PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CAREN MORAES NICHELE

USO DE AGRUPAMENTO DE INTERESSE E TRAJETÓRIA PARA CARACTERIZAÇÃO DE SESSÕES DE APRENDIZADO

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre, pelo Programa de Pós-graduação em Ciência da Computação (PPGCC) da Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Prof.ª Dr.ª Karin Becker

Porto Alegre
2006

Dados Internacionais de Catalogação na Publicação (CIP)

N594u Nichele, Caren Moraes
Uso do agrupamento de interesse e trajetória para caracterização de sessões de aprendizado / Caren Moraes Nichele. - Porto Alegre, 2006.
117 f.

Diss. (Mestrado) - Fac. de Informática, PUCRS
Orientadora: Prof.ª Dr.ª Karin Becker

1. Agrupamento de Informações (Informática). 2. Mineração de Dados (Informática). 3. World Wide Web. I. Título.

CDD 005.72

Ficha Catalográfica elaborada pelo Setor de Processamento Técnico da BC-PUCRS

PUCRS
Campus Central
Av. Ipiranga, 6681 - prédio 16 - CEP 90619-900
Porto Alegre - RS - Brasil
Fone: +55 (51) 3320-3544 - Fax: +55 (51) 3320-3548
Email: bceadm@pucrs.br
www.pucrs.br/biblioteca

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "***Uso de Agrupamento de Interesse e Trajetória para Caracterização de Sessões de Aprendizado***", apresentada por Caren Moraes Nichele, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Sistemas de Informação, aprovada em 23/08/2006 pela Comissão Examinadora:

Prof. Dr. Fernando Luís Dotti – PPGCC/PUCRS

Profa. Dra. Vera Lúcia Strube de Lima – PPGCC/PUCRS

Prof. Dr. Leandro Krug Wives – UFRGS

Homologada em 04/01/2008, conforme Ata No. 001 pela Comissão Coordenadora.

Prof. Dr. Fernando Luís Dotti
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P. 16 – sala 106 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@inf.pucrs.br
www.pucrs.br/facin/pos

AGRADECIMENTOS

À minha orientadora, professora Karin Becker, por sua toda sua dedicação, apoio e compreensão. Saiba que seu apoio nos momentos difíceis tornou este trabalho possível.

Ao meu marido, Alexandre Feijó, por sua compreensão e imensa paciência durante meus estudos. Saiba que seu carinho e encorajamento foram fundamentais durante este período de grandes mudanças nas nossas vidas.

Ao André da Fonte Lopes e Patrick Pantoja, bolsistas do projeto, pelo comprometimento e empenho na realização das tarefas de desenvolvimento relativas à implementação do ambiente de apoio definido como parte deste trabalho.

À Hewlett Packard Computadores Ltda por acreditar em minha capacidade e custear meus estudos durante os dois anos e meio de mestrado.

Ao departamento de Educação à Distância (PUCRS Virtual), por fornecer os dados e disponibilizar acesso ao especialista, permitindo assim o desenvolvimento deste trabalho.

“Pain is temporary, proud is forever”

Frase de autor desconhecido

RESUMO

Um dos principais problemas evidenciados no domínio da Educação a Distância (EAD) é a falta de percepção que os instrutores de cursos Web têm quanto à interação dos alunos durante o processo de aprendizado. Este problema é mais fortemente evidenciado no ambiente da EAD devido ao pouco contato entre os instrutores e os alunos, dadas as limitações dos canais de comunicação, e à falta de semântica no registro das páginas acessadas, em relação ao seu significado no domínio da aplicação.

A Mineração do Uso da Web (MUW) oferece técnicas de mineração de dados que permitem descobrir padrões de utilização da Web para melhor entender e servir as necessidades das aplicações. O processo de MUW é composto de etapas, a saber: pré-processamento, descoberta de padrões e análise de padrões. Várias técnicas podem ser aplicadas na etapa de descoberta de padrões. A técnica de agrupamento, foco deste trabalho, destaca-se por agregar valor nesta questão, pois tende a estabelecer grupos de usuários que mostram padrões de comportamento semelhantes. O agrupamento de sessões Web tem impulsionado uma grande área de pesquisa que visa caracterizar os usuários com base na navegação na Web. Porém, nenhum trabalho foi encontrado que aborde a similaridade entre as páginas considerando a semântica dos eventos da aplicação quando computando a similaridade entre as sessões Web. Além disso, a correta aplicação da técnica de agrupamento é uma tarefa complexa que envolve desde a preparação dos dados até a escolha do algoritmo de agrupamento, além de estar fortemente associada à complexidade do processo de descoberta de conhecimento.

Dados os problemas identificados, este trabalho propõe mecanismos de agrupamento e de interpretação de padrões que facilitem, respectivamente, a aplicação da técnica de agrupamento e a análise dos grupos por pessoas leigas, visando auxiliar na caracterização das sessões de aprendizado em um ambiente de EAD. Estes mecanismos fazem uso de uma taxonomia como forma de agregar semântica aos eventos do domínio, reduzindo assim a necessidade de retorno à etapa de pré-processamento. O mecanismo de agrupamento proposto visa facilitar a aplicação da técnica de agrupamento e aumentar a qualidade dos grupos, considerando para isso a similaridade entre as páginas com base na semântica dos eventos do domínio. O mecanismo de interpretação proposto permite representar os grupos visualmente, de modo condizente com o objetivo do agrupamento, bem como inspecionar dinamicamente os grupos formados considerando os diferentes níveis de abstração das páginas no domínio da aplicação. Foi desenvolvido um ambiente de apoio para auxiliar o intrutor durante a execução das etapas da MUW visando a facilitar a aplicação do agrupamento e a análise das sessões de aprendizado.

Palavras-Chaves: Agrupamento. Similaridade entre sessões. Mineração do Uso da Web.

ABSTRACT

The Web Usage Mining (WUM) applies data mining techniques to discover web usage patterns from Web server logs. The WUM process is composed by three major phases: pre-processing (where data is collected, cleaned and transformed), pattern discovery (in which mining algorithms are applied), and pattern analysis (where resulting patterns are analyzed).

The categorization of visitor's behavior based on their interaction in the web site is a key issue in WUM. In the E-learning area this topic becomes more relevant due to the lack of face-to-face contact between students and professors, given the physical distance, as well as the semantical gap between URLs and corresponding application events. Clustering, which subject of this research, is a mining technique that aims at grouping objects on basis of high inter-group similarity and low inter-group similarity. Several works leverage clustering techniques with the purpose of characterizing web user behavior during navigation. However, most of the works do not consider the meaning of visited URLs in the application domain, when measuring similarity between web sessions. Page semantics is frequently considered in the pre-processing phase, in data enrichment tasks, in which URLs are mapped into domain concepts. This approach is static in the sense that a new perspective of a URL (e.g. more generalized concept), to obtain better clustering results, often implies re-processing data. In addition to that, the correct clustering technique execution is a complex task which includes data preparation and transformation according to the mining objectives in such way interesting patterns can be found.

Considering these problems, this research proposes a clustering mechanism and an interpretation mechanism as a way to characterize student's behavior in a Web course. These mechanisms aim make the clustering technique execution and group analysis easy to a non data mining expert person. The proposed mechanisms are based in a domain taxonomy representing the domain events for addressing the semantic gap between URLs and application events. The clustering mechanism considers the similarity between visited pages as a way to improve the quality of clustering results. The proposed interpretation mechanism allows visualize the characteristics for each group, according to the clustering objective, as well as inspects groups dynamically considering the different levels of abstraction for application events in the domain taxonomy. These mechanisms establish the basis for categorization of web user behavior, for which a prototype was developed.

Key-words: Clustering. Web session similarity. Web Usage Mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CGI	Common Gateway Interface. É um padrão de comunicação entre as aplicações externas e servidores Web.
CLF	Common Log Format. Formato de armazenamento de acessos utilizado por servidores Web.
EAD	Educação a Distância.
HTML	HyperText Markup Language. Linguagem utilizada para construção de páginas na Web.
HTTP	Hypertext Transfer Protocol.
KDD	Knownledge Discovery in Database.
MUW	Mineração do Uso da Web.
PUCRS	Pontifícia Universidade Católica do Rio Grande do Sul.
Site	Tipicamente é um domínio que disponibiliza uma coleção de páginas através de um servidor Web.
URL	Uniform Resource Locator. Padrão de nomenclatura utilizado para identificar a localização de um objeto, tipicamente uma página Web.
Web	Abreviação de WWW.
WebCT	Web Course Tool. Ferramenta Web responsável pela criação da infra-estrutura e navegabilidade dos cursos de educação a distância instalados na PUCRS VIRTUAL [PUC06] à epoca de desenvolvimento desta dissertação.
WWW	World Wide Web.
XML	Extensible Markup Language.

SUMÁRIO

1 INTRODUÇÃO

A Web (WWW) vem crescendo rapidamente e o fluxo de acessos, o tamanho e a complexidade das páginas Web vêm acompanhando este crescimento na mesma proporção. A popularização do uso da Web como meio de pesquisa e informação é um dos fatores que contribui para este contínuo crescimento. Como conseqüência, podemos notar o aumento na complexidade das tarefas relacionadas à Web, tais como construção de páginas, infra-estrutura e planejamento de servidores, busca de informações, etc.

No domínio da Educação a Distância (EAD) baseada na Web, o contato entre os instrutores e os alunos não é tão intenso quanto em sala de aula, uma vez que grande parte do curso é ministrada de forma assíncrona e/ou distribuída na Web. Dadas as limitações no processo de aprendizado impostas pelo canal de comunicação, os instrutores têm dificuldade em avaliar o comportamento dos seus alunos durante o processo de aprendizado, bem como perceber se os materiais preparados para as aulas e os serviços oferecidos pelo ambiente de apoio ao curso estão sendo adequadamente utilizados. Os recursos estatísticos oferecidos por algumas ferramentas de gerenciamento de cursos Web (ex: WebCT[1], TelEduc[2], ATutor[3], AulaNet[4], etc) apresentam limitações analíticas que dificultam a real compreensão das sessões de aprendizado em cursos Web. As sessões de aprendizado, doravante denominadas simplesmente sessões, podem ser vistas como a seqüência de páginas acessadas por um mesmo aluno durante a navegação no curso Web.

Neste contexto, surge a Mineração Web que, através das técnicas de mineração de dados, vem auxiliar na extração de conhecimento da Web. A Mineração Web pode ser dividida em três classes [SRI00, MOB04]: Mineração do Conteúdo, Mineração da Estrutura e Mineração do Uso. A Mineração do Uso da Web (MUW) [SRI00], em particular, utiliza as técnicas de mineração com o objetivo de descobrir padrões de utilização da Web para melhor entender e servir as necessidades de aplicações Web. O processo de MUW é composto de etapas, a saber: pré-processamento (onde ocorre a coleta, limpeza, identificação e enriquecimento das sessões), descoberta de padrões (onde as técnicas de mineração são aplicadas) e análise de padrões (onde ocorre a interpretação dos resultados)[MOB04]. Todas as fases estão fortemente relacionadas, tornando o processo interativo e iterativo, sendo que o sucesso de uma fase é dependente do sucesso das anteriores. Assim, cada etapa deve ser desenvolvida de

[1] http://www.webct.com
[2] http://teleduc.cinted.ufrgs.br
[3] http://www.atutor.ca
[4] http://www.aulanet.uniovi.es

forma adequada, com objetivos condizentes com o que os dados disponíveis podem revelar, para que os resultados obtidos sejam válidos e passíveis de interpretação.

O grupo de Sistemas de Informação da PUCRS vem desenvolvendo trabalhos [MAC03b, MAR04b, TRI04, VAN04b, VAN05] na área de MUW voltados para a concepção e construção de ambientes de apoio para a análise e monitoração do processo de aprendizado na EAD, usando como estudo de caso os cursos da PUCRS VIRTUAL [PUC06]. Embora estes trabalhos abordem alguns dos principais problemas envolvidos durante as etapas da descoberta de conhecimento, não resolvem por completo a questão da caracterização e compreensão das sessões.

A técnica de agrupamento tem grande potencial e pode auxiliar na compreensão do processo de aprendizado na EAD, uma vez que as sessões são agrupadas de acordo com seu grau de similaridade. O agrupamento, ou “clustering”, agrupa as sessões em grupos de modo que as sessões dentro de um mesmo grupo tenham um alto grau de semelhança entre si, e que sejam diferentes das sessões pertencentes aos demais grupos [HAN00]. No contexto da MUW, existem dois grandes tipos de agrupamentos interessantes a serem descobertos [SRI00]: agrupamento de páginas e agrupamento do uso. O agrupamento de páginas descobre grupos de páginas que têm conteúdo relacionado, contribuindo assim com uma informação valiosa para as ferramentas de pesquisa na Web. Já o agrupamento do uso, ou agrupamento de sessões ou transações, tende a estabelecer grupos de usuários que mostram padrões de comportamento semelhantes. Por sua vez, o agrupamento do uso pode ser dividido, de acordo com os objetivos da MUW, em: agrupamento de interesse e agrupamento de trajetória. O agrupamento de interesse considera somente os acessos em comum entre as sessões. Já o agrupamento de trajetória leva em consideração o caminho percorrido pelos usuários durante a navegação na Web, ou seja, considera a seqüência e a re-visita das páginas acessadas nas sessões.

Vários trabalhos [MOB00a, MOB01, MOB02, HEE01, HEE02, WAN02, BAN00, BAN01, FU00] utilizam técnicas de agrupamento de sessões visando caracterizar os usuários com base em sua navegação na Web para tentar entender seu comportamento, ou tentar inferir seus próximos acessos com base em interesses comuns. Embora diversas formas de estabelecer a similaridade entre as sessões sejam propostas na literatura, estas não levam em conta a similaridade entre as páginas acessadas em relação à semântica dos eventos associados às páginas do domínio da aplicação. Além disso, a correta aplicação da técnica de agrupamento é uma tarefa complexa que envolve desde a etapa de pré-processamento (de modo a preparar e enriquecer os dados, bem como reduzir os ruídos e a dimensionalidade dos dados) até a etapa de descoberta de padrões onde é feita a escolha do algoritmo de

agrupamento (de modo a obter resultados condizentes com o objetivo do agrupamento), além de estar fortemente associada à complexidade do processo de descoberta de conhecimento como um todo. As tarefas executadas na etapa de pré-processamento ajudam a preparar as sessões, obtendo assim agrupamentos de melhor qualidade e de mais fácil interpretação. Tais tarefas são consideradas pontos críticos na MUW, e em especial no agrupamento, pois, a geração de uma nova perspectiva dos dados, para obter agrupamentos mais significativos, implica retorno à etapa inicial da MUW.

Dados os problemas identificados, este trabalho tem como objetivos: a) melhorar a qualidade dos padrões resultantes do agrupamento de sessões, considerando para isso a similaridade entre as páginas com base na semântica dos eventos associados às páginas do domínio da aplicação, b) facilitar a aplicação da técnica de agrupamento de acordo com o objetivo da MUW, bem como c) facilitar na interpretação dos grupos obtidos. Para tanto, este trabalho propõe mecanismos que facilitam a aplicação da técnica de agrupamento (mecanismo de agrupamento) e a interpretação dos resultados (mecanismo de interpretação), visando auxiliar na caracterização das sessões dos usuários. Estes mecanismos fazem uso de uma taxonomia como forma de agregar semântica aos eventos do domínio dinamicamente, reduzindo assim a necessidade de retorno à etapa de pré-processamento. O mecanismo de agrupamento proposto estende o agrupamento de sessões descrito por [BAN01] em dois aspectos: a) considera a similaridade entre as páginas durante o cálculo de similaridade entre as sessões, e b) permite lidar tanto com o agrupamento de interesse quanto com o agrupamento de trajetória. Além disso, permite, com base no objetivo da mineração, identificar quais tarefas de pré-processamento podem ser aplicáveis para a preparação das sessões. Já o mecanismo de interpretação proposto permite representar os padrões resultantes de maneira condizente com os objetivos da mineração bem como facilitar a interpretação dos mesmos considerando os diferentes níveis de abstração das páginas no domínio da aplicação.

Assim, a contribuição deste trabalho é mostrar como o agrupamento pode ser aplicado, de acordo com o objetivo de mineração e com a semântica dos eventos associados às páginas, de forma a facilitar a compreensão do comportamento dos alunos durante o processo de aprendizado em cursos Web.

Este documento está dividido em 9 capítulos. O capítulo 2 apresenta uma breve fundamentação teórica sobre a técnica de agrupamento. O capítulo 3 apresenta a MUW e os conceitos básicos que sustentam este trabalho. O capítulo 4 descreve os mais importantes trabalhos relacionados com agrupamento de sessões discutindo suas principais contribuições e limitações. O capítulo 5 apresenta os principais objetivos da

abordagem proposta, descreve os mecanismos de agrupamento e interpretação propostos, bem como algumas peculiaridades quanto à representação dos conceitos e das sessões necessárias à aplicação dos mecanismos propostos. Os capítulos 6 e 7 descrevem os mecanismos de agrupamento e interpretação propostos. O capítulo 6 descreve o mecanismo de agrupamento, detalhes sobre sua implementação, bem como considerações sobre sua efetividade através de uma análise comparativa com as abordagens existentes. O capítulo 7 descreve o mecanismo de interpretação, as facilidades oferecidas para visualização e inspeção dos resultados através do uso de abstrações em relação aos eventos da aplicação, bem como considerações sobre sua efetividade. O capítulo 8 apresenta a arquitetura do ambiente de apoio à caracterização de sessões de aprendizado onde os mecanismos de agrupamento e interpretação estão inseridos, bem como o protótipo desenvolvido. O capítulo 9 apresenta um estudo de caso realizado no contexto da EAD para avaliar os mecanismos propostos. Por fim, o capítulo 10 discorre sobre as conclusões, limitações, e trabalhos futuros. As referências bibliográficas pesquisadas e os anexos encontram-se no final deste documento.

2 AGRUPAMENTO

Este capítulo apresenta uma síntese dos principais conceitos relacionados à técnica de agrupamento, em particular, representação dos dados, medidas de distância, e categorias de técnicas de agrupamento.

2.1 Conceitos Básicos

De acordo com [HAN00], agrupamento é o processo que aglomera dados em grupos, de modo que os objetos dentro de um mesmo grupo tenham um alto grau de similaridade entre si, e que sejam diferentes dos objetos pertencentes aos demais grupos. O grau de similaridade entre os objetos é obtido utilizando uma medida de distância entre os objetos. A medida de distância calcula quanto os objetos estão próximos (similaridade) ou distantes (diferença) entre si, usando os atributos que os representam (também chamados “características”, ou ainda, “variáveis”).

