou conteúdo da dimensão secundária, se existir.

A Tabela 12 ilustra o mapeamento de duas URLs para os conceitos da ontologia na tabela *Mapeamento*. A URL especificada na primeira linha é mapeada para o conceito *Calendário*, na dimensão de conteúdo, sendo esta dimensão predominante. Na dimensão de serviço, a mesma URL é mapeada para o conceito *Visualizar*. A URL representada na segunda linha, é mapeada somente para o conceito *Correio Eletrônico* na dimensão de serviço, que é a predominante.

Tabela 12. Mapeamento das URLs para conceitos da Ontologia

URL	Dimensão predominante	Conceito Principal	Conceito Secundário
/SCRIPT/Cursos/scripts/student/serve_calendar?START+homepage+1010412547	Conteúdo	Calendário	Visualizar
/SCRIPT/CCD_16_02JAN/scripts/student/serve_mail?LIST+All	Serviço	Correio Eletrônico	

7.1.2 Módulo de Definições

O Módulo de Definições suporta as funcionalidades de: Importar Padrões Seqüenciais, Definir Dimensão de Interesse, Definir Critério de Agrupamento, Preparar Padrões para Interpretação e Recuperação. Estas funcionalidades são disponibilizadas no protótipo pela:

área de Importação dos Padrões (Figura 47-A), área de Definição da Dimensão de Interesse (Figura 47-B) e área de Definição do Critério de Agrupamento (Figura 47-B). A Figura 47 permite visualizar como estas áreas estão organizadas na interface do protótipo que implementa este módulo.

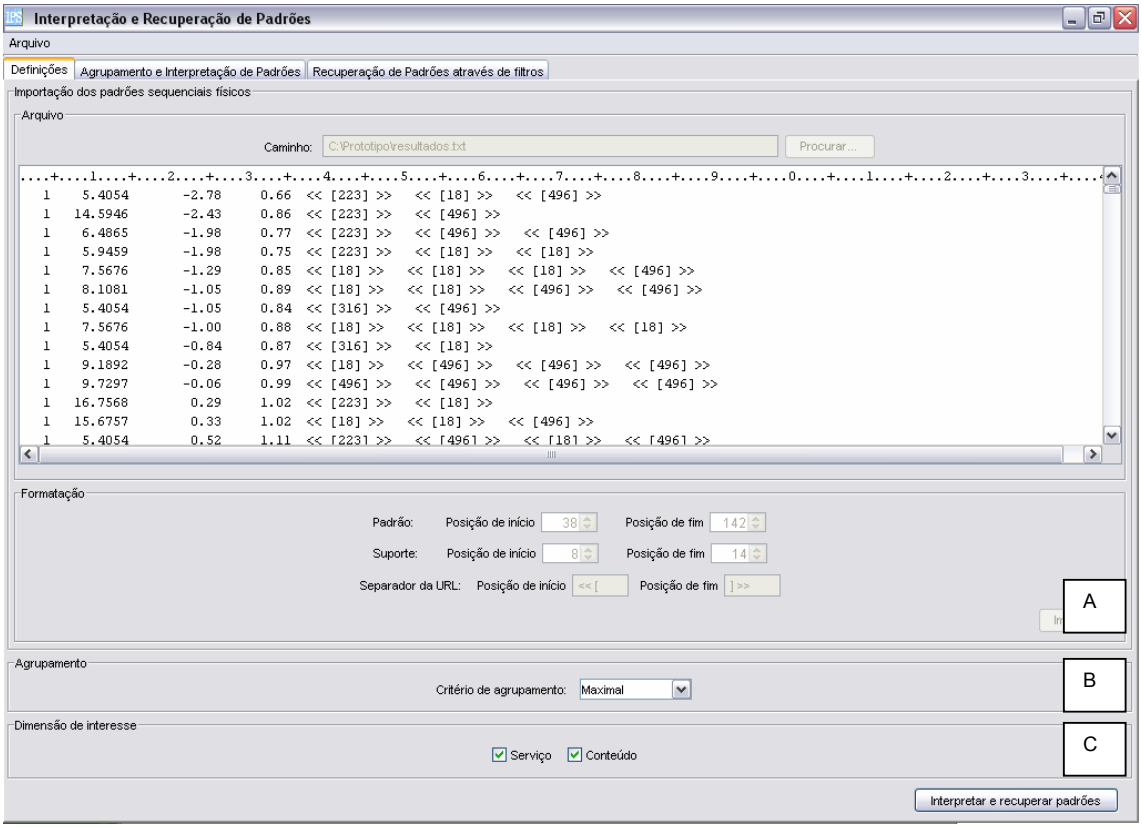


Figura 47: Interface do Módulo de Definições

Cada área requisita um conjunto de parâmetros. Depois de definidos estes parâmetros e importados os padrões seqüenciais físicos, os módulos restantes são habilitados, permitindo a interpretação e recuperação dos padrões. Para isso, basta clicar no botão "*Intepretar e Recuperar padrões"*, localizado no canto inferior da interface do Módulo de Definições.

7.1.2.1 *Área de Importação dos Padrões*

Esta área permite selecionar o arquivo texto que contém as informações sobre os padrões seqüenciais; visualizar o arquivo texto; especificar os delimitadores das informações; e finalmente importar os padrões seqüenciais físicos.

a) Arquivo de Padrões

Assume-se como entrada um arquivo texto contendo um conjunto de padrões seqüenciais físicos, com o respectivo valor de suporte. Cada linha do arquivo texto define um padrão seqüencial físico. Cada padrão possui no mínimo um valor de suporte, e as URLs que compõem cada padrão devem estar separadas por um caracter qualquer. Não existe uma posição determinada no arquivo texto para o armazenamento das informações. Isso não impede que outros algoritmos sejam utilizados para gerar padrões seqüenciais, desde que, as restrições quanto à formatação deste arquivo sejam obedecidas.

A Figura 48 representa um arquivo texto que contém alguns padrões gerados pelo algoritmo *AprioriAll*, tal como implementado na ferramenta *Intelligent Miner*. Neste exemplo, as URLs que compõem o padrão seqüencial foram substituídas por uma abreviação para facilitar a visualização dos padrões seqüenciais físicos. A segunda coluna especifica os valores de suporte e a última os padrões seqüenciais, onde as URLs são separadas pelos delimitadores "[" e "]". Outras informações existentes no arquivo texto referente ao padrão seqüencial são desconsideradas por serem irrelevantes para este trabalho.

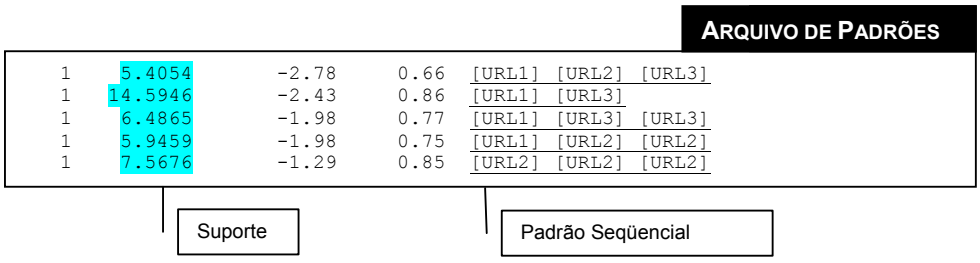


Figura 48: Exemplo de um conjunto de padrões seqüenciais

b) Seleção do Arquivo de Padrões

A Figura 49 representa a área do protótipo responsável pela importação dos padrões seqüenciais físicos, organizada em duas subáreas. Primeiramente, o protótipo permite selecionar um arquivo texto. Este deve conter as informações mínimas referentes aos padrões seqüenciais físicos, a saber, valor de suporte e URLs que o compõem. Uma vez selecionado o arquivo pelo analista, esta subárea permite a visualização do arquivo em uma pequena tela, sendo que a primeira linha representa uma régua que determina as posições que os caracteres

assumem ao longo do arquivo texto, como visualizado na Figura 49-A. Esta régua serve como auxílio na definição de outros parâmetros requisitados.

b) Formatação do Arquivo de Padrões

Após a seleção do arquivo, a subárea inferior (Figura 49-B) permite especificar as informações referentes à formatação do arquivo texto. Os dois primeiros campos determinam a posição de início e fim dos padrões seqüenciais contidos no arquivo texto. Os dois campos seguintes delimitam a posição de início e fim que armazenam o valor de suporte especificado para cada padrão seqüencial físico. Já, os últimos campos desta subárea especificam os caracteres utilizados para diferenciar as URLs que formam o padrão seqüencial. Depois de informados todos os parâmetros, finalmente os padrões seqüenciais físicos podem ser importados, acionando o botão *Importar*.

Figura 49: Área de Importação dos Padrões

Cabe ressaltar que atualmente o protótipo disponibiliza apenas a importação dos padrões seqüenciais físicos, assumindo que os demais dados (e.g. Ontologia de Domínio, mapeamento e *log* pré-processado) são inseridos usando diretamente os recursos disponíveis para tal no sistema de gerência de banco de dados. Versões futuras do protótipo estenderão as funcionalidades para a importação de todos dados.

7.1.2.2 *Área de definição do Critério de Agrupamento*

Esta área permite definir o critério segundo o qual os padrões importados serão agrupados. Nesta versão do protótipo, apenas o algoritmo maximal está disponível. Porém, o protótipo prevê a extensão para outros algoritmos de agrupamento. A Figura 50 representa esta área no protótipo.

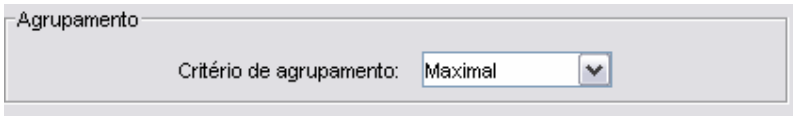


Figura 50: Área de definição do Critério de Agrupamento

7.1.2.3 *Área de Definição da Dimensão de Interesse*

A área de Definição da Dimensão de Interesse permite selecionar a dimensão segundo a qual os padrões seqüenciais físicos importados serão interpretados e recuperados, ou seja, a ela é utilizada na geração dos padrões conceituais base. A Figura 51 ilustra esta área no protótipo.

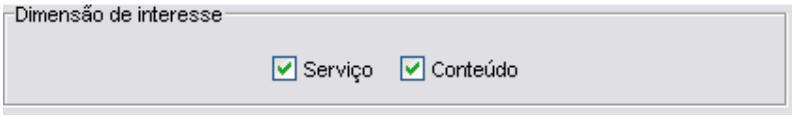


Figura 51: Área de Definição da Dimensão de Interesse

7.1.3 Módulo de Agrupamento e Interpretação de Padrões

O Módulo de Agrupamento e Interpretação de Padrões suporta as funcionalidades de: Inspecionar Agrupamentos, Visualizar Padrão Conceitual, Visualizar Padrão Conceitual Textual, Visualizar Padrão Conceitual Gráfico, Selecionar Padrão Conceitual Textual, Configurar Análise Exploratória, Executar Operação *Roll-up*, Executar Operação *Drill-down*, Verificar Padrões Contidos, Verificar Padrões Detalhe. Estas funcionalidades são

disponibilizadas no protótipo pela: área de Agrupamento de Padrões (Figura 52-A), área de Padrões Contidos (Figura 52-B), área de Análise Exploratória (Figura 52-C), e área de Padrões Detalhe (Figura 52-D). A Figura 52 representa como estas áreas estão organizadas na interface do protótipo que representa este módulo. Cabe ressaltar que as áreas dos Padrões Contidos e dos Padrões Detalhe são visualizadas a partir de uma requisição do analista.

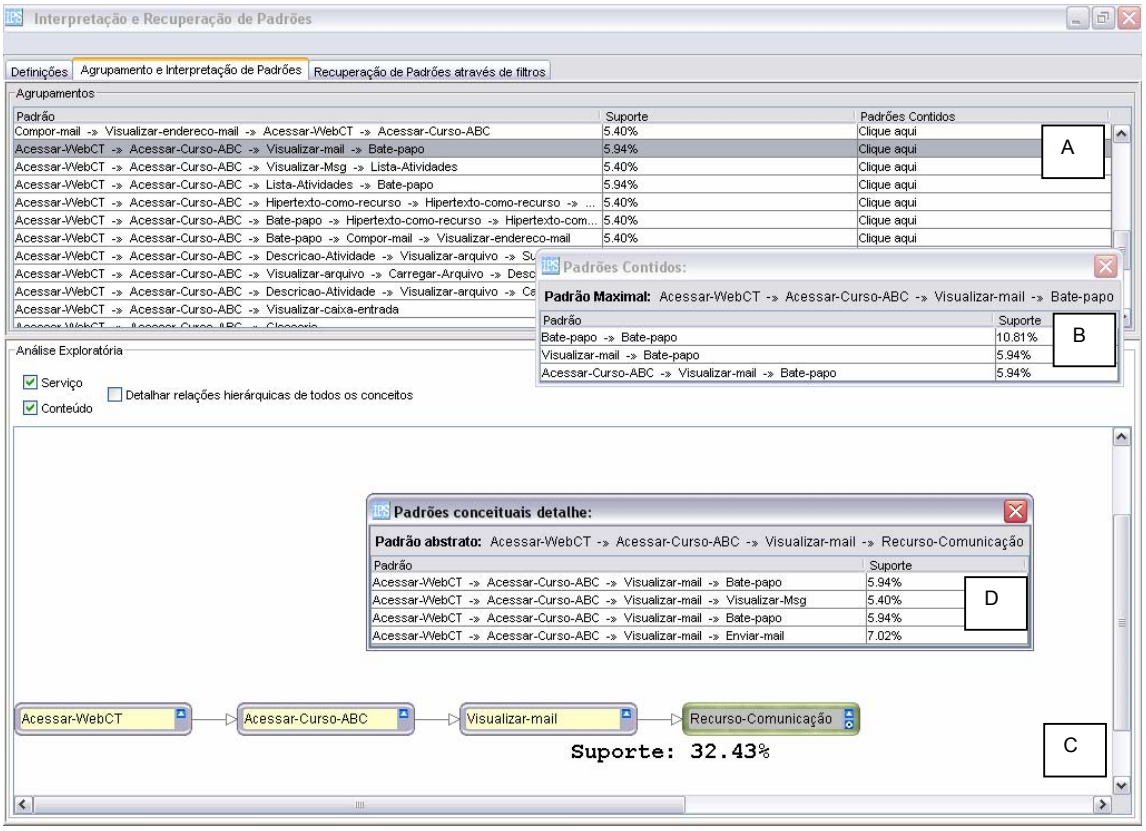


Figura 52: Interface do Módulo de Agrupamento e Interpretação de Padrões

7.1.3.1 *Área de Agrupamentos de Padrões*

A área de Agrupamento de Padrões identifica grupos de padrões conceituais base gerados pelo algoritmo de agrupamento selecionado na área de Definição de Critério de Agrupamento. Cada linha corresponde a um agrupamento e o padrão que identifica o agrupamento é representado textualmente. Considerando o critério maximal, os grupos são identificados pelo padrão maximal e seu respectivo suporte. Ao selecionar uma linha, o padrão maximal é representado graficamente na área de Análise Exploratória.