2.1.1 Representação dos Dados

Os objetos a serem agrupados podem representar inúmeras entidades do mundo real, como: pessoas, animais, documentos, páginas Web, etc. A maior parte dos algoritmos de agrupamento utiliza dois tipos de estruturas para representar os dados manipulados: matriz de dados e matriz de correlação.

2.1.1.1 Matriz de Dados

Também conhecida como “objeto-por-variável”, esta matriz representa n objetos com p variáveis. A estrutura utilizada é uma matriz de tamanho n por p, ilustrada na Figura 1.

$$\begin{bmatrix} x_{1i} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Figura 1 – Matriz de dados

2.1.1.2 Matriz de Correlação

Também conhecida como “objeto-por-objeto”, esta matriz pode armazenar o conjunto de diferenças ou similaridades entre todos os pares de n objetos. A estrutura utilizada é uma matriz de tamanho n por n, onde d(i,j) é a medida de distância entre

os objetos i e j. Esta matriz é normalmente simétrica devido à propriedade de simetria da medida de distância. Ou seja, o triângulo superior da matriz é exatamente idêntico ao triângulo inferior da matriz.

Se a medida de distância considerar a proximidade entre os objetos então a matriz é dita “matriz de similaridade”. Na matriz de similaridade, ilustrada pela Figura 2, quanto mais d(i,j) se aproxima de 1, mais os objetos i e j se assemelham. Se os objetos são a mesma entidade, i=j, então d(i,j)=1. Se os objetos são entidades diferentes, i•j, e sua similaridade é d(i,j)=1, então estes objetos têm as mesmas características, mas não significa que são necessariamente idênticos.

$$\begin{bmatrix} 1 & & & & \\ d(2,1) & 1 & & & \\ d(3,1) & d(3,2) & 1 & & \\ \vdots & \vdots & \vdots & 1 & \\ d(n,1) & d(n,2) & \cdots & \cdots & 1 \end{bmatrix}$$

Figura 2 – Matriz de similaridade

Caso contrário, se a medida de distância considerar a diferença entre os objetos a matriz é dita “matriz de diferença”. Na matriz de diferença, ilustrada pela Figura 3, quanto mais d(i,j) se aproxima de 0, mais os objetos i e j se assemelham (ou estão “próximos” um do outro). Se os objetos são a mesma entidade, i=j, então d(i,j)=0.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & 0 & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Figura 3 – Matriz de diferença

2.1.2 Tipos de Dados

Os principais tipos de dados que caracterizam os objetos a serem agrupados estão classificados dentre as seguintes classes de variáveis [HAN00]:

- Variáveis contínuas: são medidas contínuas de uma escala linear, como por exemplo: peso, altura, latitude, temperatura, etc. Em um mesmo conjunto de dados podem existir variáveis contínuas que representam os dados utilizando unidades de medida diferentes;

- Variáveis binárias: têm somente dois valores possíveis: 0 e 1, onde 0 significa que a variável está ausente e 1 significa que a variável está presente. As variáveis binárias podem ainda ser classificadas como simétricas (0 e 1 têm o mesmo peso) ou assimétricas (0 e 1 têm pesos diferentes);
- Variáveis nominais: também denominadas discretas, são generalizações das variáveis binárias podendo assumir mais que dois estados. Estes estados podem ser denotados por letras, símbolos ou números inteiros, mas sem uma ordem específica entre eles;
- Variáveis ordinais: semelhantes às variáveis nominais, exceto pelo fato que os estados das variáveis ordinais seguem uma seqüência definida;
- Variáveis escalares: fazem uma medida positiva em uma escala não linear, como em uma escala exponencial;
- Variáveis mistas: misturam vários tipos de variáveis (contínua, binária, nominal, ou escalares). Este tipo de variável expressa bem a realidade, pois de maneira geral, no mundo real, as bases de dados podem apresentar todos os tipos de variáveis.

Assim, a matriz de dados bem como a medida de distância devem ser adequadas de acordo com o tipo de variáveis que representam os objetos a serem agrupados.

2.1.3 Propriedades das Medidas de Distância

Uma medida de distância calcula quão próximos ou distantes os objetos estão uns dos outros. Esta medida é dada através de variáveis que representam estes objetos, ou seja, depende dos tipos de dados envolvidos. A medida de distância deve obedecer aos seguintes princípios matemáticos, onde A e B são dois objetos e a distância entre eles é um número representado por d(A, B):

- Número não negativo: $d(A,B) \geq 0$;
- Simetria: $d(A,B)=d(B,A)$;
- Autosimilaridade: $d(A,A)=0$ considerando a diferença entre os objetos, ou $d(A,A)=1$ se considerada a proximidade entre os objetos;
- Separação: $d(A,B)=0$ considerando a diferença entre os objetos, ou $d(A,B)=1$ considerando a proximidade entre os objetos, somente se $A=B$;
- Desigualdade Triangular: $d(A,B) \leq d(A,C) + d(B,C)$.

Um estudo mais aprofundado sobre os tipos de dados e as medidas de distância apropriadas para cada tipo de dado é relatado em [NIC04a].

2.2 Categorias de Técnicas de Agrupamento

Existe um grande número de algoritmos de agrupamento disponíveis na literatura. A escolha de um determinado algoritmo depende de dois importantes fatores: o tipo de dado e o objetivo da mineração. De acordo com [HAN00], em geral, a maioria dos algoritmos de agrupamento, podem ser classificados dentre as seguintes categorias: particional (partitioning), hierárquica (hierarchical), baseada em densidade (density-based), baseada em grade (grid-based), e baseada em modelo (model-based).

As próximas seções apresentam resumidamente as categorias de agrupamento particional, hierárquica e baseada em grafo. Um estudo mais detalhado sobre as categorias particional e hierárquica, bem como seus principais algoritmos de agrupamento foi desenvolvido em [NIC04a]. Maiores detalhes sobre as demais categorias de agrupamento podem ser encontrados em [HAN00].

2.2.1 Particional

A partir de um conjunto de n objetos, um algoritmo de agrupamento particional classifica os elementos em k grupos, onde: 1) cada grupo deve conter ao menos um objeto, e 2) cada objeto deve pertencer a somente um grupo. Resumidamente, dado k (número de partições para construir), os métodos particionais criam uma partição inicial, e aplicam iterativamente uma técnica de realocação que tenta melhorar a partição criada movendo elementos de um grupo para outro. Os mais populares algoritmos particionais são k-Means e k-Medoids. Entre as principais dificuldades identificadas nesta categoria estão a definição do número ideal de partições, a dimensionalidade dos dados, e a busca pelo valor médio dos objetos ou o objeto que representa o centro de cada grupo.

2.2.2 Hierárquico

Um algoritmo de agrupamento hierárquico organiza os objetos em forma de árvore, fazendo desta forma uma composição hierárquica dado um conjunto de n elementos. Dependendo de como a composição hierárquica é formada, o método hierárquico pode ser classificado como:

- Aglomerativo: este algoritmo inicia com cada elemento formando um grupo separado e estes grupos são sucessivamente unidos com base em sua similaridade até que todos os grupos estejam unidos em um único grupo (nível mais alto da hierarquia) ou uma condição de término seja alcançada;
- Divisivo: este algoritmo inicia com todos os elementos no mesmo grupo e estes grupos são sucessivamente divididos em pequenos grupos até que cada elemento esteja inserido em um grupo ou até que uma condição de término seja alcançada.

Tipicamente o método hierárquico utiliza um dendograma (estrutura em forma de árvore) para representar a hierarquia dos grupos e o grau de similaridade para cada nível. Os quatro métodos mais utilizados na composição dos grupos hierárquicos são: Single Linkage, Complete Linkage, Group Average Linkage e Wards Linkage. Dentre os mais populares algoritmos hierárquicos pode-se citar: AGNES (AGglomerative NESting), DIANA (DIvisive ANAlysis) e BIRCH (Balanced Iteractive reducing and Clustering using Hierarchies) [ZAN96]. Maiores detalhes podem ser obtidos em [HAN01, TAN06].

2.2.3 Baseado em Grafo

O agrupamento baseado em grafo utiliza um grafo esparso onde cada elemento é representado por um vértice e suas ligações são definidas pelo valor da similaridade entre os elementos. Este modelo combina diferentes técnicas de agrupamento em duas fases distintas com o objetivo de obter melhores resultados:

- Primeira Fase: o grafo é construído e então particionado por um algoritmo de particionamento baseado em grafo.
- Segunda Fase: aplica um algoritmo hierárquico aglomerativo nos grupos identificados pela primeira fase, de modo a melhorar a qualidade dos mesmos.

Dentre os mais conhecidos algoritmos de agrupamento baseados em grafos pode-se citar Metis, hMetis e CHAMELEON [TAN06].

O algoritmo baseado em grafo hMetis, uma variação de Metis, recebe como parâmetro de entrada uma matriz de similaridade e modela os dados recebidos utilizando um grafo do tipo nearest-neighbor, onde cada sessão é representada por um vértice e está conectada somente às sessões mais similares. hMetis permite ainda especificar alguns parâmetros para a construção do grafo, como por exemplo, número máximo e mínimo de vizinhos para cada vértice, tipo de ligações entre os vértices,

valor limite para eliminação de ligações, etc. Após sua construção, o grafo é particionado em k grupos usando o algoritmo de particionamento “min-cut”.

2.3 Considerações

Considerando a MUW, o agrupamento pode ser utilizado para agrupar usuários com comportamentos similares durante a navegação do site Web, ou mesmo agrupar sessões similares para determinar comportamentos de navegação diferentes. Para tanto, é necessário estabelecer a medida de distância entre as sessões. A medida de distância escolhida depende de dois importantes fatores: o tipo de dado e o propósito da aplicação. Com base na similaridade entre as sessões, uma matriz de similaridade pode ser construída e utilizada como entrada para qualquer algoritmo de agrupamento.

Dentre os inúmeros algoritmos de agrupamento existentes, qualquer algoritmo poderia ser utilizado para agrupar as sessões de aprendizado independente do objetivo da mineração (agrupamento de interesse e agrupamento de trajetória). Entretanto, deve-se levar em conta as restrições e a aplicabilidade de cada categoria de agrupamento. Os algoritmos pertencentes à categoria particional utilizam, além da similaridade entre as sessões, o valor médio das sessões ou a sessão central do grupo. Dado que as sessões podem ter tamanho variado e re-visitas às páginas, tornando a tarefa de encontrar o valor médio ou sessão central do grupo muito complexa, os algoritmos da categoria particional são considerados alternativas de mais difícil aplicação para o agrupamento de trajetória. Já os algoritmos da categoria baseada em grafo podem ser utilizados independente do objetivo da mineração, uma vez que esta categoria requer somente o valor da similaridade entre as sessões como parâmetro para realizar o agrupamento. O mecanismo de agrupamento proposto por este trabalho adota a classe de agrupamento baseada em grafo, independente de algoritmo, por ser aplicável em ambos os objetivos de agrupamento (interesse e trajetória).

3 MINERAÇÃO DO USO DA WEB

Este capítulo apresenta os principais conceitos envolvidos na Mineração do Uso da Web, em particular aqueles necessários para o reconhecimento de sessões. Em seguida, detalha cada uma das etapas do processo de descoberta identificando, em cada uma delas, os principais problemas envolvidos.

A Mineração da Web utiliza técnicas da mineração de dados no contexto da Web e se divide em três categorias, de acordo com as fontes e tipos de dados envolvidos [COO99, SRI00]:

- Mineração do Conteúdo: consiste na descoberta de informações relevantes sobre o conteúdo das páginas Web;
- Mineração da Estrutura: consiste na descoberta de conhecimento a partir da organização do conteúdo e referências (links) entre as páginas;
- Mineração do Uso: consiste na descoberta de padrões de utilização das páginas Web. A Mineração do Uso da Web (MUW) está ligada à análise do comportamento do usuário durante sua navegação no site Web.

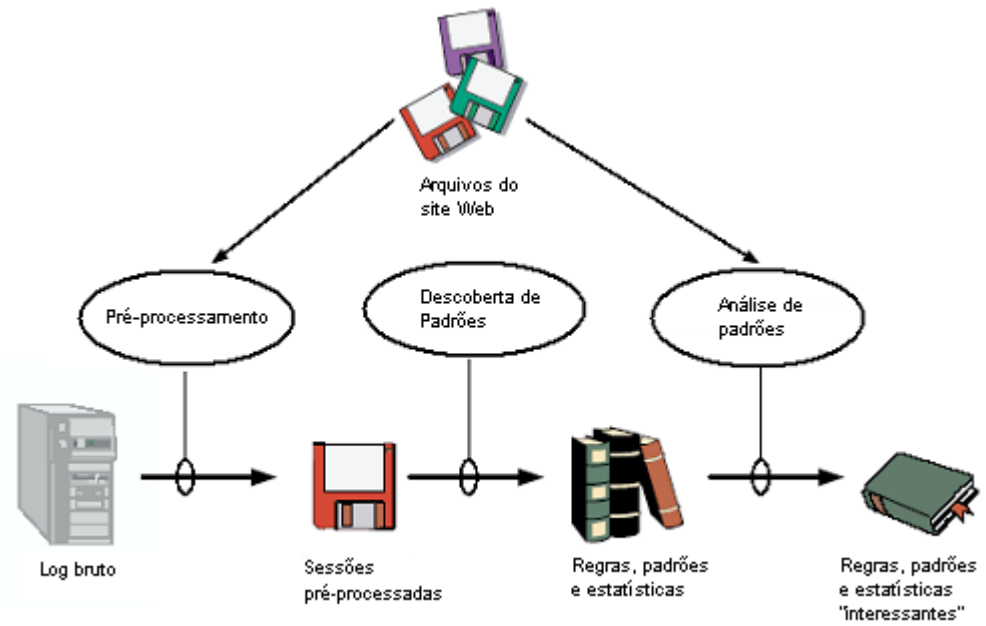

Figura 4 – Fases da Mineração do Uso da Web (adaptado de [COO99])

A Mineração do Uso da Web (MUW), foco deste trabalho, é descrita com maiores detalhes neste capítulo.

3.1 Processo da Mineração do Uso da Web

A Figura 4 ilustra as fases de pré-processamento, descoberta de padrões e análise de padrões, bem como os principais elementos envolvidos durante todo o processo da MUW. As principais tarefas envolvidas em cada uma das etapas da MUW são descritas detalhadamente nas seções seguintes.

3.1.1 Pré-processamento

A etapa de pré-processamento na Mineração do Uso da Web realiza a conversão dos dados relativos ao uso, acessos de páginas Web, em abstrações necessárias para a descoberta de padrões. A etapa de pré-processamento define tipos, modelos, e abstrações de dados com o objetivo de ajustar e melhorar a representação dos dados que serão posteriormente utilizados pelos algoritmos de mineração.

Geralmente esta etapa utiliza como fonte principal de dados arquivos de acesso provenientes dos servidores Web (log). Estes arquivos de acesso podem ser armazenados em diversos formatos, como por exemplo, Extended Common Log Format (ECLF) [W3C05]. A Tabela 1 ilustra um exemplo de arquivo de acesso que utiliza o formato ECLF. O formato CLF (Common Log Format) é mais simples e não contém os dois últimos campos (Referrer e Agent)

Tabela 1 – Exemplo de arquivo de acesso (formato ECLF)

Remote host	Auth	ID	Time/Date stamp	Method/URL/Protocol	Status	Size	Ref	Agent
16.127.37.124	aluno1	-	[20/Jul/2004:13:13:10 -0300]	“GET a.html HTTP/1.0”	200	2345	-	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)
15.20.17.2	-	-	[20/Jul/2004:13:13:10 -0300]	“GET b.html HTTP/1.0”	304	0	a.html	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)
16.127.37.124	aluno1	-	[20/Jul/2004:13:13:10 -0300]	"GET / HTTP/1.0"	200	567	-	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)

No formato ECLF o Remote host identifica o nome ou endereço IP da máquina de onde originou-se o acesso à página; Auth e ID correspondem à identificação do usuário, onde o sinal “-” indica acesso anônimo; Time/Date stamp armazena a data e hora do acesso; Method corresponde ao método utilizado (GET, POST ou HEAD); URL registra a página ou arquivo acessado; Protocol é a versão do protocolo HTTP utilizado; Status é o código de resposta do servidor Web à requisição do navegador; Size corresponde ao tamanho em bytes da página ou arquivo; Referrer identifica a URI que fez a referência à página acessada; e Agent identifica o navegador utilizado.

Entretanto, os dados contidos no arquivo de acesso não representam com total confiabilidade a real navegação dos usuários no site Web. Isso não se deve somente à freqüente falta de identificação dos usuários, mas também à ausência de registros de acessos feitos e à dificuldade de identificar o início e fim de uma sessão do usuário. O uso de cache e servidores proxy estão entre os fatores que mais contribuem para a carência de informações confiáveis nos arquivos de acesso. Detalhes sobre os principais problemas relacionados ao registro de acesso às páginas e à falta de informação no arquivo de acesso podem ser encontrados em [COO99].

Um outro aspecto importante relacionado com o arquivo de acesso é a lacuna semântica entre como é feito o registro das páginas acessadas neste arquivo e seu significado no domínio da aplicação [BER02, STU02]. Na prática, o acesso a uma página do ponto de vista do usuário pode ser registrado no arquivo de acesso do servidor Web através de várias requisições HTTP, correspondendo, por exemplo, aos diferentes elementos necessários para compor uma visão de página (ex: figuras, estilos, etc.) [COO99]. Da mesma forma, a utilização de um mesmo serviço pode ser registrada através de diferentes requisições HTTP, uma vez que diferentes parâmetros podem ser passados cada vez que o serviço é requisitado.

A seguir são descritos os passos primordiais da etapa de pré-processamento que têm como objetivo sanar a falta de informações no arquivo de acesso e melhorar a representação dos dados, conforme ilustra a Figura 5. Maiores detalhes sobre os problemas envolvidos, as tarefas realizadas e as principais técnicas utilizadas podem ser encontrados em [COO99, COO00, SRI00, MOB04].