A partir desta área, o protótipo permite visualizar os padrões contidos nestes agrupamentos na área de Padrões Contidos. A Figura 53-A representa alguns grupos de padrões gerados de acordo com o critério maximal.

7.1.3.2 *Área de Padrões Contidos*

Esta área apresenta os padrões contidos no agrupamento selecionado na área de Agrupamentos de Padrões. Cada linha representa um padrão conceitual base visualizado textualmente, com o seu respectivo valor de suporte. Qualquer padrão desta área pode ser selecionado e visualizado na área de Análise Exploratória. A Figura 53-B representa alguns padrões que fazem parte do agrupamento selecionado.

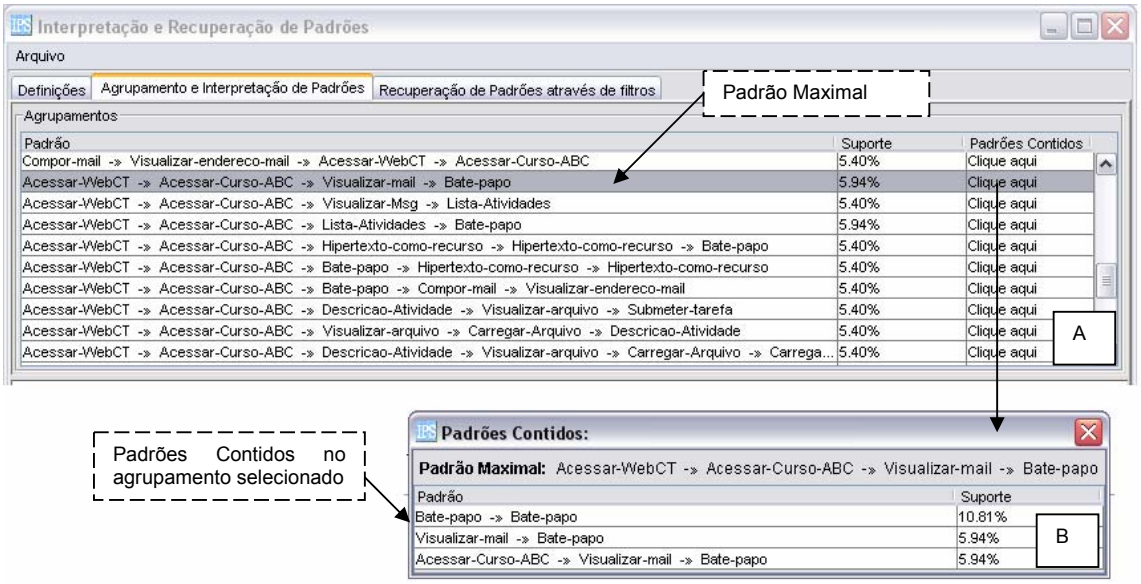


Figura 53: Áreas de Agrupamento de Padrões e Padrões Contidos

7.1.3.3 *Área de Análise Exploratória*

A área de Análise Exploratória permite: visualizar um padrão conceitual textual selecionado de qualquer área (Agrupamentos de Padrões, Padrões Contidos, Padrões Detalhe, Padrões Filtrados) através de uma representação gráfica e interativa; executar a análise exploratória sobre este padrão, compreendendo as operações de detalhamento de relacionamentos, *roll-up* e *drill-down*. Ainda, permite mudar a dimensão de interesse para interpretar este padrão conceitual específico.

A Figura 54-B representa a área de Interpretação de Padrões, onde visualiza-se um padrão conceitual base selecionado na área de Agrupamentos de Padrões (Figura 54-A), com o respectivo valor de suporte abaixo. No canto superior esquerdo da área de Análise Exploratória existem algumas opções que podem ser alteradas pelo analista para análise daquele padrão selecionado em específico. No presente exemplo, o padrão conceitual base selecionado está sendo analisado segundo a dimensão de Serviço e Conteúdo, uma vez que as duas opções estão selecionadas. Caso o analista queira analisar o mesmo padrão considerando apenas a dimensão de Serviço, basta desmarcar a opção de Conteúdo.

Figura 54: Áreas de Agrupamentos de Padrões e Análise Exploratória

A outra opção de "*Detalhar relações hierárquicas de todos os conceitos*" diz respeito à operação de detalhamento de relacionamentos hierárquicos. Esta opção, quando habilitada, permite executar a operação de detalhamento de relações hierárquicas sobre todos os conceitos ao mesmo tempo, através de uma única interação.

Na área de Análise Exploratória, os conceitos do padrão conceitual são representados por retângulos com bordas arredondadas. Os conceitos que formam um padrão conceitual base são representado por uma cor clara. Já os conceitos ascendentes são representados por uma cor escura. A Figura 55 representa uma visão ampliada de um padrão conceitual base sendo interpretado na área de Análise Exploratória.

a) Detalhamento de Relacionamentos

Observa-se na Figura 55-A que alguns conceitos possuem uma seta no canto superior direito apontando para cima. Esta seta significa que a operação de detalhamento de relacionamento hierárquico está habilitada. Clicando sobre esta seta, o conceito ascendente ao selecionado será representado logo acima, assim como a relação existente entre os dois.

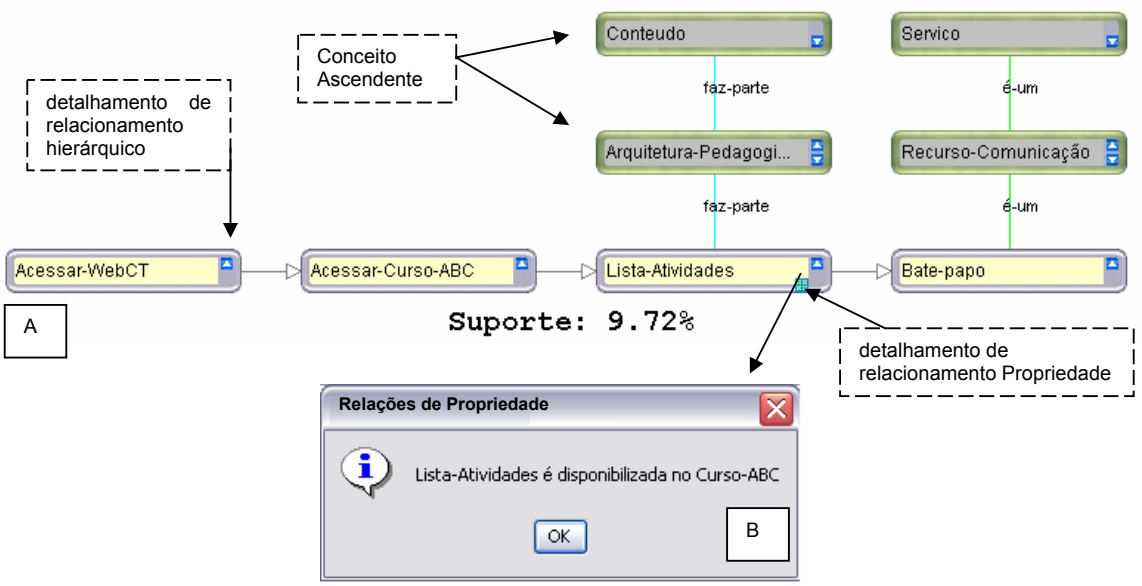


Figura 55: Explorando um padrão conceitual base

Os conceitos ascendentes possuem uma seta no canto inferior direito apontando para baixo. Ao clicar sobre esta seta, o conceito simplesmente desaparece.

Os conceitos que possuem um quadrado no canto inferior direito, expressam a existência de relacionamentos de propriedade com outros conceitos definidos pela Ontologia de Domínio. Para visualizar as informações sobre estas propriedades basta clicar sobre o quadrado que uma caixa de texto será visualizada, como representado pela Figura 55-B.

b) Operação *Roll-up*

A operação *roll-up* é realizada quando o analista executar um *duplo-click* sobre um dos conceitos ascendentes obtidos pela operação de detalhamento de relações hierárquicas (Figura 56-A). Assim, cria-se um padrão conceitual abstrato (Figura 56-B), com o respectivo suporte. Figura 56 representa a criação de um padrão conceitual abstrato a partir de uma operação de *roll-up*.

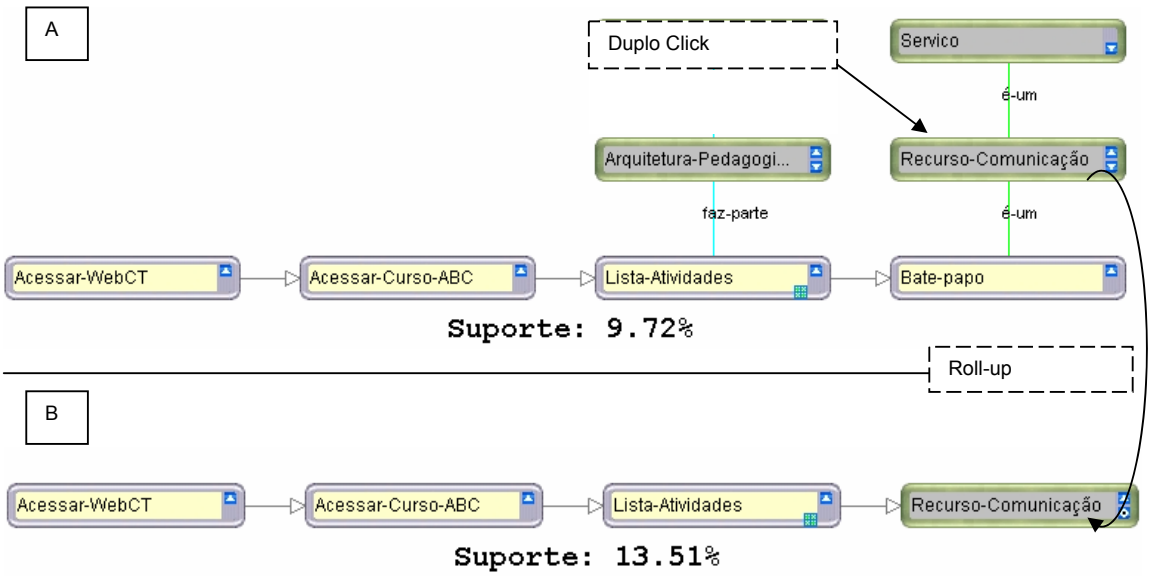


Figura 56: Operação *roll-up*

c) Operação *Drill-Down*

A operação *drill-down*, pode ser executada sobre qualquer um dos conceitos ascendentes de um padrão conceitual abstrato que possui descendentes. No padrão conceitual abstrato da Figura 56-A, a operação *drill-down* poderia ser executada somente sobre o último conceito do padrão. O que determina a operação *drill-down* estar habilitada é um círculo posicionado no canto inferior direito. Ao clicar sobre este símbolo, uma janela é criada, contendo os padrões detalhe sumarizados pelo padrão conceitual abstrato, como ilustrado pela Figura 60. Os padrões detalhe são visualizados na área de Padrões Detalhe (Figura 60-B).

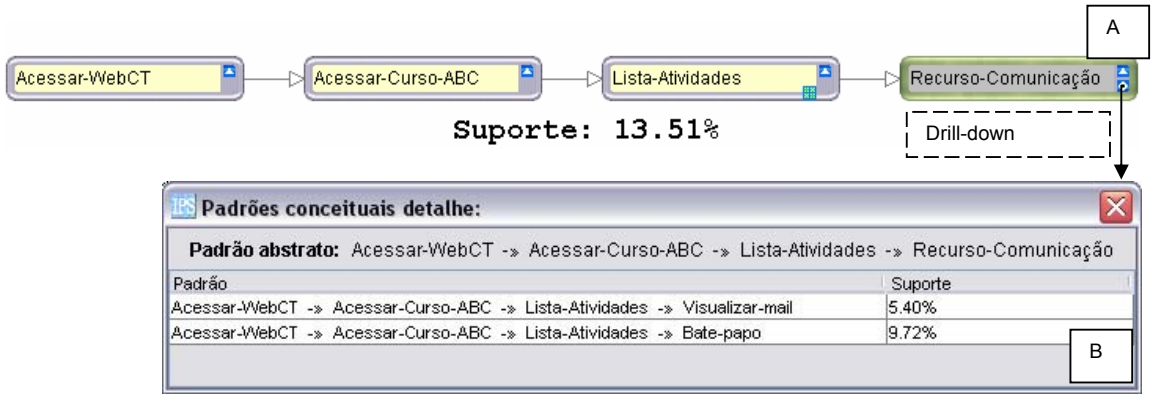


Figura 57: Operação *drill-down*

7.1.3.4 *Área de Padrões Detalhe*

A área de Padrões Detalhe representa um conjunto de padrões conceituais base que sumarizam um padrão abstrato. Ela é obtida a partir de uma interação do analista na área de Análise Exploratória (operação *drill-down*), como representado na Figura 60-A. Qualquer padrão desta área pode ser selecionado e visualizado na área de Análise Exploratória.

7.1.4 Módulo de Recuperação através de Filtros

O Módulo de Recuperação através de Filtros suporta as funcionalidades de: Visualizar Padrão Conceitual, Visualizar Padrão Conceitual Textual, Verificar Detalhamento de Relacionamentos, Interagir com a Ontologia de Domínio, Definir Filtro de Interesse, Aplicar Mecanismo de Busca, Buscar por Equivalência, Buscar por Aproximação, Verificar Padrões Conceituais Base Recuperados, Verificar Padrões filtrados. Estas funcionalidades são disponibilizadas no protótipo pela: área da Ontologia de Domínio (Figura 58-A), área de Definição de Filtros (Figura 58-B) e área de Padrões Filtrados (Figura 58-C). A Figura 58 representa como estas áreas estão organizadas na interface deste módulo.

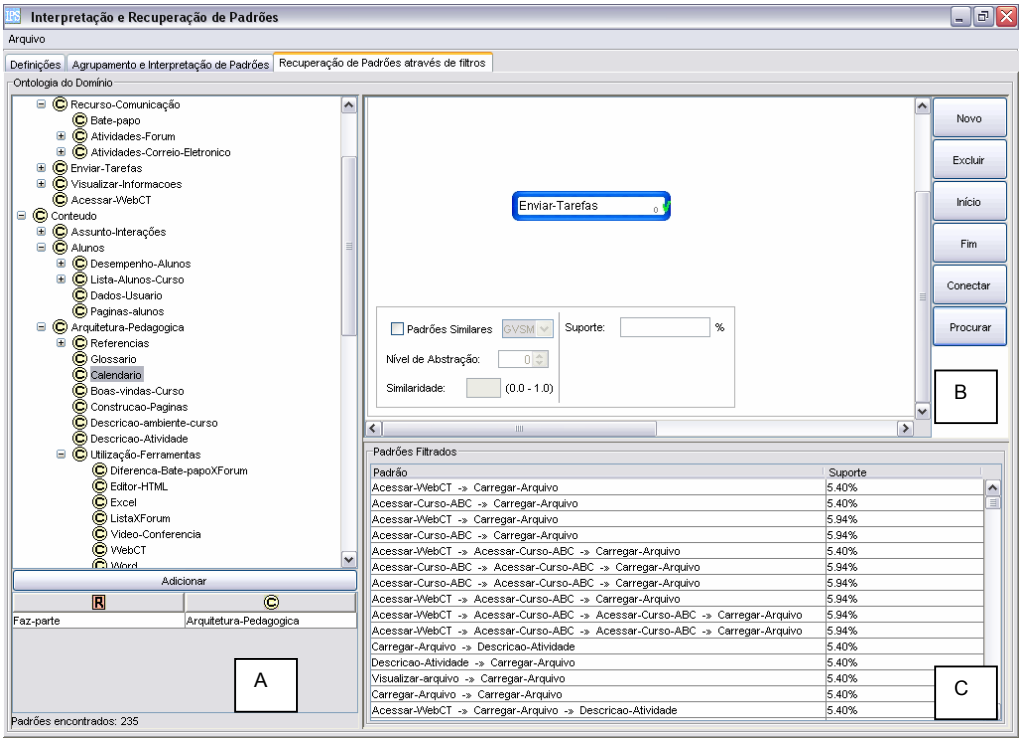


Figura 58: Interface do Módulo de Recuperação através de Filtros

7.1.4.1 *Área da Ontologia de Domínio*

A área da Ontologia de Domínio apresenta a ontologia graficamente. O analista tem a possibilidade de inspecionar os conceitos definidos, assim como as relações existentes entre eles. A Figura 59 ilustra uma Ontologia de Domínio representada graficamente. A região inferior descreve as relações do conceito selecionado com outros conceitos da ontologia.

Além de aprofundar o conhecimento sobre os principais conceitos que descrevem um domínio e suas relações, a área de Representação da Ontologia de Domínio é utilizada para auxiliar na definição da restrição conceitual que compõe um filtro. Desta forma, o analista interage e seleciona os conceitos da ontologia que expressam seu interesse pelos padrões conceituais base. O botão *Adicionar* adiciona o conceito da ontologia selecionado na área de Definição de Filtros.

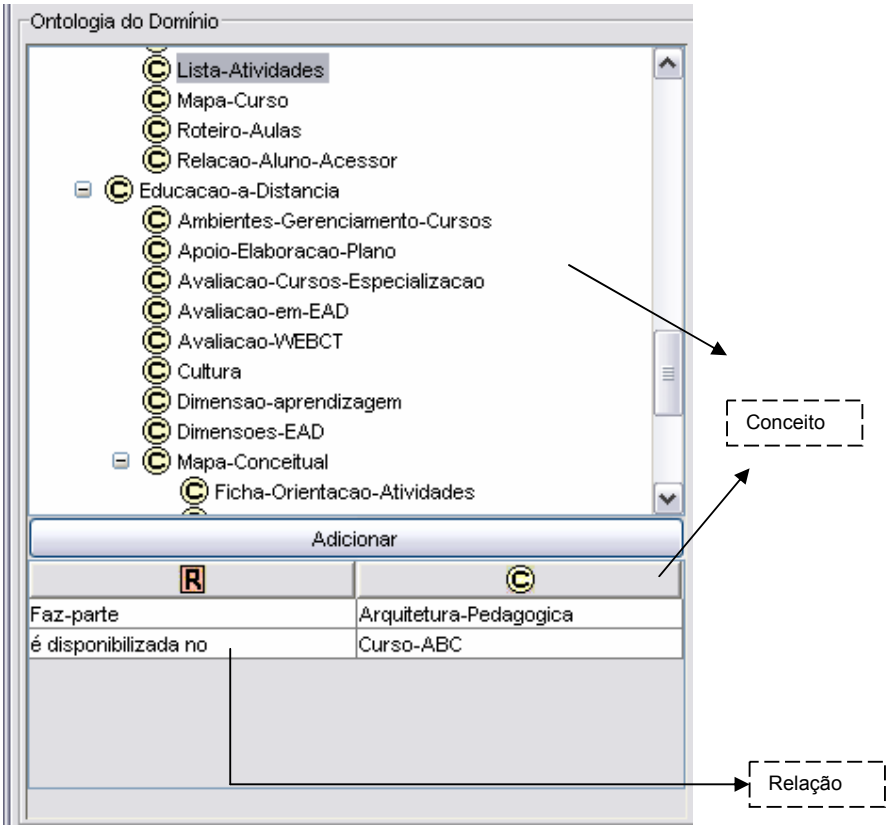


Figura 59: Área de Representação da Ontologia de Domínio.

7.1.4.2 *Área de Definição de Filtros*

Esta área permite definir filtros de interesse e escolher mecanismos de busca para recuperar os padrões de acordo com a dimensão de interesse especificada na área de definições. Filtros são criados de forma interativa expressando restrições conceituais, estruturais e estatísticas. As restrições conceituais são facilitadas pela interação com a Ontologia de Domínio representada graficamente. Os conceitos da ontologia adicionados na área de Definição de Filtros representam o interesse em determinados eventos de domínio.

Restrições estruturais são definidas através do uso de identificadores especiais. Estes identificadores são representados por um conjunto de botões localizados à direita na área de Definição de Filtros, como representado na Figura 60. Existe o identificador de início, fim e de conexão entre conceitos.

Como restrições estatísticas, o protótipo considera apenas um valor mínimo para o suporte dos padrões recuperados.

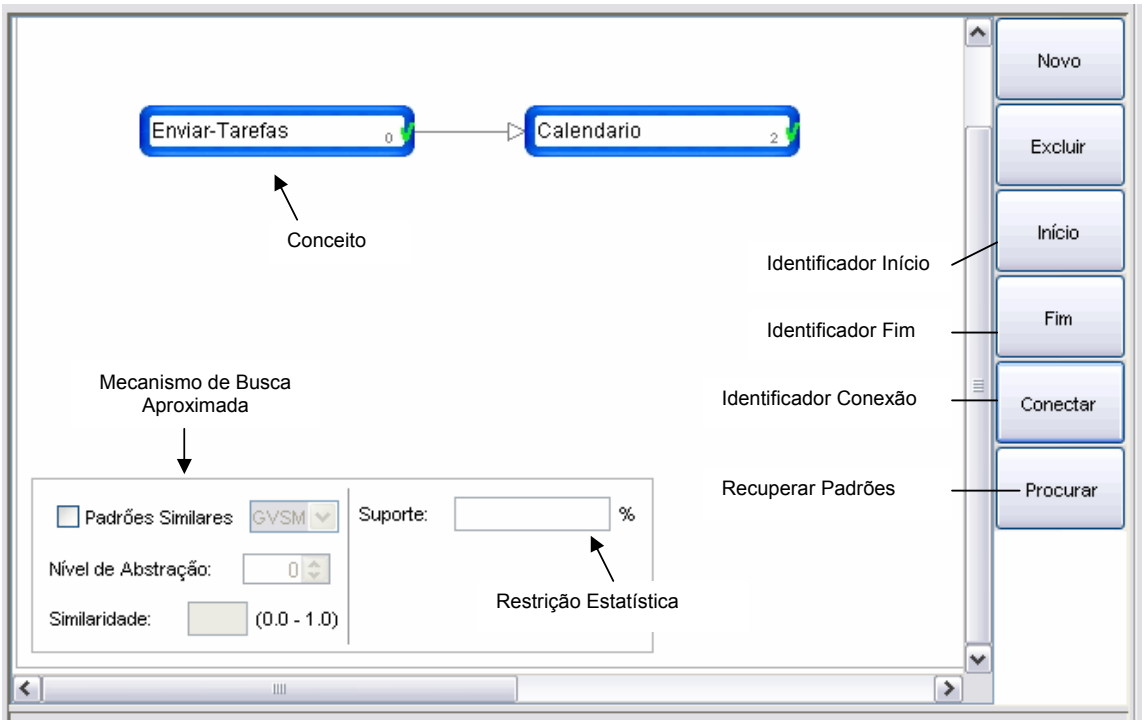


Figura 60: Área de Definição de Filtro

Quanto aos mecanismos de busca, a opcção padrão é por equivalência. Se o usuário especificar a busca por padrões similares, deve informar o restante dos parâmetros de acordo como interesse. Atualmente, apenas o algoritmo GVSM está disponível. Porém, outros algoritmos de similaridade podem ser incluídos.

O botão *Localizar* recupera padrões conceituais base de acordo com o mecanismo de busca. Demais botões (*Novo* e *Excluir*) fornecem funcionalidades adicionais .

7.1.4.3 *Área de Padrões Filtrados*

A área de Padrões Filtrados, visualizada na Figura 61, mostra os padrões recuperados de acordo com as restrições definidas pelo filtro, com o seu respectivo suporte. Se o mecanismo de busca for o aproximado, o padrão ainda apresenta o valor para a medida de similaridade do padrão em relação ao filtro definido. Ainda, se o analista selecionar um padrão nesta área, ele passa a ser visualizado graficamente na área de Análise Exploratória.

Padrão	Suporte
Acessar-WebCT -» Carregar-Arquivo	5.40%
Acessar-Curso-ABC -» Carregar-Arquivo	5.40%
Acessar-WebCT -» Carregar-Arquivo	5.94%
Acessar-Curso-ABC -» Carregar-Arquivo	5.94%
Acessar-WebCT -» Acessar-Curso-ABC -» Carregar-Arquivo	5.40%
Acessar-Curso-ABC -» Acessar-Curso-ABC -» Carregar-Arquivo	5.94%
Acessar-Curso-ABC -» Acessar-Curso-ABC -» Carregar-Arquivo	5.94%
Acessar-WebCT -» Acessar-Curso-ABC -» Carregar-Arquivo	5.94%
Acessar-WebCT -» Acessar-Curso-ABC -» Acessar-Curso-ABC -» Carregar-Arquivo	5.94%
Acessar-WebCT -» Acessar-Curso-ABC -» Acessar-Curso-ABC -» Carregar-Arquivo	5.94%
Carregar-Arquivo -» Descricao-Atividade	5.40%
Descricao-Atividade -» Carregar-Arquivo	5.40%
Visualizar-arquivo -» Carregar-Arquivo	5.40%
Carregar-Arquivo -» Carregar-Arquivo	5.40%
Acessar-WebCT -» Carregar-Arquivo -» Descricao-Atividade	5.40%
Acessar-Curso-ABC -» Carregar-Arquivo -» Descricao-Atividade	5.40%
Acessar-WebCT -» Descricao-Atividade -» Carregar-Arquivo	5.40%

Figura 61: Área de Padrões Filtrados

8 ESTUDO DE CASO EM UM AMBIENTE DE ENSINO A DISTÂNCIA

O presente capítulo descreve um estudo de caso realizado no contexto da Educação a Distância para avaliar os mecanismos de recuperação e interpretação de padrões propostos para fase de Análise de Padrões.

O processo de MUW é aplicado nas mais diversas áreas. Uma delas é a Educação a Distância (EAD), na qual tem crescido a utilização do ensino baseado na *Web*. A aplicação do processo de MUW nos dados relativos às interações dos estudantes através destes ambientes permite detectar padrões de utilização dos mesmos, bem como padrões de aprendizagem. Através deste conhecimento é possível refletir sobre a adequação da estrutura do *site*, em termos de serviço e conteúdo, e ainda compreender como os usuários navegam pelo *site* de acordo com os diferentes modelos de aprendizagem para aquisição de competências.

O domínio da EAD foi o escolhido para a realização do estudo de caso desta pesquisa devido à facilidade na obtenção do *log* juntamente com o departamento de EAD da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS). Outro motivo pela escolha deste ambiente é a possibilidade de uma comparação desta proposta com o trabalho de Machado [MAC03], que aplica a MUW sobre o mesmo *log*. Machado vivenciou sérios problemas de interpretação e recuperação de padrões na fase de Análise de Padrões, os quais serviram para a motivação desta pesquisa. Estes problemas são descritos com maiores detalhes na Seção 8.3.1.