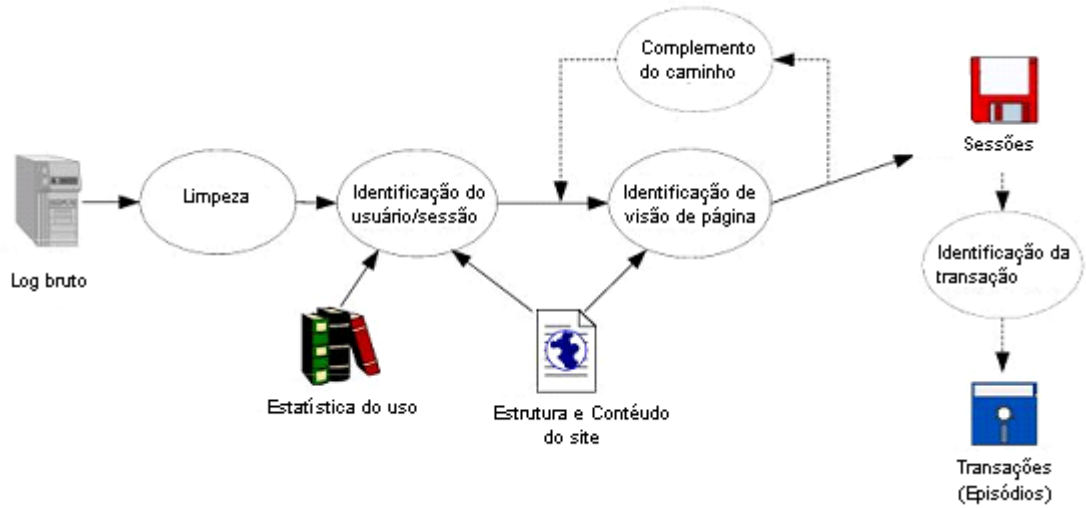

Figura 5 – Detalhes da etapa de pré-processamento do uso (adaptado de [COO00])

- Limpeza dos Dados: remove do arquivo de acesso as entradas desnecessárias ou irrelevantes. A eliminação de tais entradas é feita com base na extensão dos arquivos acessados como, por exemplo, entrada referente a

acesso de figuras, sons, estilos, animações, vídeos, páginas não encontradas, etc;

- Identificação do Usuário: identifica o usuário que acessou cada página. Esta identificação não necessita saber a identidade do usuário e sim poder distinguir entre diferentes usuários. Algumas técnicas para lidar com acessos anônimos são descritas em [COO99, SRI00, MOB04];
- Identificação de Sessões: separa as entradas contidas no arquivo de acesso em sessões individuais por usuário. Uma vez que um usuário pode visitar inúmeras vezes o mesmo site Web, deve-se considerar que o arquivo de acesso pode conter múltiplas entradas para um mesmo usuário. Deste modo, a identificação de sessão busca "quebrar" a seqüência de páginas acessadas (clickstreams) por um mesmo usuário. A sessão do usuário também pode ser controlada pelo servidor Web, neste caso é responsabilidade do servidor Web enviar cada URI com uma identificação de sessão [SRI00];
- Identificação de Visão de Páginas: identifica quais registros contidos no arquivo de acesso contribuíram para a formação e visualização de uma página no navegador do usuário. Esta identificação está fortemente relacionada à estrutura interna da página (hyperlinks para outras páginas ou arquivos), além de requerer um conhecimento detalhado da estrutura do site [MOB01]. Este tarefa fica ainda mais complicada com o uso de páginas com frames [COO03];
- Complemento dos Caminhos de Navegação: completa as entradas da sessão do usuário com as páginas de acesso que estão faltando devido ao uso de proxy ou cache pelo servidor;
- Identificação de Transações: refina as sessões de usuários em transações (ou episódios) menores considerando acessos semanticamente significativos dentro das sessões. Na MUW, esta definição é dependente do objetivo das aplicações às quais destina-se a análise. Este trabalho somente fará distinção entre sessão e transação quando pertinente e necessário ao entendimento do mesmo.

Além disso, as páginas de um site Web são ainda classificadas de acordo com a sua funcionalidade em páginas auxiliares e páginas de conteúdo e/ou serviço, permitindo assim definir diferentes granularidades na identificação de sessões [COO99]. As páginas auxiliares descrevem um tipo especial de página que apenas viabiliza a navegação até os conteúdos/serviços. Já as páginas de conteúdo/serviço, como o próprio nome sugere, são páginas que oferecem conteúdos e/ou serviços do

site Web. A utilização de páginas auxiliares e páginas de conteúdos é proposto por [COO99] como uma maneira de definir dois tipos de transações: transações de conteúdo e transações auxiliar-conteúdo. Na primeira são eliminados os acessos às páginas auxiliares. Na segunda, cada sessão é formada pelos caminhos utilizados até atingir um conteúdo/serviço Web.

3.1.1.1 Enriquecimento

O enriquecimento dos dados é uma tarefa clássica da etapa de pré-processamento, sendo necessária, por exemplo, para a identificação de transação ou identificação de visão de páginas. O enriquecimento dos dados pode ser realizado visando enriquecer as informações do usuário e/ou da sessão. Para tanto, é necessário um conhecimento do domínio da aplicação, da semântica das páginas, da topologia do site Web, bem como dos bancos de dados operacionais.

Considerando que os dados manipulados pela MUW são sessões provenientes de servidores Web, pode-se dizer que o principal problema encontrado na MUW está relacionado com a lacuna semântica entre como é feito o registro das páginas acessadas nos arquivos do servidor Web e seu significado no domínio da aplicação [BER02]. Ou seja, sem semântica, os mesmos conteúdos/serviços são tratados como diferentes requisições HTTP.

Neste contexto, surgem as abordagens semânticas que visam fornecer suporte às etapas da MUW através da representação do conhecimento do domínio, transformando os dados disponíveis em unidades significativas de eventos ao domínio da aplicação. Segundo [STU02], os eventos de aplicação podem ser classificados em eventos atômicos e eventos complexos. Os eventos atômicos podem ser:

- Eventos de conteúdo: descrevem acesso ao conteúdo disponível no site Web. Por exemplo, no contexto da EAD, eventos de conteúdo podem descrever: listas de atividades de aula, materiais oferecidos, textos, vídeos, etc;
- Eventos de serviço: descrevem acesso aos serviços disponíveis no site Web. Por exemplo, no contexto da EAD, eventos de serviço podem descrever: entrega de atividades ao professor, consulta à biblioteca, bate-papo, fórum, e-mail, etc. Estes acessos geralmente identificam requisições HTTP atendidas por aplicações Web, contendo a URL do serviço e os possíveis parâmetros, como ilustra a Figura 6.

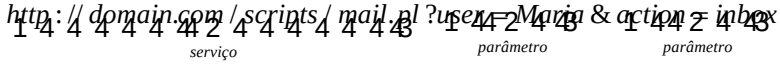

Figura 6 – Exemplo de evento de serviço

As abordagens semânticas podem ser classificadas quanto à representação do conhecimento do domínio utilizado durante o processo de integração com a MUW, a saber: taxonomias e ontologias [VAN04a]. Este trabalho adota o enriquecimento dos dados através da abordagem semântica de taxonomia como forma de auxiliar na identificação de características em comum entre os eventos da aplicação visando proporcionar uma melhora significativa no cálculo de similaridade entre as sessões. A taxonomia foi escolhida por se tratar de uma abordagem simples, capaz de agregar semântica aos eventos de aplicação com base nas relações hierárquicas entre as páginas do curso Web.

A abordagem semântica de taxonomia descreve uma maneira simples de representação de conhecimento, através de uma hierarquia conceitual formada de classes e sub-classes de conceitos ligados por relacionamentos de generalização/especialização conhecida como “hierarquia conceitual” [STU02]. Assim, a representação do conhecimento através de hierarquias conceituais está limitada a relacionamentos do tipo é-um. As hierarquias conceituais podem ser criadas automaticamente, pela organização dos conteúdos/serviços no site, com base em metadados da própria página, pela ontologia do site (ex: web semântica), ou mesmo manualmente com o conhecimento do especialista do domínio. A tradução das requisições HTTP para conceitos na hierarquia conceitual é dada pela dimensão de interesse desejado na abstração dos eventos de aplicação.

O uso da hierarquia conceitual nas etapas de pré-processamento e descoberta de padrões afeta diretamente a etapa de análise de padrões, uma vez que o mapeamento das requisições HTTP por conceitos em um nível da hierarquia conceitual e a redução da dimensão das sessões produzem melhores resultados no agrupamento e por conseqüência, facilitam a interpretação dos resultados. Entretanto, dependendo da dimensão de interesse escolhida durante a etapa de pré-processamento em relação à hierarquia conceitual, os resultados podem continuar de difícil interpretação.

Poucos trabalhos empregam o uso de taxonomia, ou mesmo ontologia, como apoio na análise exploratória durante a análise de padrões. Os poucos trabalhos existentes adotam o uso de taxonomias, que enriquecem os dados de maneira estática durante a etapa de pré-processamento (e.g. [BER00, DAI02, OBE03]). Os trabalhos de [VAN04a, VAN04b, VAN05] exploram as relações hierárquicas de uma ontologia estruturada para a visualização e análise dos padrões seqüenciais fornecendo subsídios para o analista interagir considerando os diferentes níveis de abstrações desta ontologia estruturada. A grande vantagem é que este enriquecimento é dinâmico, no sentido de que é o analista quem define o nível de abstração no qual deseja ver o conceito, e quando deseja obter conceitos mais ou

menos específicos. Com isto, padrões mais generalizados ou especializados são gerados sob demanda, de acordo com as necessidades de interpretação.

3.1.1.2 Representação das Sessões

Ao final da etapa de pré-processamento tem-se como resultando um conjunto de n conceitos, $P = \{p_1, p_2, \text{K}, p_n\}$, e um conjunto m de sessões de usuários (transações ou episódios). Conceitualmente, cada sessão s é representada como a seqüência natural de acessos com tamanho l de pares ordenados de conceito-peso $\left(p_j^s, w\left(p_j^s\right)\right)$, onde $j \in \{1, \text{K}, n\}$, considerando a ordem e re-visita das páginas [MOB04], como ilustra a Figura 7.

$$s = \left\{\left(p_1^s, w\left(p_1^s\right)\right), \left(p_2^s, w\left(p_2^s\right)\right), \text{K}, \left(p_l^s, w\left(p_l^s\right)\right)\right\}$$

Figura 7 – Representação das sessões para o agrupamento de trajetória

Este tipo de representação de sessão é utilizado quando o objetivo da mineração é descobrir padrões de trajetória dos usuários durante a navegação no site Web. Por outro lado, quando o objetivo da mineração é descobrir o interesse em comum dos usuários, cada sessão s pode ser vista como um vetor n-dimensional $\vec{s}$, onde a ordem dos acessos durante a navegação não é levada em consideração, como ilustra a Figura 8.

$$\vec{s} = \left\{w\left(p_1^s\right), w\left(p_2^s\right), \text{K}, w\left(p_n^s\right)\right\}$$

Figura 8 – Representação das sessões para o agrupamento de interesse

O peso atribuído a cada conceito, $w\left(p_j^s\right)$, pode variar com base no objetivo da mineração, a saber, binário ou tempo de visualização. O peso binário representa a existência ou não do conceito na sessão, já o peso pelo tempo de visualização determina quanto tempo o usuário demorou na sua visita ao conceito.

Ambos os tipos de pesos podem ser utilizados na representação das sessões para o agrupamento de interesse. Já na representação de sessões para o agrupamento de trajetória, o peso binário não é aplicável, dado que este tipo de representação contém somente os conceitos acessados durante a trajetória do usuário.

3.1.2 Tranformação das Sessões

Em complementação às tarefas convencionais de pré-processamento válidas para qualquer técnica de agrupamento, existe uma variedade de outras tarefas de transformação das sessões que podem ser realizadas de acordo com o objetivo da mineração [MOB01, MOB02, MOB04]. Estas tarefas que transformam os dados das sessões têm como objetivo reduzir os ruídos e a dimensionalidade dos dados, visando melhorar a qualidade dos resultados na MUW. No contexto do agrupamento na MUW os ruídos significam sessões que se mostram muito diferente das demais sessões identificadas, por exemplo, sessões muito pequenas, sessões muito grandes ou mesmo sessões com páginas desconhecidas.

A seguir são descritas as principais tarefas de transformações das sessões relacionadas às necessidades específicas das técnicas de agrupamento, a saber, filtro de importância, normalização, estatísticas do uso, e redução do caminho de navegação. Sendo as duas últimas particulares à MUW.

3.1.2.1 Filtro de Importância

Utilizar pesos binários é interessante devido à sua eficiência e facilidade em termos de armazenamento e cálculo de coeficientes de similaridade entre as sessões. Entretanto, o uso de pesos binários se torna ineficiente na identificação de padrões mais precisos de navegação. Por exemplo, um usuário pode acessar uma determinada página apenas para verificar seu conteúdo e saber que o mesmo não o interessa. Ou seja, embora o usuário tenha acessado uma página, esta pode não representar o real interesse do usuário se comparado seu tempo de acesso com as demais páginas acessadas na sessão. Assim, de acordo com [MOB01], a remoção de páginas da sessão que representem acessos irrelevantes ao interesse do usuário pode ser feita através de filtros de importância (Significance Filtering).

Entretanto, deve-se notar que o filtro de importância é relativo às características de navegação de cada usuário, da estrutura do site Web, bem como de conteúdo da página [MOB04]. Assim, o tempo de visualização a uma página gasto por um usuário detalhista e atento tende a ser maior que ao tempo gasto por um usuário mais dinâmico ou apressado. Além disso, o tempo de visualização de uma página auxiliar (páginas de menu, páginas de entrada, etc.) é menor do que o tempo de visualização de uma página de conteúdo ou orientada a produtos ou serviços.

Desta maneira, testes de significância estatística (Statistical Significance Testing) ajudam a capturar algumas das semânticas comportamentais descritas acima. Basicamente o peso associado a uma página na sessão deve ser 0 se o total de

tempo gasto nesta página é significativamente abaixo do tempo médio desta mesma página em todas as sessões nas quais ela está presente. Por exemplo, supondo que existem 6 páginas em um site Web: A, B, C, D, E e F, e a seguinte sessão s_1, ilustrada pela Figura 9-A, e um filtro de importância definido em 15 segundos, temos a sessão resultante s_1' conforme ilustra a Figura 9-B. Uma vez transformado, o vetor de sessão passa a conter somente as páginas que atingiram o limite de importância (ex.: páginas C e E).

(A) Sessão original

	A	B	C	D	E	F
s_1	11	0	22	5	127	0

(B) Sessão resultante

	A	B	C	D	E	F
s'_1	0	0	22	0	127	0

Figura 9 – Exemplo da aplicação do filtro de importância

3.1.2.2 Normalização

No contexto da MUW, a normalização dos pesos é uma tarefa que tem por objetivo tentar amenizar os fatores que potencialmente influenciam na distorção do tempo de visualização das páginas, tais como características físicas das páginas (e.g. tamanho físico dos arquivos que compõem a visualização da página), classificação de cada página (conteúdo ou auxiliar), bem como características navegacionais de cada usuário. Mobasher et al. [MOB01] apresentam dois tipos de normalização de pesos aplicados às sessões: normalização de sessões e normalização de páginas.

- Normalização de sessão: normaliza os pesos entre as páginas pertencentes a uma mesma sessão. Este tipo de normalização é útil para capturar a importância de uma página para um determinado usuário em relação às demais páginas por ele acessadas na mesma sessão;
- Normalização de página: normaliza o peso de uma página entre todas as sessões. Este tipo de normalização é útil para capturar o peso relativo da página associada a um usuário em relação aos pesos da mesma página para todos os demais usuários.

A normalização de sessão ou a normalização de página, ou sua combinação, pode ser aplicada não importando o tipo de representação das sessões.

3.1.2.3 Estatísticas do Uso

As estatísticas do uso oferecem ao analista informações que detalham as propriedades e características das sessões identificadas pela etapa de pré-

processamento. Assim, com base nestas informações o analista tem subsídios para aplicar filtros de suporte que possibilitam a remoção de sessões ou a remoção de páginas das sessões (Support Filtering [MOB00a]).

- Remoção de Páginas: a remoção de páginas remove os acessos às páginas com um determinado grau de suporte (URL Support [MOB00a]) das sessões. A remoção de páginas pode ser classificada em remoção de baixo suporte e remoção de alto suporte. A remoção de baixo suporte remove páginas com número igual ou inferior a uma porcentagem de acessos. Já a remoção de alto suporte remove páginas com número igual ou superior a uma porcentagem de acessos. Ambos os tipos de remoção de páginas podem ser combinados com o objetivo de obter um melhor resultado.

  A remoção de baixo e alto suporte não é aplicável quando utilizado o peso pelo tempo, dado que o tempo de acesso da página tem grande impacto no cálculo de similaridade entre as sessões. Na prática, a remoção de baixo suporte de 0% pode ser realizada quando utilizado o peso pelo tempo como forma de reduzir a dimensionalidade dos dados. Por exemplo, considerando as sessões ilustradas pela Figura 10-A e um suporte de 100%, remoção com alto grau de suporte equivale à remoção de uma mesma página acessada em todas as sessões como ilustra a Figura 10-B.

(A) Sessões Originais

	A	B	C	D	E	F
s_1	1	0	1	0	0	0
s_2	1	1	1	1	0	0
s_3	1	0	0	1	0	0

(B) Sessões Resultantes

	B	C	D	E	F
s'_1	0	1	0	0	0
s'_2	1	1	1	0	0
s'_3	0	0	1	0	0

Figura 10 – Exemplo da remoção de alto suporte

- Remoção de Sessões: remove as sessões muito pequenas (poucas páginas) ou grandes demais (muitas páginas) com o objetivo de reduzir os ruídos. Este tipo de remoção aplica-se para ambos os pesos binário e tempo de acesso. Para tanto, algumas das informações estatísticas são de suma importância para o analista, como por exemplo: tamanho da menor e maior sessão, média do tamanho das sessões e desvio padrão. Por exemplo, considerando 10 sessões onde a menor sessão tem tamanho 1, a maior sessão tem tamanho 51, a média de tamanho das sessões é 36, e o desvio padrão é de 19 páginas, pode-se eliminar as sessões menores que 17 e maiores que 50.

A remoção de páginas com baixa ou alta freqüência nas sessões e a remoção de sessões pequenas ou grandes demais contribui tanto para a redução efetiva da

dimensionalidade dos dados como também para a redução de ruídos [BAN01, FU00, MOB01], resultando em grupos de maior qualidade e mais fácil interpretação.

3.1.2.4 Redução do Caminho de Navegação

Quando considerado o caminho do usuário durante a navegação, a dimensão das sessões é representada pelo número de páginas acessadas. Assim, dependendo deste número, este tipo de representação pode apresentar alta dimensionalidade dos dados. A redução do caminho de navegação [BAN01, FU00] junta as páginas contíguas na seqüência e soma os tempos de visualização (quando utilizado o peso pelo tempo). A redução do caminho de navegação associado ao nível desejado de abstração das páginas no domínio da aplicação oferece resultados ainda mais interessantes, dado que o caminho pode sofrer uma redução mais significativa mantendo, contudo, as unidades significativas de eventos no mesmo. Por exemplo, considerando as sessões ilustradas pela Figura 11-A e peso pelo tempo de visualização, a redução do caminho de navegação é exemplificada pelas sessões na Figura 11-B onde as páginas contíguas são unificadas e seus tempos de visualização somados.