Nas seções seguintes, o ambiente de EAD da PUCRS, gerenciado pela ferramenta WebCT (*Web Course Tool*), é descrito em maiores detalhes, assim como o estudo de caso realizado para este trabalho. O estudo de caso tem como objetivo avaliar como a abordagem proposta auxilia o analista na fase de Análise de Padrões. Para isso, é descrito um cenário de uso do protótipo, complementando com uma comparação do processo de MUW que utiliza os mecanismos propostos para auxiliar na fase de Análise de Padrões, com o trabalho de Machado [MAC03] que não utiliza nenhuma ambiente de apoio a esta fase.

8.1 Ambiente de Ensino da EAD da PUCRS

Na proposta de EAD criada pela PUCRS, cada curso ou projeto apresenta uma construção própria e comporta capacitação, assessoramento e monitoramento de professores, monitores, tutores e dos próprios alunos distantes, no sentido de facilitar o trânsito e a construção de ambientes orientados à aprendizagem [MED01]. Tais ambientes são gerenciados pelo programa do WebCT em qualquer curso ou projeto que se instala na PUC-Virtual.

WebCT é uma plataforma composta de um conjunto de ferramentas que facilita a criação e manutenção de cursos educacionais baseados em interfaces *Web* [WCT02]. O acesso ao WebCT é realizado a partir de um servidor central, que registra no *log* do servidor *Web* todo e qualquer acesso às páginas que compõem um ambiente de ensino [GOL96].

O WebCT suporta um ambiente de ensino e aprendizado integrado, contendo uma série de ferramentas educacionais tais como sistema de conferência, bate-papo, correio eletrônico, acompanhamento do aluno, suporte para projetos colaborativos, auto-avaliação, questionários, distribuição e controle de notas, glossário, controle de acesso, calendário do curso, geração automática de índices e pesquisa, entre outras. A Figura 63 ilustra um ambiente de ensino construído pelos recursos disponíveis pelo WebCT.

Figura 62: Ambiente de ensino construído pelos recursos do WebCT

O WebCT disponibiliza alguns recursos para monitoração do comportamento dos alunos referente à navegação do *site*. Um deles é a estatística de dados de caráter geral, tais

como o número de *hits* (acessos) por página, as páginas acessadas mais freqüentemente, e o tempo médio de acesso de cada página, etc. Outro recurso que pode ser utilizado para medir a adequação da estrutura e do conteúdo do curso, é através de comentários enviados pelos próprios alunos através de mensagens.

Porém, estas formas de monitoração e acompanhamento fornecidas pelo WebCT não são suficientes para uma avaliação consistente do uso dos recursos pelos alunos. Assim, a carência de informações mais detalhadas para estimar e expressar a eficiência do uso dos recursos de ensino e a falta de mecanismos de acompanhamentos mais efetivos dos comportamentos dos usuários são alguns dos fatores que estimulam a aplicação do processo de MUW para ambientes de EAD.

8.2 *Log* do WebCT

A ferramenta WebCT registra os acessos a todos os recursos ofereciso por um ambiente de ensino. O formato dos arquivos de *log* gerado é do tipo CLF. Cada transação indica quais páginas *Web* ou scripts foram requisitados, quando e de onde partiu esta solicitação e, ainda podem trazer a identificação do usuário. A Figura 63 ilustra uma amostra de *log* gerado pelo WebCT. Nota-se que a maioria das páginas acessadas em um ambiente WebCT corresponde a chamadas de scripts.

Log Web

```
200.176.25.110 - aluno1 [10/Jan/2002:00:00:06 -0200] "GET /ESP_SE_01130/competencia/07_01/conselhos.doc HTTP/1.1" 200 31744
200.176.8.249 - -       [10/Jan/2002:00:10:17 -0200] "GET / HTTP/1.1" 200 189
200.176.8.249 - -       [10/Jan/2002:00:10:18 -0200] "GET /webct/public/home.pl HTTP/1.1" 200 1977
200.176.8.249 - aluno2  [10/Jan/2002:00:10:39 -0200] "GET /webct/homearea/homearea HTTP/1.1" 200 20032
200.176.8.249 - -       [10/Jan/2002:00:11:20 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/dropbox_stud_home.pl HTTP/1.1" 401 899
200.248.5.164 - -       [10/Jan/2002:00:11:21 -0200] "GET /webct/homearea/homearea HTTP/1.1" 401 866
200.176.8.249 - aluno2  [10/Jan/2002:00:11:25 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/dropbox_stud_home.pl HTTP/1.1" 200 28552
200.176.8.249 - aluno2  [10/Jan/2002:00:11:26 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/serve_layout.pl?LOGO HTTP/1.1" 200 52
200.176.8.249 - aluno2  [10/Jan/2002:00:11:30 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/dropbox_stud_home.pl?START+++ HTTP/1.1" 200 8817
200.248.5.164 - aluno3  [10/Jan/2002:00:11:32 -0200] "GET /webct/homearea/homearea HTTP/1.1" 200 12498
200.248.5.164 - -       [10/Jan/2002:00:11:58 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/serve_home HTTP/1.1" 401 881
200.248.5.164 - aluno3  [10/Jan/2002:00:12:02 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/serve_home HTTP/1.1" 200 20172
200.248.5.164 - aluno3  [10/Jan/2002:00:12:05 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_home?_homepage+START HTTP/1.1" 200 4105
200.248.5.164 - aluno3  [10/Jan/2002:00:12:14 -0200] "GET /Curso_ABC_02JAN/AmbienteCurso.pdf HTTP/1.1" 200 16485
200.248.5.164 - aluno3  [10/Jan/2002:00:13:41 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_home?1010410417+view HTTP/1.1" 200 6340
200.176.25.110 - aluno1 [10/Jan/2002:00:13:42 -0200] "GET /ESP_SE_01130/competencia/07_01/paulo_freire_texto.pdf HTTP/1.1" 200 41563
200.248.5.164 - aluno3  [10/Jan/2002:00:13:49 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_home?1010412547+view HTTP/1.1" 200 7930
200.176.20.28 - aluno2  [10/Jan/2002:00:15:15 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/dropbox_view.pl?START+++1010521868 HTTP/1.1" 200 2148
200.176.25.110 - aluno1 [10/Jan/2002:00:16:38 -0200] "GET /ESP_SE_01130/competencia/07_01/paulo_freire_texto.pdf HTTP/1.1" 304 -
200.176.25.110 - aluno1 [10/Jan/2002:00:17:16 -0200] "GET /ESP_SE_01130/competencia/07_01/reunioes_prod.doc HTTP/1.1" 200 31744
200.248.5.164 - aluno3  [10/Jan/2002:00:17:39 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_home?1010431964+view HTTP/1.1" 200 8736
200.248.5.164 - aluno3  [10/Jan/2002:00:18:41 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_home?1010422540+view HTTP/1.1" 200 8498
200.248.5.164 - aluno3  [10/Jan/2002:00:18:51 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_mail?START HTTP/1.1" 200 5938
200.248.5.164 - -       [10/Jan/2002:00:18:52 -0200] "GET /web-ct/style/stylesul.txt HTTP/1.1" 200 2419
200.248.5.164 - aluno3  [10/Jan/2002:00:19:01 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_chat.pl?START+1010422540 HTTP/1.1" 200 570
200.248.5.164 - aluno3  [10/Jan/2002:00:19:03 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_chat.pl?CLIENT+1010422540 HTTP/1.1" 200 831
200.248.5.164 - aluno3  [10/Jan/2002:00:19:03 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_chat.pl?SERVER+1010422540 HTTP/1.1" 200 714
200.248.5.164 - -       [10/Jan/2002:00:19:11 -0200] "GET /web-ct/code/Client.class HTTP/1.1" 200 6068
```

Figura 63: Amostra do *Log* do WebCT

8.3 Processo de MUW na EAD

O Grupo de Sistema de Informação da PUCRS vêm desenvolvendo diversos trabalhos relacionados ao processo de MUW nos ambientes de ensino à distância ([e.g. MAC03, MAR04]), onde um curso da PUC-Virtual é utilizado no estudo de caso. Estes trabalhos consideram como fonte de dados o *log* de um curso de extensão gerenciado pelo WebCT denominado neste trabalho como "Curso_ABC". O Curso_ABC ocorreu de 08 e 18 de janeiro de 2002, e contou com a participação de 15 alunos.

A topologia das páginas do Curso_ABC estão representadas na Figura 64. Observa-se as páginas de apresentação do curso (4 *links*) e as páginas relacionadas aos recursos utilizados para desempenho e cumprimento das atividades propostas pelo curso (5 *links*). As páginas marcadas em itálico, possuem ramificações para outras páginas do *site*.

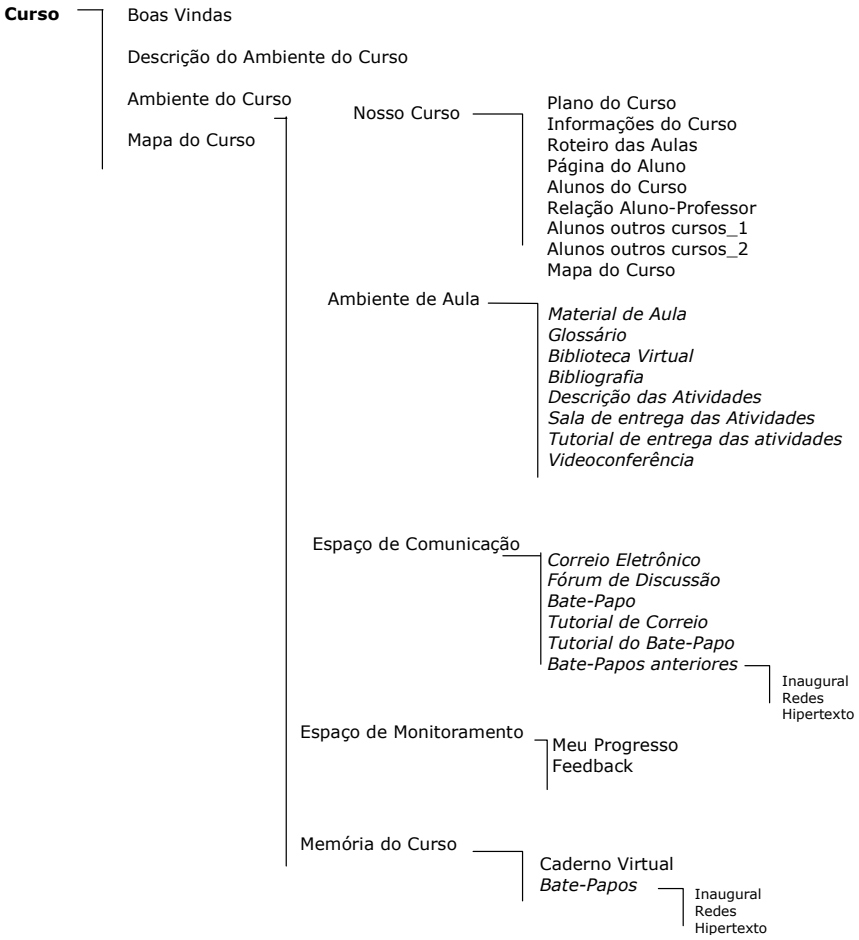


Figura 64: Topologia do Curso_ABC

A seção seguinte descreve algumas particularidades relacionadas ao trabalho de Machado [MAC03], que serviu de motivação para a presente pesquisa, e principalmente como referencial de avaliação da abordagem proposta neste trabalho.

8.3.1 Abordagem de Machado [MAC03]

O trabalho de Machado [MAC03] buscou estabelecer um modelo de processo para a MUW voltado à EAD. Através da análise do comportamento de navegação, deseja-se monitorar as atividades dos alunos durante o curso on-line, assim com avaliar a utilização dos recursos oferecidos pelo *site* educacional.

Para atingir os objetivos propostos, o processo de MUW foi executado diversas vezes visando demonstrar os diferentes tipos de padrões que poderiam ser extraídos. Para tal, aplicou-se as técnicas de associação e padrões seqüenciais sobre o *log* preprocessado, ambas disponibilizadas na ferramenta *Intelligent Miner* [IBM04]. Como resultado obteve-se uma grande quantidade de padrões de difícil interpretação e de pouca representatividade, o que despertou pouco interesse no analista, no caso uma pessoa vinculada à PUC-Virtual, conhecedora do curso em questão.

Visando aumentar a representatividade dos padrões e também a semântica associada a eles, optou-se pela associação de taxonomias ao processo de MUW, como descrito nos trabalhos de Agrawal e Srikant [SRI95, SRI97]. O resultado da fase de Mineração de Dados foi um aumento no conjunto de padrões formados por dezenas de milhares de padrões. Considerando o excessivo volume, o analista selecionava alguns padrões aleatoriamente que pareciam ser mais relevantes, os quais na maioria das vezes eram padrões generalizados por apresentar maior representatividade. O analista ratificava então o interesse nos padrões relacionados a título de exemplo, e neste momento geralmente demonstrava interesse nos mais específicos. Contudo, como a ferramenta não oferecia qualquer tipo de apoio, o relacionamento manual de regras relacionadas tornou-se uma tarefa de extrema complexidade. Desta forma, os recursos para a fase de Análise de Padrões eram muito limitados, comprometendo os resultados do processo de MUW.

8.4 Estudo de Caso

O estudo de caso descrito neste trabalho também foi realizado no contexto da PUC-Virtual considerando a mesma fonte de dados utilizada na abordagem de Machado [MAC03] e uma técnica de Mineração de Dados utilizada, a saber a de padrões seqüenciais físicos. A diferença entre as abordagens está na fase de Análise de Padrões, onde este estudo de caso oferece mecanismos de apoio à atividade de análise de padrões propostos por esta pesquisa.

8.4.1 Preparação de Dados

Primeiramente, os dados contidos nos arquivos de *log* passaram pela etapa de Preparação de Dados. Os dados utilizados para este estudo de caso foram selecionados abrangendo um período determinado para a execução de uma atividade proposta para Curso_ABC, compreendendo 6 dias de curso.

Utilizando a ferramenta de pré-processamento desenvolvida por Marquardt [MAR04], a este *log* foram aplicadas as seguintes operações:

- Limpeza: Eliminação dos registros de acesso com extensões não significativas (e.g. .gif, .jpeg, .css);
- Filtragem: Foram eliminados todos os registros sem identificação de usuário que não se referem ao Curso_ABC; e páginas que não oferecem recursos aos estudantes, ou seja, páginas que servem somente como elo de conexão entre outras páginas;
- Identificação de Sessão: Os acessos formam organizados em sessões de atividades. Machado [MAC03] define uma sessão de atividade como o conjunto de recursos acessados por um estudante para a execução de uma atividade específica, proposta pelo curso.
- Transformação - Mapeamento de Acessos: A cada URL foi associada a um identificador numérico para facilitar a manipulação dos dados na fase de Mineração.

O *log* preparado resultou em 3410 registros.

8.4.2 Ontologia de Domínio e Mapeamento

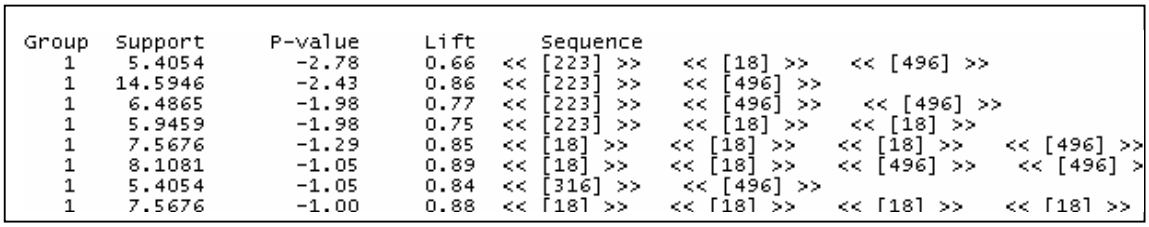
A Ontologia de Domínio para o Curso_ABC foi criada manualmente para avaliação dos mecanismos propostos por esta abordagem. Cabe ressaltar que o presente trabalho não tem por objetivo avaliar a ontologia definida e nem o método utilizado para isto, mas sim explorar a utilização desta como suporte à fase de Análise de Padrões.

Para a construção da Ontologia de Domínio foram identificadas as URLs distintas que compunham o *log* preparado. No total foram identificadas 87 URLs. Cada URL foi acessada e os eventos de domínio representados foram identificados. Estes eventos de domínio eram representados pelos conceitos de serviço e conteúdo que passavam a fazer parte da Ontologia de Domínio. Assim, à medida que as URLs eram acessadas, a Ontologia de Domínio sofria refinamentos e o mapeamento, que determina a quais conceitos da ontologia a URL estava associada, também era definido.

Para finalizar, a ontologia sofreu o último refinamento, onde alguns conceitos e relacionamentos foram criados de acordo com o conhecimento do domínio adquirido. No total a ontologia criada é formada por 127 conceitos e 130 relações.

8.4.3 Descoberta de Padrões de Uso da *Web* na EAD

A técnica de Mineração de Dados selecionada para a descoberta de padrões é o algoritmo *AprioriAll*, descrito na Seção 2.2.2.1 e disponível na ferramenta *Intelligent Miner*. O valor mínimo de suporte (*minsup*) informado para a geração dos padrões seqüenciais físicos foi 5%. No total foram descobertos 5530 padrões. A Figura 65 ilustra uma pequena amostra do arquivo texto exportado pela ferramenta *Intelligent Miner*, o qual armazena as informações sobre os padrões seqüenciais físicos, incluindo o respectivo suporte. O cabeçalho do arquivo texto deve ser removido para a importação das informações pelo protótipo na fase de Análise de Padrões.



Group	Support	P-value	Lift	Sequence
1	5.4054	-2.78	0.66	<< [223] >> << [18] >> << [496] >>
1	14.5946	-2.43	0.86	<< [223] >> << [496] >>
1	6.4865	-1.98	0.77	<< [223] >> << [496] >> << [496] >>
1	5.9459	-1.98	0.75	<< [223] >> << [18] >> << [18] >>
1	7.5676	-1.29	0.85	<< [18] >> << [18] >> << [18] >> << [496] >>
1	8.1081	-1.05	0.89	<< [18] >> << [18] >> << [496] >> << [496] >
1	5.4054	-1.05	0.84	<< [316] >> << [496] >>
1	7.5676	-1.00	0.88	<< [18] >> << [18] >> << [18] >> << [18] >>

Figura 65: Amostra de um arquivo texto obtido pela ferramenta *Intelligent Miner*

8.5 Análise de Padrões: Cenário de Uso

Esta seção descreve um cenário de uso do protótipo que disponibiliza os mecanismos de interpretação e recuperação de padrões utilizados pelo analista na fase de Análise de Padrões, considerando as atividades desenvolvidas anteriormente.

8.5.1 Definições iniciais

Primeiramente, o analista interage com a área de Importação dos Padrões do protótipo (Módulo de Definições), onde seleciona o arquivo texto contendo os padrões seqüenciais de URLs resultantes da fase de Mineração de Dados. Nesta mesma área ele delimita as posições que determinam o padrão seqüencial físico e seu valor de suporte. O analista ainda especifica os caracteres que identificam uma URL.

Importados os dados, o analista seleciona o critério que será utilizado para agrupar os padrões seqüenciais físicos, a saber critério maximal (área de Definição de Agrupamentos), e define a dimensão de interesse na qual os padrões conceituais serão analisados (área de Definição de Dimensão de Interesse). Neste exemplo, a dimensão escolhida foi conteúdo e serviço.

8.5.2 Inspecionando Agrupamentos e Interpretando Padrões

Na área de Agrupamentos de Padrões, o analista visualiza os padrões conceituais agrupados obedecendo ao critério maximal e representados de acordo com a dimensão de interesse de serviço e conteúdo. Neste estudo de caso foram gerados 40 agrupamentos para 5530 padrões seqüenciais físicos.

Inicialmente, o analista explora estes agrupamentos visíveis na área de Agrupamentos de Padrões. No exemplo ilustrado pela Figura 66-A, o analista seleciona um agrupamento que julga interessante onde o padrão maximal expressa que 9,72% das sessões de usuários acessaram a ferramenta WebCT seguidos pelo Curso_ABC, requisitaram a lista de atividades e posteriormente seguiram para bate-papo. Nota-se que a interpretação do padrão é facilitada pela utilização dos conceitos da ontologia. O analista ainda verifica os padrões contidos no agrupamento na área de Padrões Contidos (Figura 66-B), mas decide aprofundar a interpretação no padrão maximal.

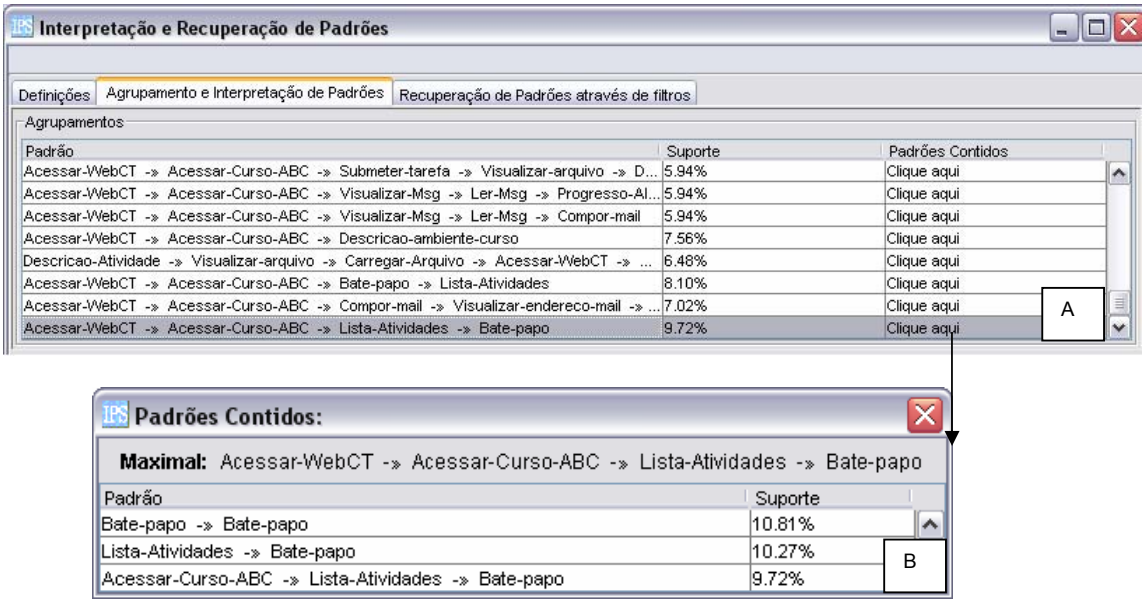


Figura 66: Inspecionando área de Agrupamentos de Padrões e Padrões Contidos

Ao selecionar o padrão maximal na área de Agrupamentos de Padrões, este é automaticamente representado graficamente na área de Análise Exploratória de acordo com a dimensão de interesse serviço e conteúdo, como representado pela Figura 67-B.

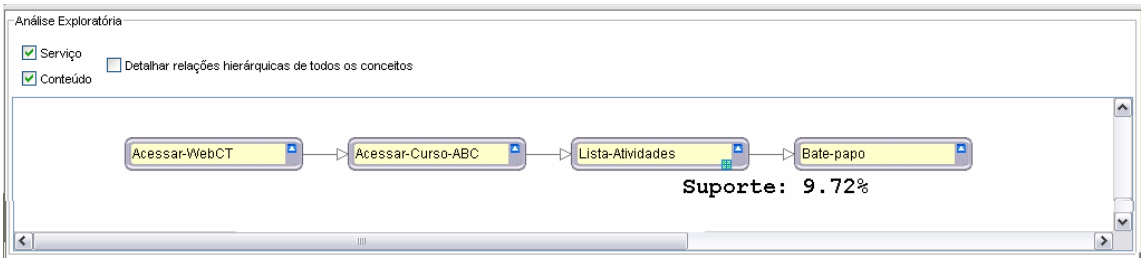


Figura 67: Área de Análise Exploratória

Antes de aprofundar a interpretação do padrão, o analista resolve verificar o padrão considerando apenas a dimensão de conteúdo, buscando saber quais conteúdos estavam sendo manipulados pelos usuários. Para isso ele apenas desmarca a opção serviço, localizada logo acima da representação gráfica do padrão conceitual base. A Figura 68 representa o padrão conceitual visualizado nesta dimensão. Ele verifica que os conteúdos envolvidos no padrão referem-se aos dados dos usuários e a lista de atividades. Existem também conceitos que não possuem conteúdo disponível.

Surgiu então, uma curiosidade de verificar o padrão em termos de serviços acessados. Para isso o analista apenas desmarca a opção conteúdo e marca a opção serviço. O padrão visualizado (Figura 69) expressa que os usuários acessaram o WebCT, acessaram o Curso_ABC, visualizaram informações e foram para o bate-papo. Ao final, o analista resolve voltar à dimensão de interesse serviço e conteúdo para realizar a análise exploratória.

Figura 68: Padrão conceitual na dimensão de interesse em conteúdo

Figura 69: Padrão conceitual na dimensão de interesse em serviço

Para realizar a análise exploratória do padrão, o analista tem que interagir com ele. Primeiramente, ele realiza algumas operações de detalhamento de relacionamentos hierárquicos habilitada para os 4 conceitos que compõem o padrão conceitual base (Figura 70). Ele solicita o detalhamento de informações hierárquicas sobre o conceito *Lista-Atividades* e descobre que esta faz parte da arquitetura pedagógica do Curso_ABC, através da existência de uma relação de composição entre os conceitos *Lista-Atividades* e *Arquitetura-Pedagógica*. Visando adquirir maiores informações sobre o conceito *Arquitetura-Pedagógica* através de outra operação de detalhamento de relacionamento hierárquico, ele compreende que este faz parte do conteúdo. Da mesma forma, operações detalhamento de relacionamentos

hierárquicos foram realizadas sobre os outros conceitos, descobrindo que bate-papo é um recurso de comunicação e que este por sua vez é um serviço.

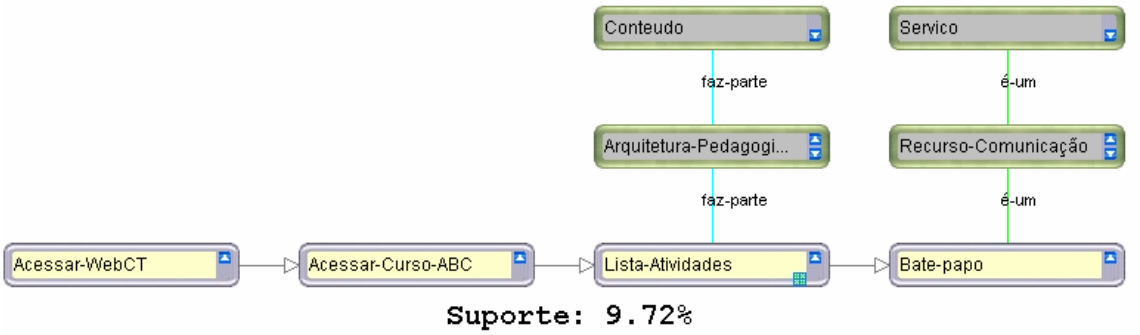


Figura 70: Realizando operações de detalhamento de relações hierárquicas

Ainda, o analista visualiza que existe uma relação de propriedade entre o conceito *Lista-Atividades* e outro conceito da ontologia. Curioso, ele realiza a operação de detalhamento de relacionamentos de propriedade. Uma mensagem é mostrada, informando que a lista de atividade é disponibilizada no Curso_ABC, como mostrado na Figura 71.