(A) Sessões Originais

	P1	P2	P3	P4	P5	P6
s_1	(A,10)	(A,10)	(A,10)	(C,10)	(A,10)	(A,10)
s_2	(A,10)	(D,10)	(B,10)	(B,10)	(C,10)	-
s_3	(D,10)	(D,10)	(A.10)	-	-	-

(B) Sessões Resultantes

	P1	P2	P3	P4	P5	P6
s'_1	(A,30)	(C,10)	(A,20)	-	-	-
s'_2	(A,10)	(D,10)	(B,20)	(C,10)	-	-
s'_3	(D,20)	(A,10)	-	-	-	-

Figura 11 – Exemplo de redução do caminho de navegação

3.1.3 Descoberta de Padrões

A mineração de dados oferece algoritmos desenvolvidos para inúmeras áreas, entre elas: estatística, mineração de dados, reconhecimento de padrões, inteligência artificial, etc. No contexto da MUW podemos citar: análise estatística, regras associativas, padrões seqüenciais, agrupamento, classificação, modelagem de dependências e regressão.

A técnica de agrupamento, foco deste trabalho, aplicada na etapa de descoberta de padrões oferece algoritmos que agrupam sessões que possuam características similares. No que se refere à MUW, existem dois tipos interessantes de agrupamento [SRI00, MOB04]: agrupamento do uso e agrupamento de páginas. O agrupamento do uso tende a estabelecer grupos de sessões de usuários que apresentem padrões de navegação semelhante, quer na sua trajetória, quer em seus

interesses. Por outro lado, o agrupamento de páginas descobre grupos de páginas que têm conteúdo relacionado.

O mapeamento das URLs para conceitos de uma hierarquia conceitual durante a etapa de pré-processamento, associado às tarefas de transformação das sessões, tende a aumentar a qualidade dos grupos resultantes, uma vez que a primeira agrega semântica aos acessos (melhorando o cálculo de similaridade) e a segunda reduz a dimensionalidade das sessões e os possíveis ruídos.

3.1.4 Análise de Padrões

A última fase da MUW é a análise de padrões. O principal objetivo desta fase é identificar somente os padrões relevantes encontrados na fase da descoberta de padrões. A metodologia utilizada para realizar esta tarefa é geralmente ditada pela aplicação à qual se destina a mineração de dados. A análise de padrões fornece informações úteis que podem ser aplicadas em diferentes áreas de aplicações da MUW como, por exemplo, personalização do site Web, reestruturação do site Web, ferramentas de recomendação, bem como caracterização do perfil dos usuários. Técnicas de visualização, tais como padrões gráficos, legenda de cores para identificar diferentes valores, ajudam a destacar padrões ou mesmo evidenciar tendências nos dados.

No contexto da MUW, mais especificamente no agrupamento de sessões, os padrões resultantes são grupos de sessões, onde cada sessão contém páginas visitadas pelo usuário. No agrupamento de trajetória, as abordagens de visualização existentes descrevem cada grupo através dos acessos que compõem cada sessão [BAN01], utilizando cadeias de Markov [MOB04], bem como através da árvore de navegação [GUN03]. No agrupamento de interesse, as abordagens de visualização existentes descrevem cada grupo através de vetores de atributos e pesos, ou através do perfil agregado [MOB01, MOB02] que representa cada grupo pela média consolidada das sessões que pertencem ao grupo.

Entretanto, a falta de um embasamento semântico para as páginas do site Web, aliada ao processo manual de interpretação dos resultados, é principal fator que torna a etapa de análise de padrões ainda mais árdua para o analista. Ou seja, a interpretação do significado dos acessos pertencentes às sessões do grupo, bem como a caracterização do grupo fica restrita aos conhecimentos do analista.

As abordagens de visualização existentes para os padrões de agrupamento de sessões e suas limitações são descritas com maiores detalhes no Capítulo 4.

3.2 Considerações

Este capítulo apresentou a Mineração do Uso da Web, bem como os principais elementos envolvidos durante todo o processo de descoberta de conhecimento. Dentre os principais problemas encontrados na MUW estão a pobreza e a ausência de informação relevantes nos arquivos de acesso, devido à natureza sintática dos arquivos de acesso. Este problema é evidenciado mais freqüentemente em sites Web com cache e/ou servidores proxy, páginas com frames, ou sem autenticação de usuários. Além disso, a falta de semântica no registro das páginas acessadas, em relação ao seu significado no domínio da aplicação, dificulta a aplicação do agrupamento e agrava ainda mais a pobreza dos resultados. Como conseqüência, devido ao trabalho exigido para suprir e prover as informações necessárias, a etapa de pré-processamento se torna sem dúvida a mais trabalhosa na MUW.

O objetivo da mineração dita a forma como as sessões serão representadas e utilizadas pelos algoritmos de agrupamento. Se o objetivo da mineração é descobrir padrões de trajetória durante a navegação, então as sessões são vistas como uma seqüência de páginas acessadas, com re-visitas e ordem entre os acessos. Quando o objetivo da mineração é descobrir os interesses em comum, as sessões são representadas como um vetor, onde a ordem dos acessos não é levada em consideração.

Após a identificação das sessões dos usuários, as sessões são transformadas com a finalidade de reduzir os ruídos e a dimensionalidade dos dados, visando grupos com mais qualidade.

A fase de mineração de dados oferece técnicas de mineração para diversas áreas. No contexto da MUW, existem dois tipos de agrupamentos interessantes: agrupamento do uso e agrupamento de páginas. O agrupamento do uso, foco deste trabalho, via estabelecer grupos de sessões de usuários que tenham padrões de navegação similares, tanto em sua trajetória quanto em seus interesses em comum.

Mesmo que o enriquecimento semântico seja aplicado na etapa de pré-processamento e utilizado durante a descoberta de padrões para o cálculo de similaridade entre as sessões, esta representação é estática no sentido que ela pode ser explorada somente na dimensão de interesse da hierarquia conceitual em que as requisições HTTP foram traduzidas. Ou seja, a necessidade de abstração dos acessos para outras dimensões de interesses na hierarquia ou a baixa qualidade dos grupos resultantes implica volta para a etapa inicial do processo de descoberta e re-execução de todas as etapas da MUW novamente.

4 TRABALHOS RELACIONADOS

Este capítulo apresenta os principais trabalhos relacionados ao agrupamento de sessões, descrevendo as abordagens de agrupamento, interpretação dos padrões resultantes do agrupamento, bem como trabalhos de MUW no contexto da EAD.

A maioria dos trabalhos de agrupamento de sessões foca na representação das sessões de acordo com a finalidade do agrupamento, ou no cálculo de similaridade entre as sessões. A representação das sessões tem grande impacto em como a similaridade é computada em cada abordagem.

As seções seguintes apresentam o agrupamento de interesse e o agrupamento de trajetória descrevendo como os principais trabalhos na literatura tratam o agrupamento das sessões e interpretação dos resultados em cada uma destas abordagens.

4.1 Agrupamento de Interesse

No agrupamento de interesse, foco dos trabalhos [FU00, HEE02, MOB01, MOB02], são considerados exclusivamente os acessos em comum entre os usuários. Ou seja, neste tipo de agrupamento de sessões, a trajetória do usuário não é levada em conta. Neste tipo de agrupamento, cada sessão é vista como um vetor no espaço n-dimensional de páginas do site Web, $P = \{p_1, p_2, \text{K}, p_n\}$, com um peso associado (binário ou tempo de acesso) a cada página, como ilustrado anteriormente na Figura 8. Dada esta representação, o conjunto das m sessões, $S = \{\vec{s}_1, \vec{s}_2, \text{K}, \vec{s}_m\}$, pode ser visto como uma matriz de dados $m \text{x} n$, onde m representa as sessões e n seus atributos.

O trabalho [HEE02] aborda ainda outras questões, como por exemplo, identificar o número ideal de grupos, e complementar as sessões com informações de várias fontes de dados provenientes da utilização, topologia e/ou conteúdo, estabelecendo o conceito de vetores de modalidade para cada sessão.

4.1.1 Similaridade entre as Sessões

O agrupamento de interesse utiliza a matriz de dados e emprega medidas de distância convencionais (ex: Euclidiana, Coseno, Coeficiente Jaccard) aplicadas aos vetores de atributos. Os trabalhos [MOB01, HEE02] utilizam o Coseno como medida de distância. Já [FU00], faz uso da medida de distância Euclidiana para computar a similaridade entre as sessões.

O enriquecimento dos dados durante a etapa de pré-processamento melhora o resultado do agrupamento, uma vez que os acessos a um mesmo conteúdo/serviço, antes tratados como diferentes requisições HTTP, são traduzidos por conceitos na hierarquia conceitual que representa o domínio. Entretanto, nenhum destes trabalhos utiliza explicitamente a similaridade entre os conceitos durante o agrupamento das sessões.

4.1.2 Interpretação dos Resultados

Além da interpretação convencional de agrupamento, a qual geralmente representa os grupos através dos vetores de atributos e pesos [FU00], os grupos resultantes do agrupamento de interesse podem ser representados por um perfil agregado [MOB01, MOB02].

Na interpretação convencional, mesmo utilizando o enriquecimento dos dados na etapa de pré-processamento, o analista deve avaliar os atributos e pesos atribuídos para as sessões pertencentes ao grupo de modo a tentar interpretar as características de formação do grupo. Conseqüentemente, a interpretação é dependente do conhecimento do especialista do domínio.

		A	B	C	D	E	F
Grupo 0	s8	0	0	1	1	0	0
	s4	0	0	1	1	0	0
	s7	0	0	1	1	0	0
Grupo 1	s0	1	1	0	0	0	1
	s3	1	1	0	0	0	1
	s6	1	1	0	0	0	1
	s9	0	1	1	0	0	1
Grupo 2	s2	1	0	0	1	1	0
	s5	1	0	0	1	1	0
	s1	1	0	1	1	1	0

(A) Grupos resultantes

Perfil Agregado	
Peso	Página
100 %	B
100 %	F
75 %	A
~~25 %~~	~~C~~

(B) Perfil Agregado do Grupo 1

Figura 12 – Exemplo de perfil agregado (adaptado de [MOB04])

No perfil agregado [MOB01, MOB02], os grupos são representados por perfis agregados baseados no vetor médio de cada grupo dado pelo método chamado PACT (Profile Aggregations based on Clustering Transactions), onde o peso de cada página neste vetor é dado pela razão da soma dos pesos desta página nas sessões pertencentes ao grupo em relação ao número total de sessões pertencentes ao grupo. Se utilizado o peso binário, então o peso das páginas no vetor agregado representa a

percentagem de sessões no grupo no qual a página foi acessada. Caso contrário, se for utilizado o peso por tempo de acesso, o peso das páginas no vetor agregado representa o tempo médio. As páginas resultantes no vetor do perfil agregado podem ser ordenadas pelo peso obtido e filtradas de acordo com um limite estipulado. Por exemplo, considerando as sessões ilustradas pela Figura 12-A e um valor limite de 75%, o perfil agregado obtido para o grupo 1 é ilustrado pela Figura 12-B.

4.2 Agrupamento de Trajetória

No agrupamento de trajetória [BAN01, GUN03, WAN02], o caminho do usuário durante a navegação na Web é considerado. Assim, o agrupamento de trajetória representa as sessões em termos das páginas visitadas, $P = \{p_1, p_1, \mathrm{K}, p_n\}$, pesos (binário ou tempo de acesso), ordem dos acessos, e re-visita das páginas, conforme ilustrado anteriormente na Figura 7. Dada esta representação, o conjunto das m sessões, $S = \{\vec{s}_1, \vec{s}_2, \mathrm{K}, \vec{s}_m\}$, pode ser visto como uma matriz de dados $m \times m$.

4.2.1 Similaridade entre as Sessões

O agrupamento de trajetória emprega técnicas de programação dinâmica (Dynamic-Programming-Based) [WAN02, GUN03] e variações destas [BAN01] para alinhar as seqüências de dados e então medir a similaridade entre as sessões.

Na programação dinâmica pura [WAN02, GUN03] é utilizado o melhor escore obtido pelo alinhamento entre cada par de sessões para construir uma matriz de similaridade entre as sessões, considerando possivelmente a similaridade entre as páginas e o tempo de visualização das páginas. Embora [WAN02] leve em conta a similaridade entre as páginas quando computando o escore de alinhamento entre as sessões, esta similaridade é dada por uma estrutura de símbolos que mapeia as URLs em termos da localização física na topologia do domínio. Assim, o cálculo de similaridade entre as páginas e a interpretação dos grupos resultantes é dependente da organização física das páginas. Entretanto, esta abordagem fica prejudicada para sites Web onde a localização física das páginas não segue uma organização semântica.

Na técnica conhecida por Weighted Longest Common Subsequences (WLCS) [BAN01], que representa uma variação da programação dinâmica, é utilizada uma medida de similaridade baseada na interseção entre duas sessões, denominada LCS (Longest Common Subsequence). O WLCS calcula a similaridade entre as sessões considerando o tempo de acesso às páginas pertencentes à subseqüência em comum e a importância do tempo gasto nesta subseqüência para ambas as sessões.

Entretanto, é importante notar que WLCS assume total igualdade entre as páginas para construir o LCS entre as sessões. O cálculo de similaridade entre sessões utilizando WLCS é detalhado a seguir na seção 4.2.1.1.

4.2.1.1 Weighted Longest Common Subsequences (WLCS)

WLCS [BAN01] estabelece a similaridade entre as sessões considerando dois fatores: a similaridade de sua região em comum com base no tempo de acesso às páginas, e a importância desta região dentro de cada sessão. WLCS encontra a intersecção entre duas sessões aplicando o algoritmo Longest Common Subsequence (LCS) [HIS77]. Assim, dada uma região em comum para duas sessões s_1 e s_2, WLCS obtém duas funções $l^{s_1}(i)$ e $l^{s_2}(i)$, onde $i = \{1, \mathrm{K}, L\}$ e L é o tamanho da intersecção, as quais mapeiam os índices das páginas que pertencem ao LCS no vetor que representa cada sessão. Considerando o tempo de acesso às páginas, WLCS obtém a similaridade entre as sessões combinando dois componentes: similaridade e importância.

- Componente de Similaridade: determina o quanto duas sessões são similares na sua região em comum (LCS). Para cada página pertencente ao LCS este componente calcula a razão entre o menor e maior tempo de visualização nas duas sessões (função min-max). Então, obtém a razão da soma dos pesos min-max pelo tamanho do LCS conforme ilustra a Fórmula 1.

$$S' = \frac{\sum_{i=1}^{L} \frac{\min\left(w\left(p^{s_1}_{l^{s_1}(i)}\right), w\left(p^{s_2}_{l^{s_2}(i)}\right)\right)}{\max\left(w\left(p^{s_1}_{l^{s_1}(i)}\right), w\left(p^{s_2}_{l^{s_2}(i)}\right)\right)}}{L} \quad (1)$$

- Componente de Importância: determina o quanto a região em comum (LCS) é importante para ambas as sessões. A importância do LCS é dada pela fração de tempo gasto no LCS em cada uma das sessões $\frac{T^{s_1}_{LCS}}{T^{s_1}}$ e $\frac{T^{s_2}_{LCS}}{T^{s_2}}$ respectivamente. A importância para ambas as sessões é calculado então pela média[5] destas duas frações de tempo, como ilustra a Fórmula 2.

$$S'' = \sqrt{\frac{T^{s_1}_{LCS}}{T^{s_1}} \times \frac{T^{s_2}_{LCS}}{T^{s_2}}} \quad (2)$$

A similaridade final entre as duas sessões, ilustrada pela Fórmula 3, é dada pelo produto dos componentes de similaridade e importância.

$$S = S' \times S'' \quad (3)$$

[5] [BAN01] utiliza a média geométrica.

4.2.2 Interpretação dos Resultados

Duas propostas foram encontradas na literatura para interpretação de agrupamentos de trajetória, a saber árvore agregada [GUN03] e espaço tridimensional [WAN02]. Estas técnicas complementam as técnicas convencionais, que representam cada grupo pelas sessões nele incluídas com os respectivos pesos.

Na interpretação convencional, abordada por [BAN01], a interpretação de cada grupo é dependente da interpretação dada pelo especialista do domínio, bem como da semântica utilizada no mapeamento das páginas para conceitos durante o enriquecimento dos dados na etapa de pré-processamento.

Quando utilizado um espaço tridimensional [WAN02], dois eixos correspondem aos grupos obtidos em ordem decrescente de tamanho, sendo que as sessões pertencentes aos grupos são também colocadas nestes mesmos eixos seguindo o mesmo ordenamento. O terceiro eixo identifica a similaridade entre as sessões. Assim, considerando o caso ideal, onde a similaridade entre sessões dentro do mesmo grupo é 1 e entre sessões de diferentes grupos é 0, pode-se observar que somente a diagonal do espaço tridimensional tem valores no terceiro eixo (eixo da similaridade). De acordo com [WAN02], a presença de diagonais no espaço indica bons agrupamentos, enquanto altos valores de similaridade fora da diagonal indicam agrupamentos inadequados. A utilização do espaço tridimensional [WAN02] oferece uma medida de qualidade para os agrupamentos obtidos, entretanto não oferece ao analista nenhuma facilidade para caracterização dos grupos.

Já a interpretação que utiliza a árvore agregada [GUN03] tenta caracterizar cada grupo resultante através do caminho de navegação das sessões pertencentes ao grupo. A árvore de navegação utilizada por [GUN03] estende a árvore agregada descrita pelo trabalho de Spiliopoulou et al. [SPI99], uma vez que inclui o tempo de acesso como uma informação adicional. A árvore agregada é composta de nodos, sendo que o nodo inicial (null) representa o início da navegação das sessões. Os demais nodos guardam quatro informações: a página, a ocorrência da página na trajetória, o tempo de acesso ([GUN03] assume pesos normalizados para cada nodo), e a representabilidade estatística da ocorrência da página na trajetória (número de sessões representadas pelo caminho da árvore até o nodo em questão). Por exemplo, a Figura 14 ilustra um exemplo de árvore agregada representando a trajetória de 5 sessões de usuários, ilustradas pela Figura 13, pertencentes a um mesmo grupo. A árvore mostra que 2 usuários iniciaram a visita pela página “A” e 3 pela página “B”. Dos usuários que iniciaram pela página “A”, 1 seguiu pelas páginas “B”, “E” e “F” consecutivamente. O outro usuário que iniciou pela página “A” seguiu pelas páginas “D” e “B”. Dos outros 3 usuários que iniciaram pela página “B”, 1 passou pelas

páginas “C” e “E”, e 2 passaram por “D” e “B”, sendo que destes últimos, um acabou a sua trajetória pela página “C” e o outro pela página “E”.