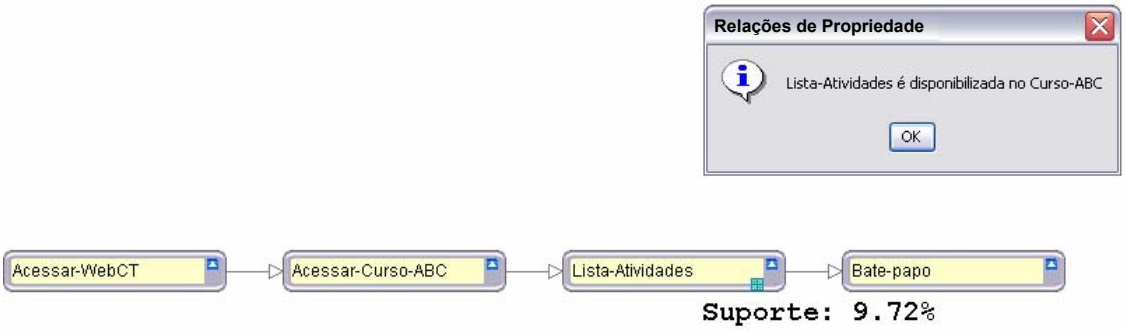


Figura 71: Explorando o significado das relações de propriedade

Para o analista, o padrão foi completamente compreendido de forma intuitiva através do padrão conceitual base e das informações que adquiriu com as sucessivas operações de detalhamento de relacionamentos hierárquicos e de propriedade. Neste ponto da interpretação, começam a surgir hipóteses sobre as informações contidas no padrão. O analista supõe que ao acessar a lista de atividades, os alunos encontraram dúvidas sobre as atividades requisitadas pelo Curso_ABC e por isto acessaram o bate-papo para questionamentos. Considerando esta hipótese, o analista resolveu verificar quais os outros recursos de comunicação foram

utilizados pelos alunos após o acesso a lista de atividades. Desta forma, ele realiza a operação de *roll-up* sobre o conceito *Recurso-Comunicação*, substituindo o conceito *Bate-papo* para o *Recurso-Comunicação*. O padrão abstrato obtido é ilustrado pela Figura 72. O analista verifica que o padrão abstrato criado possui suporte superior ao padrão conceitual base explorado anteriormente, ou seja, os alunos acessaram outros recursos de comunicação além do bate-papo.

Para verificar quais os padrões conceituais detalhe que sumarizam o padrão abstrato, o analista requisita a operação de *drill-down*. Como resultado, uma janela contendo os padrões conceituais detalhe é mostrada na área de Padrões Detalhe (Figura 73). A partir deste padrões, o analista verifica que os alunos foram visualizar seus e-mail.

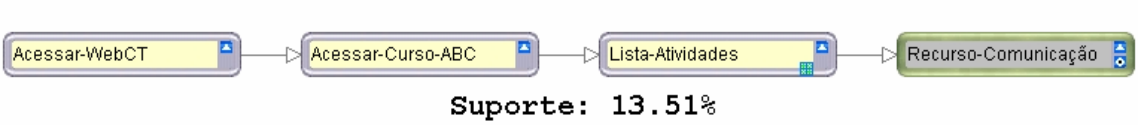


Figura 72: Exemplo de padrão abstrato

Figura 73: Padrões conceituais detalhe

Nota-se até então, que o protótipo permite uma intensa interatividade com o analista de maneira amigável e flexível. Ou seja, o analista facilmente inspeciona os padrões presentes nos agrupamentos que despertam interesse, seleciona um padrão conceitual base para aprofundar a interpretação, escolhe diferentes dimensões de interesse sem re-execução as fases anteriores, e realiza operações que compõem a analise exploratória, podendo assim aprofundar a compreensão e descobrir padrões relacionados.

8.5.3 Definindo filtros e Recuperando Padrões

Surge então, uma curiosidade de verificar se os estudantes estão preocupados com os seus desempenhos durante o curso Curso_ABC. Para verificar esta hipótese o analista define filtros de interesse interagindo com a Ontologia de Domínio representada graficamente na área da Ontologia de Domínio. Nela, ele seleciona o conceito *Progresso-aluno*, adicionando-os na área de Definição de filtros, conforme visualizado na Figura 74-A. Após definir o filtro, o analista requisita a recuperação dos padrões que estão de acordo com o filtro utilizando o mecanismo de busca equivalente. Neste caso, foram encontrados 48 padrões, visualizados na área de Padrões Filtrados (Figura 74-B).

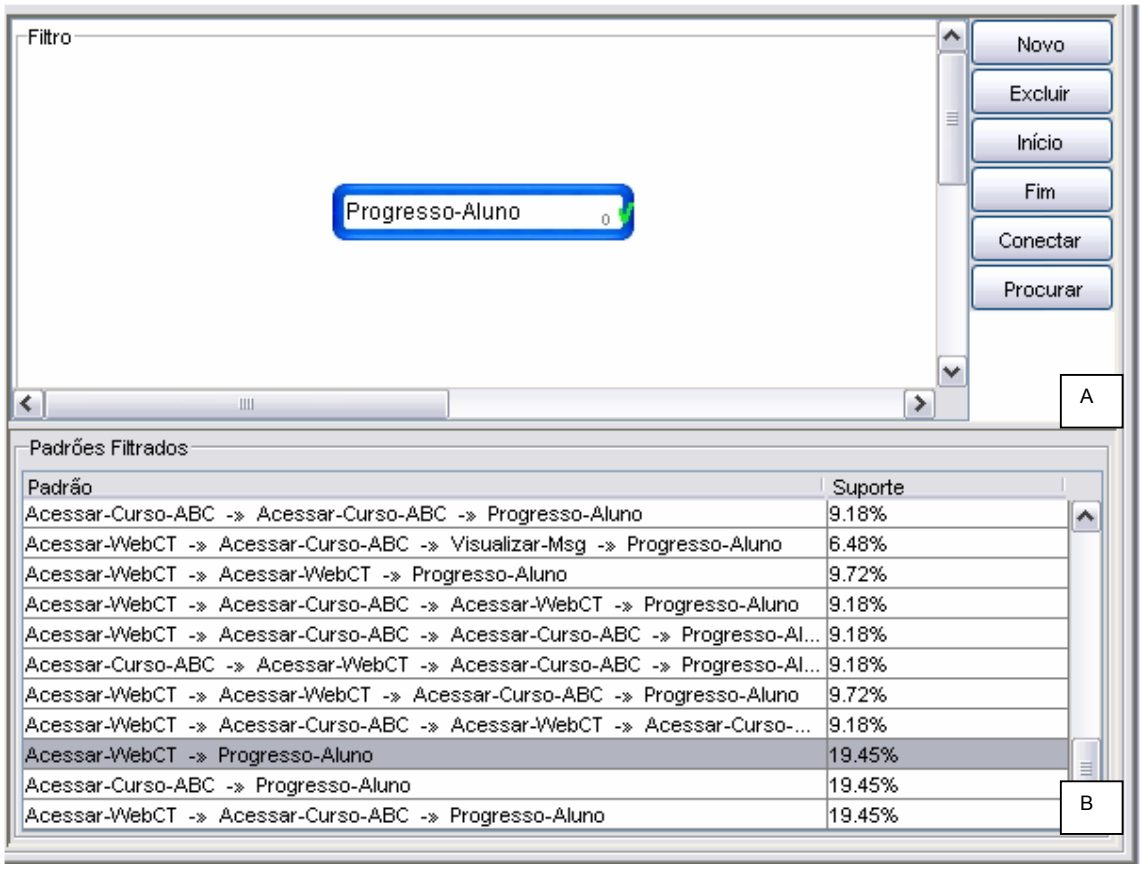


Figura 74: Definição do filtro de interesse - I

Analisando os padrões, o analista verifica que realmente os estudantes acompanham seus desempenhos no curso Curso_ABC. Porém, surge a curiosidade de verificar se os alunos são instigados a acompanhar o seu desempenho pelo acesso ao recurso de comunicação bate-

papo, que julga ser o mais utilizado. O analista supõe que o desempenho dos alunos pode ser pauta para muitas discussões na ferramenta de bate-papo.

Para isso, o analista aprimora o filtro definido, expressando que os padrões conceituais devem conter uma requisição ao recurso de bate-papo e posteriormente ao desempenho dos alunos do curso (Figura 75). Obedecendo estas restrições, 4 padrões foram recuperados. Analisando os padrões, mais uma vez o analista tem sua hipótese comprovada.

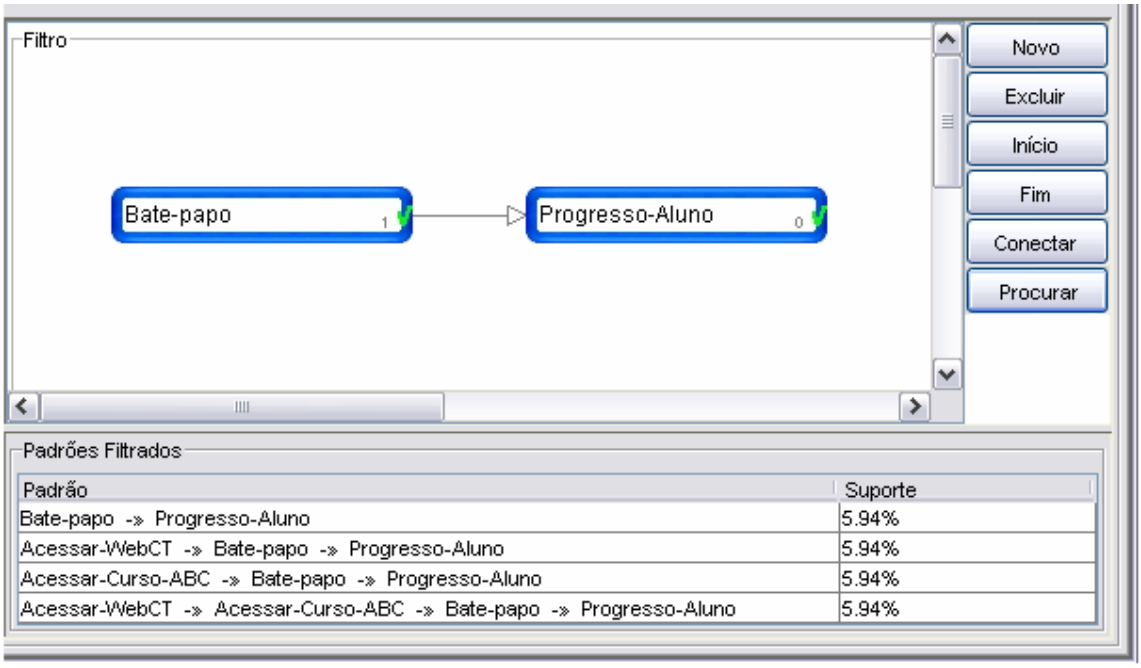


Figura 75: Definição do filtro de interesse - II

O analista então resolve ampliar o escopo da pesquisa por padrões, aplicando o método de busca por padrões aproximados, como visualizado pela Figura 76. Para isto ele escolhe a medida de similaridade escolhida foi a GVSM, o nível de abstração informado foi 1, e o valor de similaridade mínimo do padrão foi de 0,7. Para esta situação especificada, foram recuperados 342 padrões. Nota-se que o padrão conceitual que possui valor de similaridade 1 é aquele que casa com os interesses especificados pelo analista através do filtro. Os padrões com similaridade menor que 1 são aqueles que possuem algumas variações.

Por exemplo, o padrão selecionado com similaridade 0,9 é um padrão que representa uma similaridade muito próxima ao interesse especificado pelo filtro pois ao invés dos usuários acessarem o bate-papo, como especificado no filtro, eles acessaram outro recurso de

comunicação que possibilita a visualização de e-mail. Ao analisar este padrão, o analista achou-o interessante pois demonstra que os alunos também utilizam outros recursos de comunicação e não somente o bate-papo.

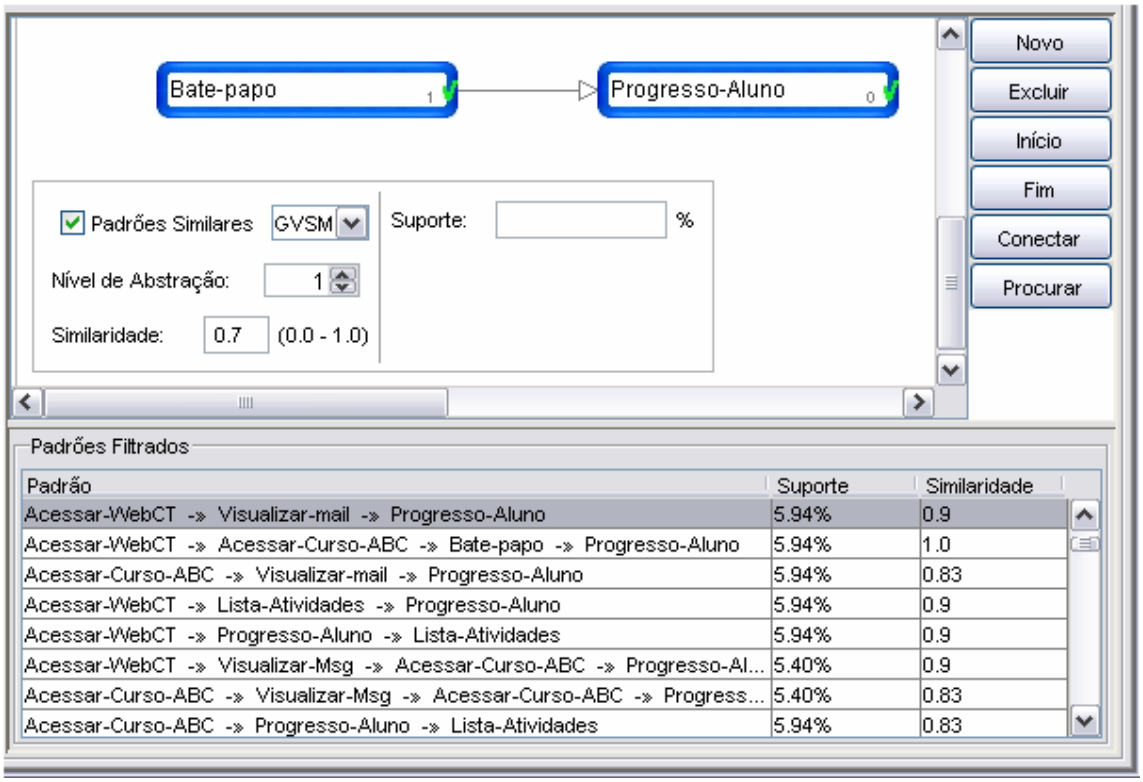


Figura 76: Aplicação do método de busca aproximada

8.6 Considerações

Os mecanismos propostos e desenvolvidos no ambiente de apoio à fase de Análise de Padrões apresentam diversas vantagens em relação a esta fase em um processo de MUW aplicado no mesmo domínio, porém sem auxilio de mecanismos para a interpretação e recuperação de padrões. Estas vantagens e desvantagens estão sumarizadas na Tabela 13.

Tabela 13. Comparação do Processo de MUW anterior com o atual.

Critérios		Análise dos Critérios	
Fase de Análise de Padrões do Processo MUW de Machado	**Fase de Análise de Padrões do Processo MUW de Vanzin**	**Fase de Análise de Padrões do Processo MUW de Machado**	**Fase de Análise de Padrões do Processo MUW de Vanzin**
Forma de visualização do padrão			
Textual	Textual e gráfica	Não existe interatividade com os padrões.	A representação gráfica permite a interatividade como padrão, auxiliando-o na interpretação e recuperação de padrões.
Composição do Padrão visualizado na fase de Análise			
Padrão é composto por um conjunto de identificadores numéricos ou por conceitos da taxonomia.	Conjunto de conceitos de serviço e conteúdo, de acordo com a dimensão de interesse.	Dificuldade de interpretar o significado de cada padrão.	Fácil interpretação, possibilitando a escolha de diferentes dimensões de interesse.
Exploração do significado do Padrão			
Não disponível.	Operação de detalhamento de relacionamentos hierárquicos e de propriedade.	Não existe a possibilidade de aprofundar a interpretação das informações expressas pelo padrão.	Permite aprofundar a compreensão do significado dos conceitos que compõem o padrão.
Obtenção dos Padrões Abstratos			
Todos os padrões abstratos foram obtidos como resultado do algoritmo de Mineração de Dados.	Os padrões abstratos são obtidos sobre demanda, através da operação de *roll-up*	Dificuldade de recuperar padrões relevantes devido ao grande número de padrões retornados com o uso de taxonomia.	Diminuição no número de padrões retornados, conseqüentemente, maior facilidade de recuperar padrões relevantes, uma vez que os padrões abstratos são obtidos sob demanda.

Critérios		Análise dos Critérios	
Fase de Análise de Padrões do Processo MUW de Machado	**Fase de Análise de Padrões do Processo MUW Atual**	**Fase de Análise de Padrões do Processo MUW de Machado**	**Fase de Análise de Padrões do Processo MUW de Vanzin**
Obtenção dos padrões que detalham os Generalizados			
Não disponível.	Os padrões que detalham os padrões generalizados são obtidos pela operação de *drill-down*.	Impossibilidade de recuperar os padrões que suportam o generalizado de forma rápida e intuitiva.	Possibilidade de recuperação rápida dos padrões que suportam o padrão generalizado (abstrato) interagindo com o mesmo.
Definição de filtros			
Composto por um vetor de páginas, como forma de expressão genérica de padrões.	Composto por um conjunto de elementos, permitindo definir restrições conceitual, estruturais e estatísticas. A restrição conceitual é definida através da interação com Ontologia de Domínio representada graficamente.	Conhecimento do domínio para a definição de filtros e limitações de restrições.	Facilidade para a definição do filtro, feito de forma interativa com a Ontologia de Domínio. O analista não precisa ter um profundo conhecimento do domínio e nem dominar uma sintaxe em particular.
Mecanismos de busca			
Mecanismo de busca equivalente.	Mecanismo de busca equivalente e por aproximação.	Impossibilidade de descobrir padrões similares ao filtro definido.	Possibilidade de recuperar padrões similares ao interesse especificado pelo filtro.
Agrupamento de Padrões			
Não utilizado.	Agrupamento de padrões de acordo com critério pré-definido.	Muito tempo consumido na análise dos padrões, uma vez que muitos são redundantes e desinteressantes.	Possibilidade de direcionar o foco em grupos de padrões específicos. Uma vez que não há um interesse num determinado grupo, vários padrões são desconsiderados, otimizando o tempo de análise.

8.7 Depoimento do Analista

O analista envolvido no processo de MUW guiado pelo trabalho de Machado [MAC03] participou de uma demonstração da utilização do protótipo desenvolvido no escopo deste trabalho. O objetivo foi avaliar a utilidade dos mecanismos propostos na fase de Análise de Padrões, considerando o mesmo domínio para o estudo de caso utilizado no trabalho de Machado.

Após a demonstração, o analista participou de uma entrevista. As questões formuladas e o parecer do analista podem ser encontrados no Anexo II deste volume. A entrevista guiou o analista a estabelecer um comparativo das atividades realizadas na fase de Análise de Padrões no processo de MUW vivenciado anteriormente, sem um ambiente de apoio a esta fase, com a fase de Análise descrita neste estudo de caso. Cabe ressaltar que o objetivo não era de encontrar padrões relevantes, e sim avaliar se os mecanismos propostos auxiliariam na busca por padrões potencialmente relevantes.

Quanto à fase do processo anterior, o analista enfatiza que era dispendido muito tempo e esforço para entender o significado dos padrões, além de que também tinha que entender alguns conceitos técnicos do processo para prosseguir a busca por padrões relevantes. O analista também dedicava muito tempo configurando as cláusulas (filtros) para recuperar padrões relevantes, e mesmo assim elas eram bem limitadas. Assim, ao final da fase de Análise sem um ambiente de apoio à interpretação e recuperação, o analista já se dava por satisfeito quanto encontrava pelo menos alguns padrões relevantes.

Para o entrevistado, a utilização do ambiente de apoio na fase de Análise de Padrões, possibilita que o analista facilmente interprete e recupere padrões potencialmente relevantes para o domínio. Ou seja, analista não necessita dispender tempo com detalhes técnicos, como por exemplo, se preocupar em conhecer como definir um filtro de interesse através de sintaxes de difícil manipulação; em compreender o significado das URL do projeto do *site*, etc.

De acordo com a visão do analista, o ambiente de apoio proposto neste trabalho demonstrou ser útil por: representar os padrões de forma intuitiva e de fácil compreensão; permitir aprofundar a interpretação do padrão de forma fácil e visual; descobrir padrões relacionados através das operações de *drill-down* e dos agrupamentos; possibilitar a

exploração de padrões inesperados através da navegação pelos agrupamentos de padrões, uma vez que nem sempre o analista tem em mente os caminhos realizados pelos estudantes; permitir a verificação de hipóteses, representadas facilmente através de filtros criados a partir de um glossário de termos (ontologia) e sem profunda preocupação com sintaxe.

Assim, com a utilização de um ambiente de apoio, os benefícios da fase de Análise de Padrões para o processo de MUW aplicado na EAD ficam mais visíveis, onde a partir dos padrões relevantes é possível aperfeiçoar a modelagem conceitual de ambiente educacional.

9 CONCLUSÕES E TRABALHOS FUTUROS

O presente trabalho centra-se na fase de Análise de Padrões, onde problemas enfrentados comprometem o resultado do processo de MUW. Os mecanismos propostos por esta pesquisa visam facilitar as atividades de interpretação e de recuperação de padrões seqüenciais de navegação fazendo uso do conhecimento representado pela Ontologia de Domínio.

Visando viabilizar os mecanismos propostos, os eventos de domínio são representados em dois níveis, a saber, Físico e Conceitual. Ainda, torna-se necessário o mapeamento entre estes níveis. A Ontologia de Domínio (nível Conceitual) e o mapeamento entre os níveis de representação dos eventos são explorados na fase de Análise de Padrões por permitir flexibilidade de interpretação e recuperação de padrões considerando diferentes dimensões de interesse, sem necessidade de retorno à fase de Preparação de Dados.

Os mecanismos de interpretação de padrões visam representar padrões seqüenciais físicos na forma de padrões seqüenciais conceituais de acordo com diferentes dimensões de interesse, e ainda permitir a análise exploratória dos padrões conceituais, através da operação de detalhamento de relacionamento, *roll-up* e *drill-down*.

Desta forma, os mecanismos de interpretação de padrões contribuem por: a) facilitar entendimento dos padrões seqüenciais amenizando o esforço do analista na interpretação do significado deste; b) não exigir do analista um profundo conhecimento do domínio, uma vez que a ontologia representa parte deste conhecimento; c) permitir aprofundar a compreensão do conhecimento suportado pelos padrões seqüenciais conceituais de forma interativa; d) descobrir conceitos e outros padrões relacionados de forma fácil e intuitiva. Uma comparação destes mecanismos de interpretação de padrões com as abordagens similares foi descrito na seção 5.3.

Os mecanismos de recuperação de padrões complementam os mecanismos de interpretação contribuindo na restrição do foco da busca por padrões relevantes, através da geração de agrupamentos ou da aplicação de filtros de interesse. A comparação destes mecanismos de recuperação de padrões em relação as abordagens relacionadas foi descrito na seção 6.4.

A geração de agrupamentos otimiza a atividade de inspeção *ad hoc* uma vez que a partir de poucos padrões o analista pode considerar ou desconsiderar todo o grupo de padrões, por estes terem características em comum. Já os mecanismos de recuperação de padrões através de filtros contribuem por: a) minimizarem a necessidade de aprendizado de uma sintaxe; b) não requerem do analista um profundo conhecimento do domínio para a sua definição, uma vez que exploram a Ontologia de Domínio para este propósito; c) suportarem diferentes tipos de restrições e serem aplicados considerando diferentes mecanismos de busca (equivalente ou aproximada).

Para avaliar os mecanismos propostos, foi definido um ambiente de apoio que disponibiliza estes mecanismos através das funcionalidades de um protótipo. O estudo de caso realizado possibilitou a comparação da fase de Análise de Padrões em dois processos de MUW aplicados no contexto da EAD, sendo que um destes processos utilizou o protótipo para apoiar a fase de Análise de Padrões. Como resultado, os mecanismos apresentaram visíveis vantagens relacionadas à interpretação e recuperação de padrões, confirmadas pelo analista que participou de ambos os processos.

Quanto a extensibilidade da abordagem proposta, cabe ressaltar que os mecanismos propostos foram avaliados considerando padrões obtidos através da aplicação do algoritmo *AprioriAll*, porém, estes mecanismos podem ser perfeitamente aplicados para classes associativas ou mesmo padrões seqüenciais retornados por algoritmos similares ao *AprioriAll*.

Quanto as limitações, a principal delas refere-se à dependência dos mecanismos na estrutura de como o conhecimento do domínio é representado e como os níveis de representação dos eventos do domínio são mapeados. A existência destas restrições se faz necessária para simplificar os mecanismos propostos, ou seja, a inexistência destas implica na extensão dos mecanismos de recuperação e interpretação.

Outras limitações referem-se: a impossibilidade de definir filtros de interesse que suportem qualquer tipo de expressão regular; a possibilidade de utilizar somente um cálculo para a medida de similaridade necessário para o mecanismo de busca por aproximação; e somente um critério para a geração dos agrupamentos de padrões.

Trabalhos futuros centram-se em estudar a viabilidade da integração dos mecanismos propostos com a Web Semântica já que esta é composta por uma camada que especifica as ontologias de domínio. O objetivo é explorar como esta camada suportaria os mecanismos propostos. Outro interessante trabalho futuro visa explorar a utilização de agentes para auxiliar os analistas na definição dos filtros de interesse de acordo com os padrões do uso da Web armazenados.

Outros trabalhos futuros propõem a validação empírica na PUC-Virtual; avaliação a aplicabilidade dos mecanismos em outros domínios; aplicabilidade de outros critérios de agrupamentos e outras medidas de similaridades para o mecanismo de busca por aproximação;

REFERÊNCIAS

[AGR93] Agrawal, R.; Imielinski, T.; Swami, A. N. "Mining association rules between sets of items in large databases". In: ACM SIGMOD International Conference on Management of Data, 1993, pp.207-216.

[AGR94a] Agrawal, R.; Srikant, R. "Mining sequential patterns". In: 11th International Conference on Data Engineering, 1994, pp.3-14.

[AGR94b] Agrawal, R.; Srikant, R. "Fast algorithms for mining association rules". In: International Conference on Very Large Data Bases, 1994, pp.487-499.

[BEC03] Becker, K.; Vanzin, M. "Discovering interesting usage patterns in web-based learning environments". In: International Workshop on Utility, Usability and Complexity of E-Information Systems, 2003, pp.57-72.

[BER97] Berry, M.; Linoff, G. "Data mining techniques: for marketing, sales, and customer support". New York , John Wiley & Sons, 1997, 454 p.

[BER00] Berendt, B.; Spiliopoulou, M. "Analysing of navigation behaviour in Web sites integrating multiple information systems". The VLDB Journal, vol. 9-1, May 2000 , pp.56-75.

[BER02a] Berendt, B.; Hotho, A.; Stumme, G. "Towards semantic Web mining". In: 1rst International Semantic Web Conference, 2002, pp.264-278.