	Sessões
Grupo 1	$s_1 = \{(A,10),(B,10),(E,10),(F,10)\}$ $s_2 = \{(A,50),(D,10),(B,10)\}$ $s_3 = \{(B,10),(D,10),(B,10),(C,10)\}$ $s_4 = \{(B,50),(D,50),(B,100),(E,10)\}$ $s_5 = \{(B,10),(C,10),(E,10)\}$

Figura 13 – Sessões pertencentes a um mesmo grupo

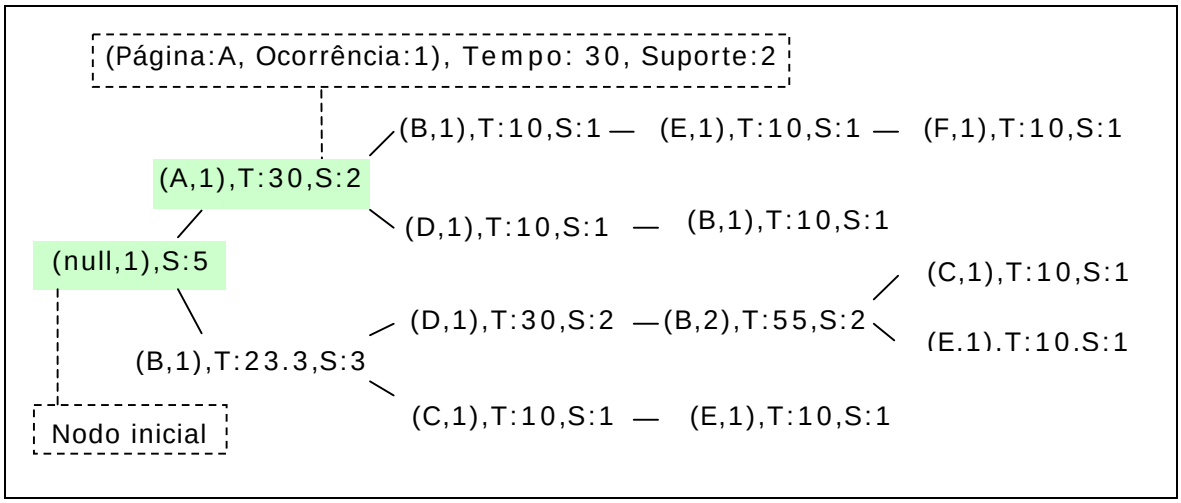

Figura 14 – Representação de um grupo utilizando árvore agregada

Embora a árvore agregada permita ao analista interpretar o grupo através da análise da trajetória entre as sessões pertencentes a um mesmo grupo, esta representação é estática. Ou seja, a representação de cada grupo pela árvore agregada pode ser explorada somente na dimensão de interesse da hierarquia conceitual em que as requisições HTTP foram traduzidas. Além disso, dependendo da diversidade e do tamanho das sessões resultantes do agrupamento a árvore agregada pode se tornar bastante complexa, com muitas ramificações, impedindo o analista de inferir quais propriedades efetivamente caracterizam o grupo.

4.3 Ambientes de Apoio ao Uso da Web no Domínio da EAD

O grupo de Sistemas de Informação da PUCRS vem desenvolvendo pesquisas e trabalhos na área de MUW no contexto da EAD, usando cursos da PUCRS VIRTUAL como estudo de caso [BEC03, MAC03a, MAC03b, MAR04a, MAR04b, VAN04a, VAN04b, VAN05, TRI04]. Os estudos de caso realizados no domínio da PUCRS VIRTUAL

mostram que os professores têm certa dificuldade em avaliar o aprendizado dos alunos e em identificar a interação dos alunos neste ambiente virtual e distribuído [MAC03b, BEC03].

O trabalho de Marquardt [MAR04a, MAR04b] trata dos principais problemas relacionados à etapa de pré-processamento dos dados aplicados à MUW em ambientes de ensino a distância. Para estes, desenvolve mecanismos de apoio que possibilitam a seleção de diferentes fontes de dados, e auxiliam na limpeza, transformação e enriquecimento dos dados de maneira a alinhá-los com os diferentes objetivos aplicados à mineração de dados na EAD. Como resultado, este trabalho implementou um protótipo customizável, chamado de LogPrep. LogPrep é organizado em termos de tarefas de pré-processamento, e cada tipo de tarefa pode ter um ou mais operadores que empregam uma técnica específica de pré-processamento para implementar cada tarefa. A customização de LogPrep é dada pela possibilidade de incorporar ao mesmo, de maneira muito simples, novas tarefas de pré-processamento, bem como novos operadores.

A fase de mineração propriamente dita é endereçada por [TRI04], que implementa um ambiente de descoberta de padrões seqüenciais, o qual inclui também mecanismos de visualização destes padrões. Este ambiente é uma implementação dos conceitos descritos em [SPI99]. A fase de mineração também é foco do trabalho de [WAN02], apresentado no capítulo 4.2, o qual mostra que domínios de EAD podem necessitar de soluções específicas para abstrair os eventos do domínio, diferentemente das soluções propostas no contexto de aplicações de comércio eletrônico. O trabalho de [WAN02] explora a topologia de um ambiente de EAD, dada por uma hierarquia de símbolos que mapeia a hierarquia conceitual do site Web, para calcular a similaridade entre as sessões. Como já salientado, esta abordagem não seria aplicável em infraestruturas genéricas de apoio ao projeto de gerência de cursos baseados na Web, como WebCT, que tem uma estrutura interna própria conveniente à organização das funcionalidades do ensino, e não baseada na semântica dos conteúdos e serviços oferecidos.

Já os trabalhos de Vanzin [VAN04a, VAN04b, VAN05] concentram-se na etapa de análise de padrões seqüênciais de navegação aplicada à MUW em ambientes de EAD. Os trabalhos exploram a representação do conhecimento no contexto da EAD descrito através de ontologias estruturadas para auxiliar a etapa de análise de padrões. Os trabalhos utilizam uma ontologia estruturada como forma de enriquecer os dados, representando os padrões seqüenciais de navegação através de padrões conceituais, e gerar padrões generalizados considerando diferentes níveis de abstrações, os quais são direcionados à interpretação de regras seqüenciais. Além

disso, os trabalhos fornecem subsídios para o usuário interagir dinamicamente na visualização dos padrões resultantes considerando os diferentes níveis de abstrações desta ontologia estruturada.

4.4 Considerações

Em geral, os trabalhos analisados apresentam conceitos e heurísticas para as fases da MUW (pré-processamento, descoberta de padrões e análise de padrões), cada qual voltado às tarefas e transformações necessárias para atingir seus objetivos motivadores considerando o agrupamento de sessões. Entretanto, nenhum destes trabalhos aborda o uso da similaridade entre as páginas considerando a semântica dos eventos da aplicação quando computando a similaridade entre as sessões ou na interpretação dos padrões resultantes do agrupamento.

Em termos de preparação dos dados, as abordagens existentes utilizam uma variedade de tarefas de transformação nas sessões, como por exemplo, abstração das páginas acessadas para eventos de aplicação, filtro de suporte, filtro de importância, normalização do peso da página, redução da dimensionalidade das sessões, em adição às tarefas de pré-processamento clássicas existentes [COO99]. Todas estas tarefas são importantes para o agrupamento de sessões, mas a aplicação de algumas destas tarefas estão relacionadas ao objetivo do agrupamento, sendo necessário para tanto identificar quais tarefas de preparação são aplicáveis e/ou mandatórias e quais são opcionais, de acordo com o objetivo de agrupamento.

No contexto da EAD, embora os trabalhos existentes abordem alguns dos principais problemas envolvidos durante as etapas da descoberta de conhecimento, não resolvem por completo a questão da compreensão e caracterização das sessões de aprendizado. Assim, o agrupamento pode agregar valor nesta questão, uma vez que as sessões dos alunos são agrupadas de acordo com o grau de similaridade dos padrões de acesso ao site Web de apoio ao curso.

O agrupamento de sessões descrito em [BAN01] é utilizado como base do mecanismo de agrupamento proposto neste trabalho uma vez que é possível adaptar o algoritmo que descobre a região de interseção entre duas sessões para considerar a similaridade entre os conceitos quando construindo o LCS. Além disso, o cálculo de similaridade de WLCS é generalizado pelo mecanismo de agrupamento proposto, com a devida preparação das sessões, para o agrupamento de interesse. Com isso, pode-se reduzir ainda mais a dimensionalidade dos dados, principal problema observado nas medidas de distância convencionais aplicadas aos vetores de atributos, uma vez que o WLCS calcula a similaridade entre as sessões com base na região em comum entre as sessões. Assim, o cálculo de similaridade de WLCS se torna uma medida mais

atrativa e vantajosa para realizar o agrupamento de interesse se comparada com as medidas convencionais. No Capítulo 5, a seção 5.5 detalha o processo para preparação das sessões que viabiliza a aplicação do mecanismo proposto para ambos os tipos de agrupamento (interesse e trajetória). O Capítulo 6 apresenta o mecanismo de agrupamento proposto que descreve entre outros itens o cálculo de simialridade entre as sessões (sim_WLCS) e o algoritmo sim_LCS, o qual estende o algoritmo LCS para considerar a similaridade entre os conceitos levando em consideração a semântica dos eventos da aplicação formalizada por uma hierarquia conceitual.

Já as abordagens de análise de padrões para o agrupamento de interesse [MOB01, MOB02] e agrupamento de trajetória [GUN03] servem de base para o mecanismo de interpretação proposto, às quais são acrescidos subsídios de interação para facilitar a interpretação dos grupos. O enriquecimento dinâmico explorado pelos trabalhos de [VAN04a, VAN04b, VAN05] para a visualização dos padrões seqüênciais é interessante, pois permite que o analista interaja considerando os diferentes níveis de abstração com que os padrões podem ser traduzidos considerando uma ontologia estruturada. Contudo, esta abordagem não pode ser aplicada diretamente à visualização dos grupos resultantes da técnica de agrupamento, pois a definição do nível de abstração da página na hierarquia conceitual está diretamente ligada ao grau de similaridade entre os conceitos e conseqüentemente ao modo como as sessões são agrupadas.

5 USO DO AGRUPAMENTO DE INTERESSE E TRAJETÓRIA PARA CARACTERIZAÇÃO DE SESSÕES DE APRENDIZADO

Este capítulo apresenta os principais objetivos da abordagem proposta, descreve sucintamente os mecanismos de agrupamento e interpretação propostos. Também são apresentadas algumas peculiaridades quanto à representação dos conceitos e das sessões necessárias à aplicação dos mecanismos propostos.

5.1 Objetivos

O objetivo principal deste trabalho é propor mecanismos de agrupamento e interpretação de padrões de navegação que facilitem, respectivamente, a aplicação da técnica de agrupamento e a análise dos grupos por pessoas leigas, visando auxiliar na caracterização das sessões de aprendizado em um ambiente de EAD.

Estes mecanismos buscam amenizar duas dificuldades existentes no agrupamento de sessões. A primeira se refere à complexidade envolvida na correta aplicação da técnica de agrupamento de acordo com o objetivo da mineração por uma pessoa leiga, associada ainda à complexidade do processo de descoberta de conhecimento como um todo. A segunda se refere à baixa qualidade do agrupamento das sessões, conseqüência de fatores como, por exemplo, a falta de semântica dos eventos de aplicação, a presença de ruídos, bem como a alta dimensionalidade dos dados. Assim, dentre os objetivos específicos deste trabalho está propor mecanismos que:

- Considerem a representação conceitual das páginas através de uma taxonomia como forma de agregar semântica aos eventos do domínio em diferentes etapas do processo de MUW;
- Facilitem a aplicação da técnica de agrupamento por pessoas leigas, permitindo, com base no objetivo da mineração, identificar quais tarefas de pré-processamento podem ser aplicáveis para a preparação das sessões;
- Permitam melhorar a qualidade do resultado do agrupamento das sessões considerando para isso a similaridade entre os conceitos durante a etapa de descoberta de padrões;
- Facilitem a interpretação dos grupos através da definição de modos distintos de visualização e da inspeção dinâmica dos grupos através do nível de interesse de abstração dos conceitos.

Complementam estes objetivos:

- Definir um ambiente de apoio à execução das fases de preparação dos dados (pré-processamento), agrupamento (descoberta de padrões) e interpretação dos grupos (análise de padrões) que incorpore os mecanismos propostos, e permita uma análise da efetividade dos mesmos;
- Desenvolver um protótipo que implemente os mecanismos propostos.

Uma visão geral dos mecanismos propostos e seus pressupostos é descrita nas próximas seções deste capítulo. Os pressupostos se referem aos requisitos necessários à aplicação da abordagem proposta em relação à representação do conhecimento do domínio e às peculiaridades das etapas de processo de MUW.

5.2 Representação Conceitual de Eventos e Nível Conceitual de Interesse

Este trabalho propõe a utilização de uma semântica de taxonomia representada por uma hierarquia conceitual, previamente definida por um especialista do domínio, como forma de agregar semântica aos eventos de aplicação. Este trabalho restringe o relacionamento de generalização de um conceito somente para um conceito ascendente, ou seja, não permite que um mesmo conceito seja descendente de dois conceitos distintos, como forma de reduzir a complexidade envolvida no cálculo da similaridade entre os conceitos.

Além disso, este trabalho adota a mesma classificação para os eventos de aplicação descrita por Stumme et al. [STU02] restrita, contudo, aos eventos atômicos de conteúdo e serviço. Neste contexto, os eventos da aplicação são representados em dois níveis: nível físico e nível conceitual [VAN04a]. O primeiro é representado pelas URLs que pertencem ao site Web, e o segundo pela hierarquia conceitual do domínio. A ligação entre estes dois níveis é definida pelo mapeamento das URLs que representam eventos da aplicação para conceitos na hierarquia conceitual. O mapeamento é dito de nível conceitual base quando as URLs do nível físico são mapeadas diretamente para conceitos no nível conceitual, dada a interpretação do especialista do domínio em relação aos eventos da aplicação. Considerando que os conceitos da hierarquia estão ligados por relacionamentos de generalização/especialização, o mapeamento do nível físico pode ser abstraído para qualquer nível de generalização da hierarquia conceitual, usando recursivamente os conceitos ascendentes do nível conceitual base. Este tipo de mapeamento é chamado mapeamento de nível conceitual de interesse. Ou seja, as URLs são mapeadas inicialmente para o nível conceitual base e então os conceitos são generalizados recursivamente através dos seus conceitos ascendentes até chegar o nível desejado de abstração.

A Figura 15 ilustra os eventos de aplicação segundo o nível físico e o nível conceitual, para um curso de EAD na Web hipotético onde os usuários podem ver o material das aulas, participar de fóruns, trocar e-mails, etc. O mapeamento do nível físico para o nível conceitual base é ilustrado na Figura 15 pela tradução das URLs www.ead.com/scripts/email.pl e www.ead.com/temas.html, que representam o nível físico, respectivamente para o conceito de serviço “Compor mensagem” da ferramenta de correio eletrônico e para o conceito de conteúdo “Temas” do material de apoio do curso, que representam por sua vez o nível conceitual base (nível conceitual de interesse 0). Uma vez mapeados para o nível conceitual base, os eventos da aplicação podem ser abstraídos para qualquer nível conceitual de interesse obedecendo aos relacionamentos de generalização da hierarquia conceitual. Isto é ilustrado na Figura 15 pela tradução do serviço “Compor mensagem” da ferramenta de correio eletrônico para o conceito “Correio” que é seu ascendente direto (nível conceitual de interesse 1), ou ainda por “Ferramentas de comunicação” se considerado dois níveis de ascendentes como nível conceitual de interesse. O nível conceitual de interesse de um conceito é dado pelo número de ascendentes usados para generalizá-lo na hierarquia, obedecendo ao número máximo de ascendentes do mesmo ou ao nível máximo de abstração na hierarquia conceitual. O nível máximo de abstração, definido pelo especialista do domínio com base na semântica dos eventos da aplicação, visa limitar a generalização dos conceitos a um determinado nível na hierarquia conceitual. Por exemplo, se não for interessante para o objetivo do agrupamento abstrair os conceitos até os níveis conceituais de interesse mais generalizados na hierarquia (e.g.: “Eventos”, “Conteúdo” e “Serviço”), então se pode dizer que o nível máximo de abstração dos conceitos na hierarquia é dado pelo seu número de ascendentes menos dois níveis de ascendente, como ilustra a Figura 15. Se o nível conceitual de interesse a ser considerado for maior que o número de ascendentes do conceito até seu nível máximo de abstração, então o conceito é mapeado pelo seu conceito ascendente no nível máximo de abstração.

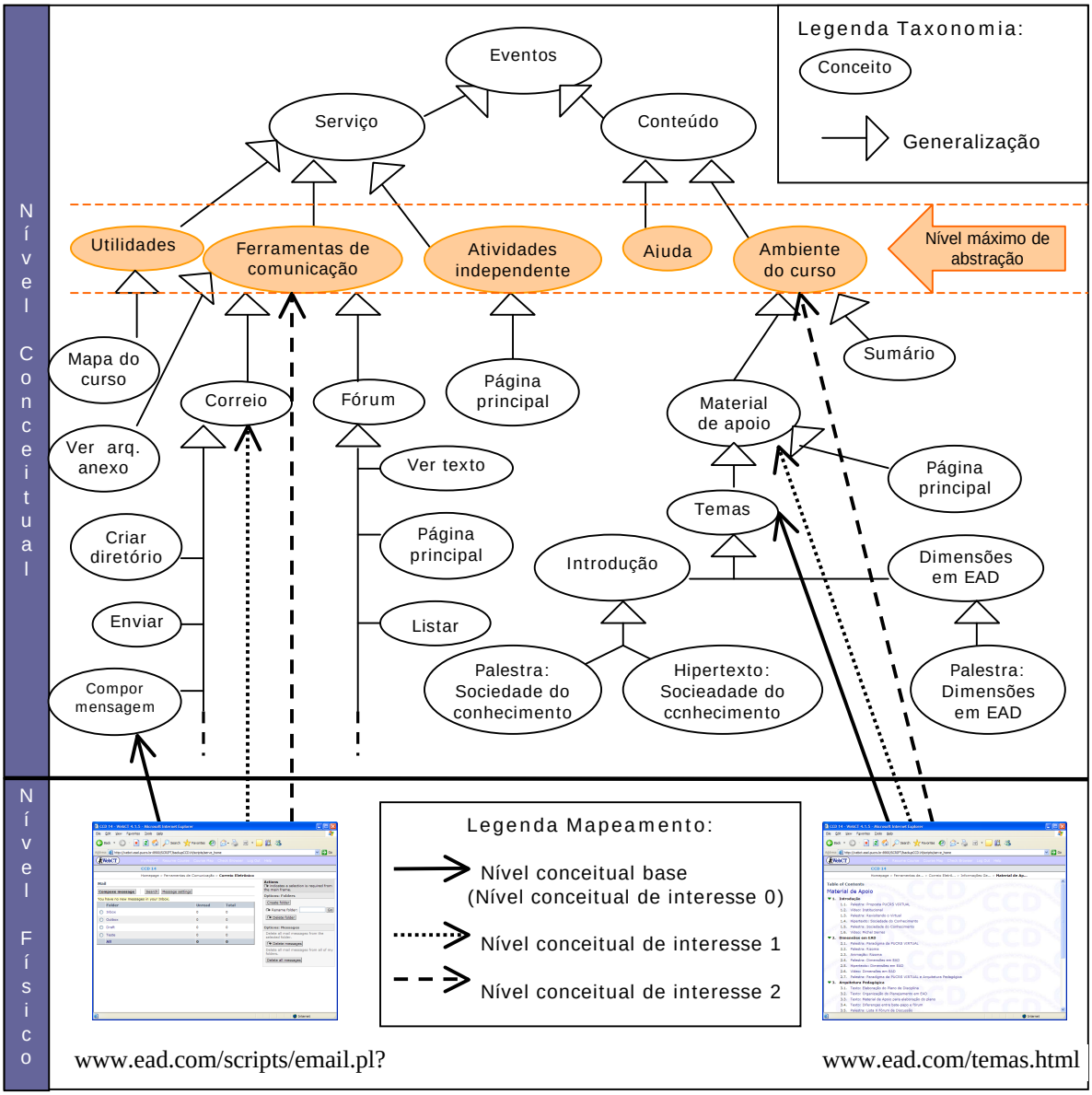

Figura 15 – Níveis de representação dos eventos da aplicação

5.3 Descrição dos Mecanismos

Os mecanismos de agrupamento e interpretação propostos fazem uso de uma taxonomia como forma de agregar semântica aos eventos do domínio dinamicamente, melhorando assim a qualidade do agrupamento e reduzindo a necessidade de retorno à etapa de pré-processamento. Esta dinamicidade é dada pela representação das sessões para todos os possíveis níveis conceituais de interesse que a hierarquia conceitual pode suportar, até o nível máximo de abstração definido. Desta maneira, para cada nível conceitual de interesse o agrupamento é realizado e os padrões resultantes de cada nível de abstração ficam disponíveis para serem interpretados,

sem a necessidade da volta à etapa inicial para gerar uma nova perspectiva dos dados.

O mecanismo de agrupamento proposto estende o agrupamento de sessões WLCS descrito na seção 4.2.1.1 em dois aspectos: a) considera a similaridade entre as páginas durante o cálculo de similaridade entre as sessões, e b) permite lidar de forma homogênea com o agrupamento de interesse e trajetória, indiferente do tipo de peso atribuído. A abordagem do mecanismo proposto é inovadora, pois leva em conta a similaridade entre os conceitos considerando as relações hierárquicas de uma taxonomia, agrupa as sessões aplicando um enriquecimento dinâmico, além de facilitar a aplicação do agrupamento por pessoas leigas. O enriquecimento dinâmico realizado pelo mecanismo de agrupamento explora as relações hierárquicas da taxonomia do domínio com o objetivo de abstrair as sessões para todas os possíveis níveis de abstração durante a etapa de descoberta de padrões para então disponibilizar os padrões resultantes de cada nível conceitual de interesse para posterior interpretação na etapa de análise de padrões, reduzindo assim a necessidade de volta à etapa inicial da MUW. O Capítulo 6 descreve como o mecanismo de agrupamento proposto funciona.

O mecanismo de interpretação proposto permite representar os grupos resultantes de maneira condizente com os objetivos da mineração, bem como facilitar a interpretação dos mesmos através da mudança dinâmica do nível desejado de abstração das sessões no domínio da aplicação. A abordagem do mecanismo proposto é inovadora por disponibilizar diferentes visualizações de acordo com o objetivo do agrupamento, bem como por oferecer uma inspeção dinâmica nos grupos resultantes. A inspeção dinâmica permite que os grupos sejam visualizados quanto ao nível conceitual de interesse sem a necessidade de voltar à etapa inicial da MUW para obter agrupamentos mais significativos. O Capítulo 7 descreve como o mecanismo de interpretação proposto funciona.

5.4 Pressupostos

Os mecanismos propostos para agrupamento e interpretação estabelecem as seguintes premissas a respeito das fases da MUW:

- O pré-processamento considera como fonte de dados um conjunto de URLs provenientes de curso Web registrado por um servidor Web, o qual é processado através de tarefas típicas como limpeza, filtragem, identificação de usuário, identificação de sessão (e possivelmente transação), e complemento de caminhos [COO99];

- Uma taxonomia existe, a qual reflete o conhecimento do domínio, e as URLs são mapeadas para conceitos no nível conceitual da hierarquia do domínio dada a interpretação do especialista do domínio em relação aos eventos da aplicação. O mapeamento das URLs, representadas no nível físico, para o nível conceitual acontece durante o enriquecimento dos dados na etapa de pré-processamento dos dados;
- O especialista do domínio define durante a criação da hierarquia conceitual o nível máximo de abstração que os conceitos serão generalizados considerando os eventos da aplicação representados nesta hierarquia;
- Como parte das tarefas típicas de pré-processamento os conceitos são classificados de acordo com o tempo de acesso (auxiliares e conteúdo) e as sessões, ou transações, são classificadas em [COO99]: sessões de conteúdo ou sessões auxiliar-conteúdo. O agrupamento de interesse utiliza sessões de conteúdo e o agrupamento de trajetória utiliza sessões do tipo auxiliar-conteúdo;
- As sessões são preparadas para o agrupamento de acordo com o objetivo de mineração (ver detalhes na seção 5.5);
- As tarefas específicas de transformação das sessões [BAN01, FU00, MOB01, MOB04] são aplicadas nas sessões identificadas provenientes da etapa de pré-processamento de maneira a reduzir os ruídos e a dimensionalidade dos dados, visando obter grupos de melhor qualidade;
- O tempo de acesso, quando utilizado, não é normalizado;
- O analista define o número de grupos na etapa de descoberta de padrões.

5.5 Representação das Sessões

Da mesma forma que [BAN01, GUN03], no presente trabalho o caminho do usuário durante a navegação no site Web é visto como uma seqüência de pares ordenados de conceitos e pesos. Como tarefa de transformação das sessões, o mecanismo de agrupamento proposto permite utilizar ou não o peso pelo tempo de acesso, de acordo com o objetivo da mineração. Quando utilizado o peso pelo tempo de acesso, este trabalho assume que o tempo não deve ser normalizado, uma vez que o algoritmo WLCS, utilizado como base do mecanismo de agrupamento proposto, tem como fator principal o tempo de acesso para calcular a similaridade entre as sessões. Por outro lado, se o tempo de acesso não é relevante para o objetivo da mineração,

então este trabalho assume o peso binário. Ou seja, todas as páginas têm peso igual, representado pela unidade de 1 (um) segundo.

O mecanismo de agrupamento proposto por este trabalho assume a mesma representação das sessões, ilustrada pela Figura 7, para ambos os objetivos de agrupamento (interesse e trajetória). O caminho de navegação original, contendo a ordem dos acessos e as re-visitas, permanece inalterado na representação da sessão do usuário quando o objetivo é o agrupamento de trajetória, conforme ilustra a Tabela 2-A. Se o tempo de acesso for irrelevante para o objetivo do agrupamento de trajetória, então se assume o peso binário, conforme ilustra a Tabela 2-B.

Tabela 2 – Exemplo de sessões para o agrupamento de trajetória

$s_1 = \{(c_{12},60),(c_{121},30),(c_{11},50),(c_{12},50)\}$ $s_2 = \{(c_{12},40),(c_{12},60)\}$ $s_3 = \{(c_{12},30),(c_{11},40),(c_{121},10)\}$	$s_1 = \{(c_{12},1),(c_{121},1),(c_{11},1),(c_{12},1)\}$ $s_2 = \{(c_{12},1),(c_{12},1)\}$ $s_3 = \{(c_{12},1),(c_{11},1),(c_{121},1)\}$
(A) Peso pelo tempo	(B) Peso binário

Já quando o objetivo é o agrupamento de interesse, as sessões sofrem uma transformação no seu caminho original para não levar em conta nem a ordem e nem a re-visita aos conceitos na mesma sessão do usuário. Aplicando tarefas específicas de transformação das sessões, os conceitos são re-arranjados na sessão, de acordo com sua localização na hierarquia conceitual, de maneira a formar uma seqüência de acessos em comum. Se o peso pelo tempo de acesso foi escolhido, então os tempos são somados para as re-visitas. Isto é ilustrado na Tabela 3-A, considerando as mesmas sessões ilustradas na Tabela 2-A. Se o tempo de acesso não é relevante para o objetivo do agrupamento de interesse, então se assume peso binário, conforme ilustra a Tabela 3-B, para o mesmo exemplo.

Tabela 3 – Exemplo de sessões para o agrupamento de interesse

$s_1 = \{(c_{11},50),(c_{12},110),(c_{121},30)\}$ $s_2 = \{(c_{12},100)\}$ $s_3 = \{(c_{11},40),(c_{12},30),(c_{121},10)\}$	$s_1 = \{(c_{11},1),(c_{12},1),(c_{121},1)\}$ $s_2 = \{(c_{12},1)\}$ $s_3 = \{(c_{11},1),(c_{12},1),(c_{121},1)\}$
(A) Peso pelo tempo	(B) Peso binário

A re-organização dos conceitos pode ser baseada em metadados da própria página, pela ontologia do site Web (ex: web semântica), pela organização dos conceitos na hierarquia conceitual (ex: caminhamento na árvore que representa a hierarquia), ou até mesmo pelo próprio especialista do domínio. Considerando que a hierarquia conceitual pode ser representada por uma árvore, o caminhamento BFS (Breadth First Search) [COR90], ou caminhamento em largura, pode ser resumido da

seguinte maneira: partindo da raiz, passa nível por nível da árvore, caminha pelos nodos da esquerda para a direita dentro de cada nível (onde o nível é definido simplesmente em termos da distância da raiz). Por exemplo, a árvore ilustrada pela Figura 16 é numerada na ordem em que o caminhamento BFS lê os nodos.

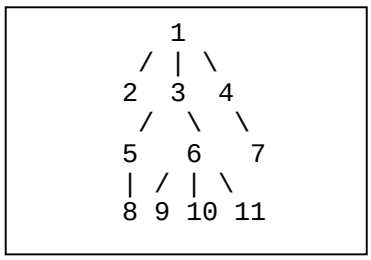

Figura 16 – Exemplo de caminhamento Breadth First Search (BFS)

Este trabalho adota a organização dos conceitos dada pelo caminhamento em largura. Pressupõe-se, que conceitos irmãos estão organizados em ordem alfabética na hierarquia conceitual. A utilização do caminhamento em largura possibilita aproximar conceitos irmãos e primos (mesma profundidade na hierarquia) na re-organização das sessões. Por exemplo, considerando a hierarquia conceitual ilustrada pela Figura 17, e a sessão s_1 representada no nível conceitual base ilustrada pela Tabela 2-A, a transformação desta sessão para a sessão ilustrada na Tabela 3-A ocorre da seguinte maneira:

- Primeiramente os conceitos da hierarquia são re-arranjados na sessão de acordo com o caminhamento em largura: (c, c_1, c_2, c_{11}, c_{12}, c_{21}, c_{121}, e c_{122}). Assim, após reordenar os conceitos, a sessão s_1 fica representada da seguinte forma: $s_1 = \{(c_{11},50),(c_{12},60),(c_{12},50),(c_{121},30)\}$.
- Por fim, as re-visitas aos conceitos são eliminadas e seus tempos somados. A representação final da sessão s_1 para o agrupamento de interesse pode ser vista como: $s_1 = \{(c_{11},50),(c_{12},110),(c_{121},30)\}$.

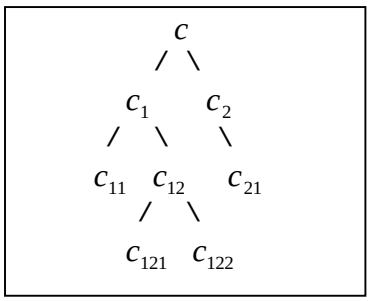

Figura 17 – Exemplo de hierarquia conceitual

A transformação em termos de re-organização dos conceitos na representação das sessões para o agrupamento de interesse é necessária, pois deste modo o mecanismo de agrupamento proposto, o qual utiliza a técnica WLCS considerando a similaridade entre os conceitos, pode ser generalizado para ambos objetivos de agrupamento (interesse e trajetória). Por exemplo, considerando as sessões s_1 e s_3 representadas no nível conceitual base, como ilustra a Tabela 2-A pode-se notar que, apesar de cada sessão apresentar originalmente um caminho de navegação distinto, quando transformadas seguindo a ordem definida pela re-organização dos conceitos, como ilustra a Tabela 3-A, estas sessões têm seus caminhos transformados de maneira que nem a ordem nem as re-visitas são levadas em conta, obtendo-se assim uma subseqüência em comum que identifica seus interesses em comum.

5.5.1 Nível Conceitual de Interesse na Representação das Sessões

Considerando que as URLs são mapeadas do nível físico para o nível conceitual base e que os conceitos da taxonomia estão ligados por relacionamentos de generalização/especialização, este trabalho assume que as sessões podem ser traduzidas para qualquer nível conceitual de interesse. Este enriquecimento é dado em termos da abstração das sessões através da generalização recursiva dos conceitos ascendentes do nível conceitual base até o nível conceitual de interesse desejado. Ou seja, de acordo com o nível conceitual de interesse, uma sessão pode ser representada de forma mais generalizada ou mais especializada. Uma sessão representada no nível conceitual base contém conceitos que identificam os eventos de aplicação de acordo com a interpretação mais detalhada do especialista do domínio. Por sua vez, uma sessão traduzida para um nível conceitual de interesse na hierarquia conceitual contém abstrações dos conceitos no nível conceitual de interesse em questão.

O enriquecimento dinâmico das sessões possibilita traduzir cada sessão para os possíveis níveis conceituais de interesse. O nível conceitual de interesse de uma sessão obedece ao número máximo de ascendente dos conceitos acessados nesta e ao nível máximo de abstração definido. A Tabela 4 ilustra as possíveis abstrações da sessão s_1 descrita na Tabela 2-A demonstrando o enriquecimento dinâmico das sessões. Por exemplo, no nível conceitual base (nci=0) a sessão s_1 é representada na Tabela 4 pelos conceitos da hierarquia que mapeiam os eventos da aplicação de acordo com a visão do especialista do domínio. Já quando outro nível conceitual de interesse, nci={1, 2, 3}, é utilizado para traduzir as sessões, cada conceito pertencente à sessão é generalizado para seu conceito ascendente correspondente ao nível conceitual de interesse em questão, diretamente ou por recursão.

Tabela 4 – Enriquecimento dinâmico das sessões

Representação das Sessões (agrupamento de trajetória)	Representação das Sessões (agrupamento de interesse)	Nível Conceitual de Interesse (nci)
$s_1 = \{(c_{12},60),(c_{121},30),(c_{11},50),(c_{12},50)\}$	$s_1 = \{(c_{11},50),(c_{12},110),(c_{121},30)\}$	nci = 0
$s_1 = \{(c_1,60),(c_{12},30),(c_1,50),(c_1,50)\}$	$s_1 = \{(c_1,160),(c_{12},30)\}$	nci = 1
$s_1 = \{(c,60),(c_1,30),(c,50),(c,50)\}$	$s_1 = \{(c,160),(c_1,30)\}$	nci = 2
$s_1 = \{(c,60),(c,30),(c,50),(c,50)\}$	$s_1 = \{(c,190)\}$	nci = 3

Com isto, conceitos que eram diferentes podem se tornar iguais, transformando-se naquele nível conceitual de interesse, em re-visitas na mesma sessão. Da mesma forma, sessões sem nenhum nível de similaridade em um determidado nível conceitual de interesse podem se tornar similares quando generalizadas para outro nível conceitual de interesse. Deve-se notar que existe um erro introduzido no agrupamento das sessões dependendo do nível de abstração das sessões. Pois, quanto mais elevado o nível conceitual de interesse em que as sessões são abstraídas, mais as sessões tendem a serem similares umas às outras.

A redução da dimensionalidade pode ainda ser aplicada ao objetivo do agrupamento de trajetória se for relevante ao objetivo do agrupamento em questão, conforme ilustra a Tabela 5. Neste caso, as re-visitas aos conceitos contíguos são unificadas com seus tempos somados (quando utilizado peso pelo tempo).

Tabela 5 – Enriquecimento dinâmico das sessões com redução da dimensionalidade

Representação das Sessões (agrupamento de trajetória)	Nível Conceitual de Interesse (nci)
$s_1 = \{(c_{12},60),(c_{121},30),(c_{11},50),(c_{12},50)\}$	nci = 0
$s_1 = \{(c_1,60),(c_{12},30),(c_1,100)\}$	nci = 1
$s_1 = \{(c,60),(c_1,30),(c,100)\}$	nci = 2
$s_1 = \{(c,190)\}$	nci = 3

Quanto mais o valor do nível conceitual de interesse se aproxima do nível conceitual base (nci = 0), mais os conceitos ficam especializados na representação das sessões. De maneira equivalente, quanto maior o valor do nível conceitual de interesse, mais os conceitos ficam generalizados na representação das sessões, tornando-se possivelmente em re-visitas no nível conceitual de interesse em questão. Assim, utilizar a redução de dimensionalidade é relevante quando se deseja representar as sessões em um nível conceitual de interesse relativamente alto na hierarquia conceitual.

6 MECANISMO DE AGRUPAMENTO

Este capítulo apresenta o mecanismo de agrupamento proposto o qual descreve uma generalização do algoritmo WLCS. Esta generalização visa agrupar sessões, independente do objetivo do agrupamento, considerando a similaridade entre os conceitos e os diferentes níveis conceituais de interesse que as sessões podem ser representadas. O capítulo também apresenta uma análise sobre a efetividade do mecanismo proposto.

O mecanismo de agrupamento propõe uma abordagem que visa facilitar a aplicação do agrupamento de sessões e aumentar a qualidade dos grupos, reduzindo assim a necessidade de retorno à etapa inicial de pré-processamento para gerar uma nova perspectiva dos dados. Esta abordagem foi definida visando complementar as carências apresentadas pelos trabalhos relatados no Capítulo 4, e está fundamentada em três componentes: similaridade entre os conceitos, similaridade entre as sessões e agrupamento dinâmico das sessões. A similaridade entre os conceitos leva em conta a proximidade entre os conceitos considerando as relações hierárquicas de uma taxonomia. A similaridade entre as sessões considera a similaridade entre os conceitos quando computando a subseqüência em comum utilizada como base para o cálculo de similaridade entre as sessões. O agrupamento dinâmico das sessões agrupa as sessões considerando todas os possíveis níveis conceituais de interesse nos quais as sessões podem ser traduzidas.