[BER02b] Berendt, B.; Mobasher, B.; Nakagawa, M.; Spiliopoulou , M. "The impact of site structure and user environment on session reconstruction in Web usage analysis". In: 4th WebKDD Workshop, 2002, pp.159-179.

[BER02c] Berendt, B. "Using site semantics to analyze, visualize, and support navigation". Data Mining Knowledge Discovery, vol.6-1, Jan. 2002, pp.37-59.

[BLA03] Blanchard, J.; Guillet, F.; Briand, H. "Exploratory visualization for association rule rummaging". In: 4th International Workshop on Multimedia Data Mining, 2003, pp.107-114.

[BRI02] Brickley, Dan; Guha, R.V. "RDF Vocabulary Description Language 1.0: RDF Schema". Capturado em: http://www.w3.org/TR/rdf-schema/, June 2003, 24 p.

[CAB97] Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J.; Zanasi, A. "Discovering data mining: from concept to implementation". New Jersey: Prentice Hall, 1998, 224p.

[CHE96] Chen, M.; Park, J. S.; Yu, P. S. “Data mining for path traversal patterns in a Web environment”. In: 6th Conference on Distributed Computing Systems, 1996, pp. 385-392.

[COO97] Cooley, R.; Srivastava, J.; Mobasher, B. “Web mining: information and pattern discovery on the World Wide Web”. In: 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997, pp.558-567.

[COO99] Cooley, R.; Mobasher, B.; Srivastava, J. “Data preparation for mining world wide Web browsing patterns”. Journal of Knowledge and Information Systems, vol.1-1, Feb. 1999, pp.5-32.

[COO99a] Cooley, R.; Tan, P.; Srivastava, J. “Websift the web site information filter system”. In: Workshop on Web Usage Analysis and User Profiling, 1999, pp 163–182.

[COO03] Cooley, R. “The use of Web structure and content to identify subjectively interesting Web usage patterns”. ACM Transactions on Internet Technology, vol. 3-2, May 2003, pp. 93-116.

[DAI02] Dai, H.; Mobasher, B. “Using ontologies to discovery domain-level Web usage profiles”. In: 2nd Semantic Web Mining Workshop, 2002, 13p.

[FAY96] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. “The KDD process for extracting useful knowledge from volumes of data”. Communications of the ACM, vol. 39-11, Nov. 1996, pp.27-34.

[GAN03] Ganesan, P.; Garcia-Molina, H.; Widom, J. “Exploiting hierarchical domain structure to compute similarity”. ACM Transactions on Information Systems, vol. 21-1, Jan. 2003, pp. 64-93.

[GOL92] Goldberg, D.; Nichols, D.; Oki, B.; Terry, D. “Using collaborative filtering to weave an information tapestry”. Communications of the ACM, vol. 35-12, Dec. 1992, pp.61-70.

[GOL96] Goldberg, M. W.; Salari, S.; Swoboda, P. “World wide web-course tool: an environment for building WWW-based courses”. In: International World Wide Web Conference on Computer Networks and ISDN Systems, 1996, pp.1219-1231.

[GRU93] Gruber, T. “A translation approach to portable ontology specifications”. Knowledge Acquisition, vol.5-2, Sept. 1993, pp.199-220.

[HAN96] Han, J. et al. “DBMiner: A system for mining knowledge in large relational databases”. In: International Conference on Data Mining and Knowledge Discovery, 1996, pp.250-255.

[HAN97] Han, J. “Olap mining: an integration of olap with data mining”. In: IFIP Conference on Data Semantics, 1997, p.1-11.

[HAN00] Han, J.; Kamber, M. “Data mining: concepts and techniques”. San Francisco, Morgan Kaufmann Publishers, 2000, 550 p.

[HIP02] Hipp, J.; Guntzer, U. “Is pushing constraints deeply into the mining algorithms really what we want?: an alternative approach for association rule mining”. SIGKDD Exploration, vol. 4-1, June 2002, pp.50-55

[IBM04] IBM Software, I. “DB2 Intelligent Miner for Data”. IBM Software, 2002. Capturado em: http://www-3.ibm.com/software/data/iminer/fordata/index.html, October, 2004, 27p.

[ISL98] Integral Solutions Limited. “Clementine User Guide-Version S”, Integral Solutions Limited, 1998. Capturado em: http://www.spss.com/clementine/, Março 2004.

[KLE94] Klemettinen, M.; Mannila, H.; Ronkainen, P.; Toivonen, H.; Verkamo, A. “Finding interesting rules from large sets of discovered association rules”. In: Third ACM International Conference on Information and Knowledge Management (CIKM), 1994. pp.401-407.

[KOS00] Kosala, R.; Blockeel, H. “Web mining research: A survey”. SIGKDD Explorations, vol.2-1, June 2000, p.1-15.

[LEE01] Berners-Lee, T.; Hendler, J.; Lassila, O. “The semantic Web”. Scientific American, vol. 284-5, May 2001, pp.35-43.

[MAC03] Machado, L. “Mineração do uso da Web na Educação a Distância: proposta para a condução de um processo a partir de um estudo de caso”. Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2003, 103 p.

[MAN95] Mannila, H.; Toivonen, H.; Verkamo, A. I. “Discovering frequent episodes in sequences”. In: Proceedings of the 1rst International Conference on Knowledge Discovery and Data Mining, 1995, pp. 210-215.

[MAR04] Marquardt, C. “Apoio ao pré-processamento de dados da mineração do uso em ambientes de ensino na Web”. Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2004, 105 p.

[MED01] Medeiros, G.; Medeiros, M.; Vargas, R.; Herrlein, M.; Colla, A.; Franciosi, B.; Wagner, P. “Um cenário educacional para a PUCRS Virtual”. Colabora - Revista Digital da CVA-RICESU, vol. 1-1, Agosto 2001, 6 p.

[MOB96] Cooley, R.; Mobasher, B.; Srivastava, J. “Web mining: Pattern discovery from World Wide Web transactions”. Technical Reports, Department of Computer Science, University of Minnesota, 1996. 25 p.

[OBE03] Oberle, D.; Berendt, B.; Hotho, A.; Gonzalez, J. “Conceptual user tracking”. In: International Atlantic Web Intelligence Conference, 2003, pp.142-154.

[POH03] Pohle, C. “Integrating and updating domain knowledge with knowledge discovery”. 2003. In: 6th International Conference for Business Informatics, 2003, pp. 15-17.

[POH02] Pohle, C.; Spiliopoulou, M. “Building and exploiting ad hoc concept hierarchies for Web log analysis”. In: 4th International Data Warehousing and Knowledge Discovery Conference, 2002, pp.83-93.

[SAH99] Sahar, S. “Interestingness via what is not interesting”. In: 5th International Conference on Knowledge Discovery and Data Mining, 1999, pp.332-336.

[SMI04] Smith, M.; Welty, C.; McGuinness, D. “OWL Web Ontology Language Guide”. Capturado em http://www.w3.org/TR/owl-guide/, July 2004, 23p.

[SIL96] Silberschatz, A.; Tuzhilin, A. “What makes patterns interesting in knowledge discovery systems”. IEEE Transactions on Knowledge and Data Engineering, vol. 8-6, December 1996, pp.970-974.

[SPI98] Spiliopoulou, M.; Faulstich, L. “WUM: a Web Utilization Miner”. In: Workshop on the Web and Data Bases, 1998, pp.109-115.

[SPI02] Spiliopoulou, M.; Pohle, C. “Modelling and incorporating background knowledge in the web mining process”. In: Exploratory Workshop on Pattern Detection and Discovery, 2002, pp.154-169.

[SRI95] Srikant, R.; Agrawal, R. “Mining sequential patterns: generalizations and performance improvements”. In: International Conference on Extending Database Technology, 1996, pp.3-17.

[SRI97] Srikant, R.; Agrawal, R. “Mining generalized association rules”. In: 21th International Conference on Very Large Data Bases, 1995, pp.407-419.

[SRI00] Srivastava, J.; Cooley, R., Deshpande, M.; Tan, P. “Web usage mining: discovery and applications of usage patterns from Web data”. SIGKDD Explorations, vol.1-2, Jan. 2000, pp.12-23.

[STU02] Stumme, G.; Hotho, A.; Berendt, B. “Usage mining for and on the semantic Web”. In: National Science Foundation Workshop on Next Generation Data Mining, 2002, pp. 77-86.

[SUR02] Sure, Y.; Angele, J.; Staab, S. “Ontoedit: guiding ontology development by methodology and inferencing”. In: International Conference on Ontologies, Databases and Applications of Semantics (ODBASE), 2002, pp.1205-1222.

[SUZ98] Suzuki, R. C.; Bonfim, T. “Aplicações de recursos computacionais no Ensino à Distância”. In: IV Congresso RIBIE, 1998, 6p.

[VAN04] Vanzin, M.; Becker, K. “Exploiting knowledge representation for pattern interpretation”. In: Workshop on Knowledge Discovery and Ontologies, 2004. pp.61-71.

[VAN04a] Vanzin, M.; Becker, K. “Tutorial sobre mineração do uso da Web”. In: 19º Simpósio Brasileiro de Banco de Dados. Brasília, 2004.

[ZAI01] Zaiane, O. R. “Web usage mining for a better web-based learning environment”. In: IEEE International Conference on Advanced Learning Technologies, 2001, pp.450-455.

[W3C03] World Wide Web Consortium, “Web Characterization Activity”. Capturado em: http://www.w3.org/, March 2004.

[WCT02] WebCT. “Web course tools official page”. 2002. Capturado em: http://www.webct.com/, Abril 2004.

ANEXO I

Ontologia de Domínio representada

em OWL

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:vcard="http://www.w3.org/2001/vcard-rdf/3.0#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:jms="http://jena.hpl.hp.com/2003/08/jms#"
    xmlns="http://a.com/ontology#"
    xmlns:rss="http://purl.org/rss/1.0/"
    xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xml:base="http://a.com/ontology">
  <owl:Ontology rdf:about="">
    <owl:imports rdf:resource="http://protege.stanford.edu/plugins/owl/protege"/>
  </owl:Ontology>
  <owl:Class rdf:ID="Hotel">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="Acomodaçao"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Academia">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="Facilidade"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Evento"/>
  <owl:Class rdf:ID="Conteúdo">
    <rdfs:subClassOf rdf:resource="#Evento"/>
  </owl:Class>
  <owl:Class rdf:ID="Quadra-Tenis">
    <rdfs:subClassOf rdf:resource="#Facilidade"/>
  </owl:Class>
  <owl:Class rdf:ID="Localizar">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#Serviço"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Serviço">
    <rdfs:subClassOf rdf:resource="#Evento"/>
  </owl:Class>
  <owl:Class rdf:ID="Detalhar">
    <rdfs:subClassOf rdf:resource="#Serviço"/>
  </owl:Class>
  <owl:Class rdf:ID="Restaurante"/>
  <owl:Class rdf:ID="Reservar">
    <rdfs:subClassOf rdf:resource="#Serviço"/>
  </owl:Class>
  <owl:ObjectProperty rdf:ID="faz-parte">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Restaurante"/>
          <owl:Class rdf:about="#Acomodaçao"/>
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="#Conteúdo"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="disponibiliza">
    <rdfs:range rdf:resource="#Facilidade"/>
    <rdfs:domain rdf:resource="#Hotel"/>
  </owl:ObjectProperty>
```

```
  <owl:ObjectProperty rdf:ID="refere-se-a">
    <rdfs:range rdf:resource="#Hotel"/>
    <rdfs:domain rdf:resource="#Detalhar"/>
  </owl:ObjectProperty>
  <rdf:Description>
    <rdf:rest rdf:parseType="Collection">
      <owl:Class rdf:about="#Hotel"/>
    </rdf:rest>
    <rdf:first rdf:resource="#Facilidade"/>
  </rdf:Description>
  <Hotel rdf:ID="Blue-Tree"/>
</rdf:RDF>
```

ANEXO II

Questões Relacionadas a Entrevista

com o Especialista do Domínio

Entrevista com o Especialista na área de EAD

1. Qual a sua percepção sobre os mecanismos de interpretação de padrões do uso da *Web* em relação ao processo de MUW realizado sem utilização de um ambiente de apoio à fase de Análise de Padrões?

Sem um ambiente de apoio à fase de Análise de Padrões, o analista acabava dedicando muito tempo em tentar entender o significado dos padrões retornados do processo de mineração. O analista também ficava muito tempo envolvido em entender os conceitos do processo.

A fase de análise se tornava desgastante, assim, o analista já se dava por feliz quando encontrava um padrão relevante.

A utilização de um ambiente de apoio à interpretação é muito válida pois libera o analista da tarefa árdua que é tentar entender o significado dos padrões retornados pelo processo de mineração. Ainda, o ambiente de apoio permite que o analista se dedique ao objetivo do processo de MUW, que é encontrar padrões relevantes.

Com um ambiente de apoio fica mais visível os benefícios do processo de MUW, onde a partir dos padrões relevantes é possível aperfeiçoar a modelagem conceitual de um curso.

2. Qual a sua percepção sobre os mecanismos de recuperação de padrões do uso da *Web* em relação ao processo de MUW realizado sem utilização de um ambiente de apoio à fase de Análise de Padrões?

Sem um ambiente de apoio à fase de Análise de Padrões, o analista dispendia muito tempo configurando as cláusulas para encontrar padrões relevantes, e mesmo assim elas eram bem limitadas.

Com um ambiente de apoio, o analista não se prende a detalhes técnicos e se concentra nos objetivos. Os mecanismos de recuperação são úteis por possibilitarem a exploração por padrões inesperados através da navegação pelos agrupamentos de padrões. Nem sempre o analista tem em mente os caminhos realizados pelos estudantes.

Outra funcionalidade interessante é analisar padrões considerando diferentes níveis de detalhe, de forma fácil e visual.

Os mecanismos de recuperação também possibilitam a verificação de hipóteses, representadas facilmente através de filtros criados a partir de um glossário de termos (ontologia). Os padrões descobertos possibilitam adaptações no ambiente educacional de acordo com as expectativas dos estudantes.