Detalhes sobre cada um dos componentes utilizados na abordagem proposta pelo mecanismo de agrupamento, bem como a análise sobre sua efetividade são descritos no restante deste capítulo.

6.1 Similaridade entre Conceitos

O mecanismo de agrupamento considera a semântica do domínio, representada por uma taxonomia, para computar a similaridade entre os conceitos em termos de sua localização na hierarquia. A função de similaridade adotada (Fórmula 4) é uma adaptação do componente de similaridade em GVSM (Generalized Vector-Space Model) [GAN03], onde c_i e c_j são conceitos na hierarquia, LCA(c_i,c_j) é o ascendente comum mais próximo de c_i e c_j, e depth(c) é o número de ascendentes do conceito c até o topo da hierarquia.

$$sim(c_i, c_j) = \frac{2 \times depth(LCA(c_i, c_j))}{depth(c_i) + depth(c_j)} \quad (4)$$

Nota-se que quanto mais os conceitos são especializados (i.e. distantes da raiz da hierarquia conceitual), maior sua potencial similaridade. Por exemplo, considerando a hierarquia exemplificada na Figura 17 a similaridade entre os conceitos c_{11} e c_{12} é dada por:

$$sim(c_{11}, c_{22}) = \frac{2 \times depth(LCA(c_{11}, c_{22}))}{depth(c_{11}) + depth(c_{22})} = \frac{2 \times depth(c_1)}{2+2} = \frac{2 \times 1}{4} = 0.5$$

Já a similaridade entre os conceitos c_{121} e c_{122}, que são especializações de c_{12}, é dada por:

$$sim(c_{121}, c_{122}) = \frac{2 \times depth(LCA(c_{121}, c_{122}))}{depth(c_{121}) + depth(c_{122})} = \frac{2 \times depth(c_{12})}{3+3} = \frac{2 \times 2}{6} = 0.66$$

6.2 Similaridade entre as Sessões

O mecanismo de agrupamento proposto estende a similaridade entre as sessões de WLCS, discutido na seção 4.2.1.1, por considerar a similaridade entre os conceitos quando computando a subseqüência em comum entre as sessões. A simples comparação entre os conceitos para encontrar a subseqüência em comum utilizada por WLCS é substituída pela similaridade entre os conceitos (Fórmula 4) considerando como parâmetro adicional o limite de similaridade (m), conforme ilustra a Fórmual 5.

$$sim(c_1, c_2) = \begin{cases} \frac{2 \times depth(LCA(c_1, c_2))}{depth(c_1) + depth(c_2)}, \text{ se } sim(c_1, c_2) \geq \mathrm{m} \\ 0, \text{ caso contrário} \end{cases} \quad (5)$$

O limite de similaridade estabelece um valor mínimo de similaridade entre os conceitos para que dois conceitos, originalmente diferentes, possam ser considerados “similares” observando seu grau de proximidade na hierarquia conceitual do domínio. Assim, é proposto o algoritmo sim_LCS, o qual adapta o algoritmo LCS Delta [HIS77] utilizado por WLCS. O pseudo-código de sim_LCS é ilustrado na Figura 18. A adaptação visa substituir a simples comparação entre os conceitos pelo grau de similaridade entre os conceitos em termos de sua localização na hierarquia, conforme demonstram as linhas 12 e 22 da Figura 18. Por exemplo, considerando a hierarquia conceitual da Figura 17, as sessões s_1 e s_2 ilustradas na Tabela 2-A e um limite de similaridade 0.5, os conceitos c_{11} e c_{12}, apesar de não serem idênticos, são considerados semelhantes quando computando a subseqüência em comum entre as sessões pois atingem o limite de similaridade estabelecido. A definição de um limite de similaridade não é uma tarefa fácil, pois é dependente de sua localização em termos da profundidade da hierarquia. A definição de um valor ideal para o limite de similaridade está fora do escopo deste trabalho. Nota-se que se o limite de

similaridade for 1, então a matriz de similaridade gerada por sim_WLCS é igual à matriz de similaridade gerada por WLCS.

```
sim_LCS(x, y, m)
Entradas:
x    seqüência de conceitos acessados na primeira sessão
y    seqüência de conceitos acessados na segunda sessão
m    limite de similaridade

Saídas:
wx   índices de x que participam da subseqüência em comum
wy   índices de y que participam da subseqüência em comum
1. int [][]T
2. int i, j
3. m = length(x)
4. n = length(y)
5. // Criação da matriz do LCS
6. for (i ← -1) to (m-1) do
7.    T[i, -1] ← 0
8. for (j ← -1) to (n-1) do
9.    T[-1, j] ← 0
10. for (i ← 0) to (m-1) do
11.   for (j ← 0) to (n-1) do
12.      if (sim(xi, yj) ≥ m) then
13.         T[i, j] ← T[i-1, j-1] + 1
14.      else
14.        T[i, j] ← max (T[i, j-1], T[i-1, j])
16. // Caminha pela matriz LCS para obter os índices das sessões
17. // que participam da subseqüência em comum
18. i ← m-1
19. j ← n-1
20. k ← T [i, j]
21. While (i > 0) and (j > 0) do
22.    if (T[i, j] = (T[i-1, j-1] + 1)) and (sim(xi, yj) ≥ m) then
23.       wx[k] ← i
24.       wy[k] ← j
25.       i ← i-1
26.       j ← j-1
27.       k ← k-1
28.   else
29.      if (T [i-1, j] > T[i, j-1]) then
30.         i ← i-1
31.      else
32.          j ← j-1
```

Figura 18 – Algoritmo sim_LCS (adaptado de LCS Delta)

O objetivo de sim_WLCS é computar a subseqüência em comum através do algoritmo sim_LCS e aplicar as mesmas fórmulas do WLCS (Fórmulas 1 a 3) de modo a obter a similaridade final entre as sessões. Vale lembrar que, caso o objetivo do agrupamento não leve em conta o fator tempo de acesso, as sessões já terão sido

transformadas para indicarem o peso 1 (um segundo) para todos os acessos. Neste caso, a similaridade final entre as sessões obtida pelo WLCS fica restrita ao valor obtido pelo componente de importância (Fórmula 2), uma vez que o componente de similaridade do WLCS (Fórmula 1) retornará sempre 1 (um).

Considerando que duas sessões podem ter mais de uma subseqüência em comum, sim_WLCS calcula a similaridade entre sessões aplicando o algoritmo sim_LCS duas vezes, invertendo as sessões como parâmetros, e atribui o valor mais alto obtido como similaridade entre as sessões. Por exemplo, considerando o limite de similaridade 0.5 e duas sessões $s_1 = \{(c_{12},60),(c_{121},30),(c_{11},50),(c_{12},50)\}$ e $s_3 = \{(c_{12},30),(c_{11},40),(c_{121},10)\}$, o mecanismo de agrupamento calcula primeiramente a similaridade entre s_1 e s_3 considerando a subseqüência obtida por sim_LCS(s_1,s_3,0.5), conforme ilustra a Figura 19.

sim_LCS(s_1,s_3,0.5)		
índice na sessão / i	$l^{s_1}(i)$	$l^{s_3}(i)$
1	3	1
2	4	2

$s_1 = \{(c_{12},60),(c_{121},30),\underline{(c_{11},50)},\underline{(c_{12},50)}\}$

$s_3 = \{\overline{(c_{12},30)},\overline{(c_{11},40)},(c_{121},10)\}$

sim_WLCS(s_1,s_3)

$$S' = \frac{\left(\frac{30}{50}+\frac{40}{50}\right)}{2} = \frac{0.60+0.80}{2} = 0.70$$

$$S'' = \sqrt{\frac{100}{190}\times\frac{70}{80}} = \sqrt{0.53\times 0.875} = 0.68$$

$$S = 0.70\times 0.68 = 0.48$$

Figura 19 – Similaridade obtida com a subseqüência dada por sim_LCS(s_1,s_3)

Então, calcula a novamente a similaridade entre s_1 e s_3 considerando a subseqüência obtida por sim_LCS(s_3,s_1,0.5), conforme ilustra a Figura 20. Neste exemplo, a similaridade final entre as sessões s_1 e s_3 é 0.48, pois foi o maior valor calculado por sim_WLCS considerando as subseqüências obtidas entre as duas sessões.

sim_LCS(s_3,s_1,0.5)		
índice na sessão / i	$l^{s_1}(i)$	$l^{s_3}(i)$
1	1	2
2	2	3

$s_3 = \{(c_{12},30),\underline{(c_{11},40)},\underline{(c_{121},10)}\}$

$s_1 = \{\overline{(c_{12},60)},\overline{(c_{121},30)},(c_{11},50),(c_{12},50)\}$

sim_WLCS(s_1,s_3)

$$S' = \frac{\left(\frac{40}{60}+\frac{10}{30}\right)}{2} = \frac{0.67+0.33}{2} = 0.50$$

$$S'' = \sqrt{\frac{90}{190}\times\frac{50}{80}} = \sqrt{0.47\times 0.62} = 0.54$$

$$S = 0.50\times 0.54 = 0.27$$

Figura 20 – Similaridade obtida com a subseqüência dada por sim_LCS (s_3,s_1)

6.3 Agrupamento Dinâmico das Sessões

O mecanismo de agrupamento agrupa as sessões considerando todos os níveis de interesse nos quais as sessões podem ser traduzidas de acordo com a hierarquia conceitual do domínio. A altura da hierarquia conceitual do domínio determina os possíveis níveis conceituais de interesse nos quais as sessões podem ser traduzidas. Esta altura é obtida pelo maior depth(c) entre os conceitos da hierarquia conceitual. Entretanto, como já salientado na seção 5.2, o analista pode definir o nível máximo de abstração para os conceitos. O nível máximo de abstração dos conceitos é independente do número de níveis conceituais de interesse que a hierarquia conceitual pode representar, pois os conceitos podem ter diferentes números de ascendentes na hierarquia.

Para cada nível conceitual de interesse, o agrupamento dinâmico das sessões realiza as seguintes tarefas, as quais são executadas nesta ordem:

1. Enriquecimento dinâmico das sessões: as sessões são traduzidas para o nível conceitual de interesse em questão mantendo as mesmas tarefas específicas de transformação das sessões [BAN01, FU00, MOB01, MOB04] escolhidas para o objetivo do agrupamento. Por exemplo, considerando o agrupamento de trajetória com peso pelo tempo de acesso e a tarefa de transformação das sessões para redução de dimensionalidade, a unificação dos conceitos contíguos é realizada e seus tempos somados nas sessões traduzidas para cada nível conceitual de interesse, conforme ilustra a Tabela 5.
2. Geração da matriz de similaridade: a matriz de similaridade correspondente ao nível conceitual de interesse é calculada utilizando os componentes de similaridade entre os conceitos e de similaridade entre as sessões para as sessões traduzidas no nível conceitual de interesse em questão, como discutido nas seções 6.1 e 6.2.
3. Agrupamento das sessões: as sessões são agrupadas através de um algoritmo baseado em grafo que utiliza a respectiva matriz de similaridade do nível conceitual de interesse correspondente. O analista é quem deve definir o número de grupos. A definição do número ideal de grupos é uma tarefa complexa [HEE02] e está fora do escopo deste trabalho. Assume-se que o mesmo número de grupos vale para agrupar as sessões em todos os níveis conceituais de interesse.

O mecanismo de agrupamento proposto aplica a mesma idéia descrita por [VAN04a] no que diz respeito à abstração dos conceitos de acordo com os diferentes

níveis nos quais os conceitos podem ser traduzidos. Entretanto, o mecanismo de agrupamento explora as relações hierárquicas de uma taxonomia do domínio com o objetivo de abstrair as sessões de acordo com o nível conceitual de interesse durante a etapa de descoberta de padrões para então disponibilizar os grupos resultantes de cada abstração para posterior interpretação na etapa de análise de padrões. Isto porque os grupos podem mudar significativamente de acordo com a abstração das sessões, diferentemente de [VAN04a], onde os padrões seqüenciais não mudam, apenas seu suporte. Assim, partindo do nível conceitual base até o nível conceitual de interesse mais abstrato, o agrupamento é realizado e os grupos resultantes de cada nível conceitual ficam disponíveis para serem interpretados, sem a necessidade da volta à etapa inicial de pré-processamento para gerar uma nova perspectiva dos dados.

6.4 Análise Comparativa

Esta seção apresenta uma análise comparativa entre o mecanismo de agrupamento proposto e as técnicas existentes, visando ilustrar sua efetividade no cálculo de similaridade entre as sessões, bem como na melhora do agrupamento das sessões. Para tanto, foram criadas algumas sessões exemplos, apresentadas na Figura 21, que simulam sessões de usuários em um curso real da PUCRS Virtual, e têm como objetivo apenas ilustrar o funcionamento do mecanismo proposto. As sessões apresentadas na Figura 21 são representadas de forma tabular, onde a primeira linha representa o primeiro acesso, e a última linha, o último acesso da sessão. Assume-se que estas sessões estão representadas no nível conceitual base e os conceitos acessados fazem parte de uma hierarquia conceitual, ilustrada na Figura 15, que descreve o domínio de um curso real de EAD da PUCRS Virtual onde os usuários podem participar de fóruns, trocar e-mails, realizar atividades, ler o material das aulas, consultar a ajuda on-line, ver o mapa do curso, etc. É importante ressaltar que os conceitos utilizados nesta análise comparativa foram extraídos observando uma representação simplificada da hierarquia conceitual real de um dos cursos de EAD da PUCRS Virtual, a qual foi validada pelo especialista do domínio da PUCRS Virtual.

S_a

Conceito	Peso (seg)
Atividades independentes: Página principal	200
Correio: Página principal	440
Correio: Compor mensagem	300
Correio: Listar todos	200
Fórum: Listar assunto específico	400
Fórum: Listar assuntos não lidos	100
Ver arquivo anexo	300
Fórum: Página principal	300
Fórum: Listar assuntos principais	300
Fórum: Ver texto	300
Fórum: Listar assunto específico	300
Fórum: Listar assuntos resumido	300
Fórum: Ver texto	300
Fórum: Listar sem encadeamento	180

S_b

Conceito	Peso (seg)
Correio: Criar diretório	200
Correio: Listar caixa de entrada	300
Correio: Carregar arquivo anexo	300

S_c

Conceito	Peso (seg)
Mapa do curso	400
Material de Apoio: Página principal	200
Palestra: Sociedade do conhecimento	500
Mapa do curso	400

S_d

Conceito	Peso (seg)
Material de Apoio: Página principal	300
Hipertexto: Sociedade do conhecimento	200

S_e

Conceito	Peso (seg)
Hipertexto: Sociedade do conhecimento	640
Correio: Compor mensagem	300
Correio: Listar todos	300
Correio: Enviar	300

S_f

Conceito	Peso (seg)
Palestra: Dimensões em EAD	300
Fórum: Listar assunto específico	300
Fórum: Listar assuntos não lidos	300
Fórum: Listar assuntos resumido	300

S_g

Conceito	Peso (seg)
Hipertexto: Sociedade do conhecimento	200
Material de Apoio: Página principal	300

Figura 21 – Sessões exemplo utilizadas na análise do mecanismo de agrupamento

6.4.1 Encontrar Similaridade entre Sessões

O mecanismo de agrupamento proposto tende a atribuir um grau de similaridade para as sessões que têm conceitos similares, e que são consideradas totalmente diferentes pelas técnicas de agrupamento de trajetória tradicionais [FU00, BAN01]. Por exemplo, considerando as sessões s_a e s_b representadas no nível conceitual base, ilustrado na Figura 21, e a igualdade entre os conceitos para construir a subseqüência em comum, a similaridade entre estas sessões calculado por WLCS é 0 (zero), dado que o LCS entre s_a e s_b é vazio.

Por outro lado, considerando o sim_WLCS, e um limite de similaridade 0.7, as mesmas sessões s_a e s_b passam a ter um grau de similaridade de 0.35, uma vez que existem conceitos considerados similares nestas, conforme ilustra a Figura 22. O valor de similaridade entre as sessões s_a e s_b está diretamente ligado ao tempo de acesso às páginas pertencentes à subseqüência em comum, conforme já mencionado na seção 4.2.1.1, bem como ao valor do limite de similaridade estipulado entre os conceitos que forma a subseqüência em comum.

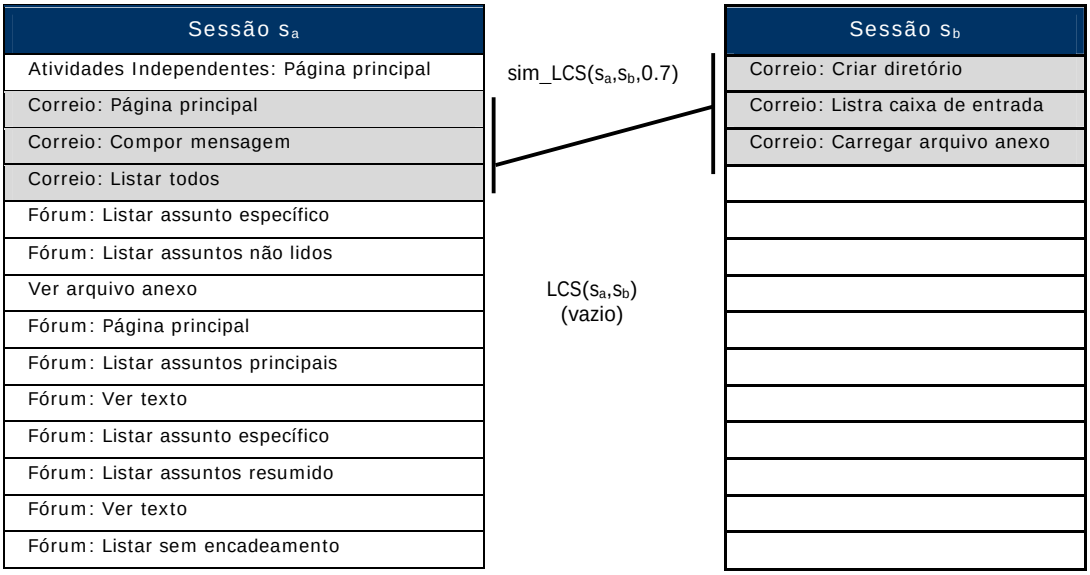

Figura 22 – Exemplo de subseqüência entre sessões com conceitos similares

6.4.2 Melhorar a Similaridade entre as Sessões

O mecanismo de agrupamento proposto também permite melhorar o grau de similaridade entre duas sessões já consideradas similares pelas técnicas tradicionais [FU00, BAN01], observando que a similaridade entre os conceitos é levada em conta quando computando a similaridade entre as sessões. Isto é possível para as sessões que incluem, além de conceitos idênticos, conceitos que atingem o limite de similaridade estipulado pelo analista. Por exemplo, considerando as sessões s_c e s_d da Figura 21 representadas no nível conceitual base, a igualdade entre os conceitos para construir a subseqüência em comum, e peso pelo tempo de acesso, a similaridade entre estas sessões calculado por WLCS é 0.19. Pois existe apenas um conceito idêntico em comum (“Material de Apoio: Página principal”) entre as duas sessões para formar a subseqüência em comum, como ilustra a Figura 23. Se considerarmos o sim_WLCS, e um limite de similaridade 0.7, a similaridade entre as sessões s_c e s_d aumenta para 0.36, visto que agora existem dois conceitos em comum para construir a subseqüência em comum. Isto é ilustrado na Figura 23, dado que a similaridade entre os conceitos “Palestra: Sociedade do conhecimento” e “Hipertexto: sociedade do conhecimento” é de 0.83 e atinge o limite de similaridade entre os conceitos estipulado.

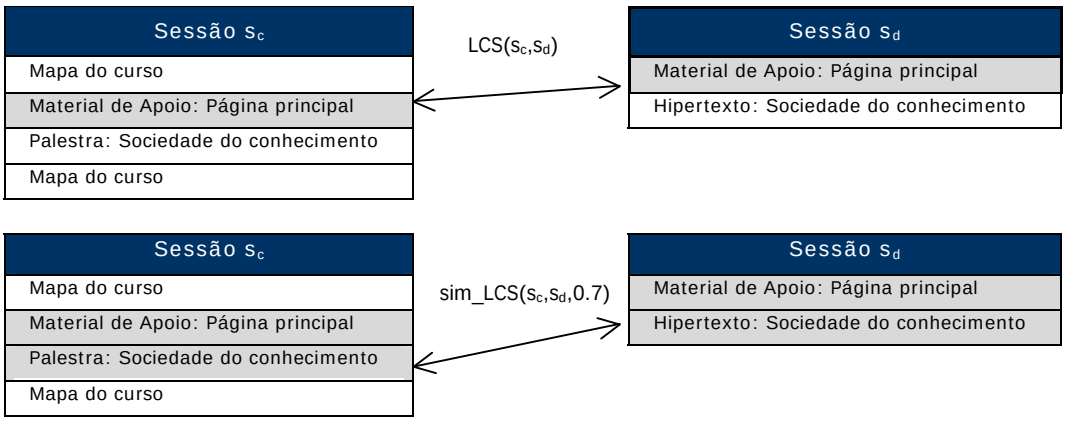

Figura 23 – Exemplo de melhora no grau de similaridade entre as sessões

6.4.3 Agrupamento de Interesse

O mecanismo de agrupamento proposto pode ser aplicado em ambos os objetivos de agrupamento, a saber interesse e trajetória. Para tanto, quando se trata do objetivo de agrupamento de interesse, o mecanismo de agrupamento proposto re-organiza os conceitos acessados dentro das sessões como parte das tarefas específicas de transformação das sessões, como discutido na seção 5.5. A re-organização dos conceitos dentro das sessões visa representar somente os interesses em comum entre as sessões, de modo que a ordem de acesso e as re-visitas não sejam levadas em conta. É importante notar que esta representação se assemelha com a representação adotada pelas técnicas tradicionais de agrupamento de interesse que representam as sessões através de vetores de atributos. A diferença é que o sim_WLCS considera somente os conceitos acessados na sessão na representação desta. Isto reduz significativamente a quantidade de atributos que representam as sessões, facilitando conseqüentemente a caracterização dos grupos. Por exemplo, considere as sessões s_d e s_g, da Figura 21, representando duas diferentes trajetórias no nível conceitual base, e peso binário. A re-organização dos conceitos realizada pelo mecanismo de agrupamento torna estas sessões idênticas do ponto de vista dos interesses dos usuários da mesma maneira que as técnicas convencionais de agrupamento de interesse [FU00, HEE02, MOB01, MOB02] fazem com os vetores de atributos, conforme discutido em 4.1.

Já se o peso pelo tempo de acesso for considerado, então a similaridade entre as sessões s_d e s_g calculada por sim_WLCS é de 0.67, enquanto que a similaridade obtida pelas medidas de convencionais de distância (ex: Coseno) aplicadas aos vetores de atributos é de 0.92. O valor de similaridade obtido por sim_WLCS não é tão alto quanto a similaridade obtida pela medida de distância Coseno, pois a similaridade calculada por sim_WLCS é mais sensível à diferença dos pesos dos

conceitos. A técnica WLCS justamente tenta explorar a relevância dos pesos de cada acesso dentro das sessões para descobrir grupos com características de navegação diferentes.

O modo como os conceitos são re-organizados dentro das sessões afeta diretamente a construção da subseqüência em comum e por conseqüência, a similaridade entre as sessões no agrupamento de interesse. Por exemplo, considerando duas sessões s_c e s_d representadas no nível conceitual base (Figura 21), ordem alfabética entre os conceitos como forma de re-organização dos conceitos dentro das sessões que representam os interesse, um limite de similaridade de 0.7, e o peso pelo tempo de acesso, é possível obter duas subseqüências em comum, como ilustra a Figura 24. A primeira subseqüência é dada pelo índice 2 da sessão s_c e o índice 2 da sessão s_d e a similaridade entre s_c e s_d é 0.19. A segunda subseqüência é dada pelo índice 3 da sessão s_c e o índice 1 da sessão s_d e a similaridade entre s_c e s_d é 0.15. O mecanismo de agrupamento proposto fica com o valor mais alto obtido por sim_WLCS.

Índice	Sessão s_c
1	Mapa do curso
2	Material de Apoio: Página principal
3	Palestra: Sociedade do conhecimento

sim_LCS(s_c,s_d,0.7) ⟷

Sessão s_d	Índice
Hipertexto: Sociedade do conhecimento	1
Material de Apoio: Página principal	2

Índice	Sessão s_c
1	Mapa do curso
2	Material de Apoio: Página principal
3	Palestra: Sociedade do conhecimento

sim_LCS(s_d,s_c,0.7) ↙↗

Sessão s_d	Índice
Hipertexto: Sociedade do conhecimento	1
Material de Apoio: Página principal	2

Figura 24 – Re-organização dos conceitos considerando puramente a ordem alfabética

Já quando considerada a organização dos conceitos pelo caminhamento em largura, adotado neste trabalho (seção 5.5), pode-se observar melhora nas subseqüências obtidas, pois o algoritmo LCS é sensível à ordem dos conceitos nas sessões. Por exemplo, é possível obter uma maior subseqüência em comum entre as mesmas sessões s_c e s_d se os conceitos “Material de Apoio: Página principal” e “Hipertexto: Sociedade do conhecimento” na sessão s_d forem re-ordenados de acordo com o caminhamento em largura. Isto é ilustrado na a Figura 25, onde o LCS entre as sessões s_c e s_d aumentou de um para dois conceitos. Neste caso, o aumento da subseqüência em comum refletiu no aumento da similaridade entre as sessões s_c e s_d (0.36).

Índice	Sessão s_c
1	Mapa do curso
2	Material de Apoio: Página principal
3	Palestra: Sociedade do conhecimento

sim_LCS(s_c,s_d,0.7)

Sessão s_d	Índice
Material de Apoio: Página principal	1
Hipertexto: Sociedade do conhecimento	2

Figura 25 – Re-organização dos conceitos considerando o caminhamento em largura

6.4.4 Agrupamento Dinâmico das Sessões

Outro importante aspecto do mecanismo de agrupamento proposto é o agrupamento dinâmico das sessões considerando os possíveis níveis conceituais de interesse nos quais as sessões podem ser traduzidas. O agrupamento dinâmico das sessões explora as relações hierárquicas da taxonomia do domínio com o objetivo de automaticamente, para todos os possíveis níveis conceituais de interesse:

a) aplicar o enriquecimento dinâmico das sessões;

b) aplicar sim_WLCS para gerar a matriz de similaridade correspondente;

c) agrupar as sessões. O conjunto de padrões resultantes de cada nível conceitual de interesse fica disponível para posterior interpretação.

A Figura 26 ilustra como as sessões s_e e s_f da Figura 21 são enriquecidas, conforme o nível conceitual de interesse. Este exemplo é detalhado no restante da seção.

nci=0 nci=1 nci=2 nci=3 nci=4

Índice	Sessão s_e	Sessão s_e	Sessão s_e	Sessão s_e	Sessão s_e
1	Hipertexto: Sociedade do conhecimento	Introdução	Temas	Material de apoio	Ambiente do curso
2	Correio: Compor mensagem	Correio	Ferramentas de comunicação	Ferramentas de comunicação	Ferramentas de comunicação
3	Correio: Listar todos				
4	Correio: Enviar				

Índice	Sessão s_f	Sessão s_f	Sessão s_f	Sessão s_f	Sessão s_f
1	Palestra: Dimensões em EAD	Dimensões em EAD	Temas	Material de apoio	Ambiente do curso
2	Fórum: Listar assunto específico	Fórum	Ferramentas de comunicação	Ferramentas de comunicação	Ferramentas de comunicação
3	Fórum: Listar assuntos não lidos				
4	Fórum: Listar assuntos resumido				

Figura 26 – Exemplo de enriquecimento dinâmico das sessões

6.4.4.1 Nível Conceitual de Interesse 0

Primeiramente o passo (a) é realizado para o nível conceitual de interesse base. Considerando que as sessões s_e e s_f já estão representadas no nível conceitual base, como ilustra a Figura 26, o enriquecimento dinâmico das sessões para o nível conceitual base não altera a representação das sessões s_e e s_f.

Em seguida, o passo (b) é executado, o qual aplica sim_WLCS e gera a matriz de similaridade. A subseqüência em comum encontrada entre as duas sessões, considerando sim_WLCS e WLCS, é identificada pelos índices dos conceitos dentro das sessões s_e e s_f, como ilustra a Figura 27. É importante notar que o WLCS não consegue encontrar uma subseqüência em comum entre as sessões s_e e s_f.

sim_LCS(s_e,s_f, 0.6)

Índice na sessão / i	$l^{s_e}(i)$	$l^{s_f}(i)$
1	1	1

LCS(s_e,s_f)

Índice na sessão / i	$l^{s_e}(i)$	$l^{s_f}(i)$
1	-	-

Figura 27 – Comparação do sim_LCS e LCS (nci=0)

Então, com base no peso escolhido e nos índices dos conceitos que fazem parte da subseqüência em comum obtida por sim_LCS, o mecanismo de agrupamento calcula a similaridade entre as sessões s_e e s_f aplicando o sim_WLCS. A similaridade obtida por sim_WLCS é 0.15 se considerado o peso pelo tempo de acesso, ou ainda 0.25 se considerado o peso binário. A similaridade entre as sessões s_e e s_f calculada por WLCS é de 0, pois neste caso não existe subseqüência em comum.

6.4.4.2 Nível Conceitual de Interesse 1

O passo (a) é realizado para o próximo nível conceitual de interesse (nci=1). As sessões s_e e s_f são traduzidas automaticamente pelo mecanismo de agrupamento para o nível conceitual de interesse 1. É importante notar que nos trabalhos de agrupamento existentes, descritos no capítulo 4, o analista precisa voltar à etapa inicial da MUW para gerar uma nova perspectiva dos dados.

Em seguida, o passo (b) é executado para gerar a matriz de similaridade. A subseqüência em comum encontrada entre as duas sessões, considerando sim_WLCS e WLCS, é identificada pelos índices dos conceitos dentro das sessões s_e e s_f, como ilustra a Figura 28. É importante notar também que, mesmo enriquecendo as sessões para o nível conceitual de interesse 1, o WLCS ainda não consegue encontrar uma subseqüência em comum entre as sessões s_e e s_f.

sim_LCS(s_e,s_f, 0.6)		
Índice na sessão / i	$l^{s_e}(i)$	$l^{s_f}(i)$
1	1	1
2	2	2

LCS(s_e,s_f)		
Índice na sessão / i	$l^{s_e}(i)$	$l^{s_f}(i)$
1	-	-

Figura 28 – Comparação do sim_LCS e LCS (nci=1)

Então, com base no peso escolhido e nos índices dos conceitos que fazem parte da subseqüência em comum comum obtida por sim_LCS, o mecanismo de agrupamento calcula a similaridade entre as sessões s_e e s_f aplicando o sim_WLCS. A similaridade obtida por sim_WLCS é 0.73 se considerado o peso pelo tempo de acesso, ou ainda 1 se considerado o peso binário. A similaridade entre as sessões calculada por WLCS é de 0, pois não existe subseqüência em comum.

6.4.4.3 Nível Conceitual de Interesse 2

O passo (a) é realizado para o próximo nível conceitual de interesse (nci=2). Seguindo o mesmo processo descrito anteriormente, as sessões s_e e s_f são traduzidas automaticamente pelo mecanismo de agrupamento para o nível conceitual de interesse 2.

Em seguida, o passo (b) é executado para gerar a matriz de similaridade. A subseqüência em comum encontrada entre as duas sessões, considerando sim_WLCS e WLCS, é identificada pelos índices dos conceitos dentro das sessões s_e e s_f, como ilustra a Figura 29. É importante notar que somente quando as sessões s_e e s_f são traduzidas para o nível conceitual de interesse 2 o WLCS consegue construir uma subseqüência em comum entre as sessões, como ilustra a Figura 29. Isso porque é possível identificar conceitos idênticos entre estas sessões somente abstraindo as mesmas dois níveis acima do nível conceitual base.

sim_LCS(s_e,s_f, 0.6)		
Índice na sessão / i	$l^{s_e}(i)$	$l^{s_f}(i)$
1	1	1
2	2	2

LCS(s_e,s_f)		
Índice na sessão / i	$l^{s_e}(i)$	$l^{s_f}(i)$
1	1	1
2	2	2

Figura 29 – Comparação do sim_LCS e LCS (nci=2)

Então, com base no peso escolhido e nos índices dos conceitos que fazem parte da subseqüência em comum comum obtida por sim_LCS, o mecanismo de agrupamento calcula a similaridade entre as sessões s_e e s_f aplicando o sim_WLCS. A similaridade obtida por sim_WLCS não sofre alteração em relação à similaridade obtida para o nível conceitual de interesse anterior (nci=1). Já a similaridade

calculada por WLCS é de 0.73, visto que a subseqüência em comum obtida é igual à do sim_WLCS. Neste nível conceitual de interesse não houve diferença na similaridade calculada pelo mecanismo de agrupamento proposto e pelo WLCS para as sessões s_e e s_f, pois os conceitos acessados nestas duas sessões são idênticos.

6.4.4.4 Nível Conceitual de Interesse 3 e 4

O agrupamento dinâmico pode continuar a ser executado para níveis de interesse mais elevados enquanto existirem conceitos dentro das sessões que ainda não atingiram o nível máximo de abstração na hierarquia conceitual. Isto é ilustrado Figura 26 na pela abstração de “Temas” para “Material de apoio” (nci=3), e de “Material de Apoio” para “Ambiente do curso” (nci=4). Apesar do enriquecimento das sessões continuar abstraindo um conceito em ambas as sessões s_e e s_f para os níveis conceituais de interesse 3 e 4, a similaridade obtida tanto para o sim_WLCS quanto para o WLCS permanece a mesma do nível conceitual de interesse 2. Isto porque a subseqüência em comum obtida e os pesos envolvidos neste não mudam em relação àqueles obtidos com nci=2.

6.4.5 Análise dos Resultados do Agrupamento

Para ilustrar como o processo de agrupamento dinâmico utilizando o sim_WLCS melhora a qualidade dos grupos em comparação à técnica WLCS, considera-se todas as sessões representadas no nível conceitual base, da Figura 21, limite de similaridade de 0.6, cinco níveis conceituais de interesse (nci={0,1,2,3,4}), a hierarquia conceitual ilustrada na Figura 15, e dois grupos para agrupar as sessões.

6.4.5.1 Nível Conceitual de Interesse 0

Os passos (a) e (b) resultam na matriz de similaridade ilustrada na Figura 30-A. Os campos em branco na matriz de similaridade significam similaridade zero.

(A) sim_WLCS(x, y, 0.6)

	s_a	s_b	s_c	s_d	s_e	s_f	s_g
s_a	1	0,35			0,29	0,33	
s_b	0,35	1			0,68		
s_c			1	0,36	0,29		0,19
s_d			0,36	1	0,13		0,60
s_e	0,29	0,68	0,29	0,13	1		0,13
s_f	0,33					1	
s_g			0,19	0,60	0,13		1

(B) WLCS(x, y)

	s_a	s_b	s_c	s_d	s_e	s_f	s_g
s_a	1				0,19	0,27	
s_b		1					
s_c			1	0,19			0,19
s_d			0,19	1	0,13		0,60
s_e	0,19			0,13	1		0,13
s_f	0,27					1	
s_g			0,19	0,60	0,13		1

Figura 30 – Comparação da matriz de similaridade (nci=0)

Pode-se observar que sim_WLCS encontra similaridade entre as sessões s_a e s_b, s_b e s_e, s_c e s_e, como ilustram os campos sombreados na Figura 30-A, pois estas

têm conceitos em comum que atingem o limite de similaridade definido para formar a subseqüência em comum. Já o WLCS não consegue encontrar nenhum grau de similaridade entre estas sessões, conforme ilustra a Figura 30-B. Além disso, o sim_WLCS melhorou o grau de similaridade entre as sessões s_a e s_e, s_a e s_f, s_c e s_d. O passo (c) é realizado e as sessões são agrupadas com base na matriz de similaridade obtida neste nível conceitual de interesse. A Figura 31 ilustra a formação dos grupos considerando a matriz de similaridade gerada por sim_WLCS e WLCS. Em negrito estão salientadas as diferenças entre os grupos.

sim_WLCS	WLCS
Material, Mapa do Curso e Introdução Grupo 0: s_c, s_d, s_g	Fórum, Correio Eletrônico, Material, Mapa do Curso Grupo 0: s_a, s_c, s_d, s_e, s_g, s_f
Ferramentas de Comunicação e Temas Grupo 1: s_a, s_b, s_e, s_f	Correio Eletrônico Grupo 1: s_b

Figura 31 – Comparação do resultado do agrupamento (nci=0)

Pode-se notar que utilizando sim_WLCS foi possível gerar um agrupamento de melhor qualidade, pois as sessões s_b, s_e e s_f com acessos principalmente a páginas de "Correio Eletrônico" e "Fórum" ficam no mesmo grupo que a sessão s_a com acesso a páginas do "Fórum". As sessões s_c, s_d e s_g com acessos a páginas de "Material do Curso", "Mapa do Curso", "Hipertexto: Sociedade do conhecimento" e "Palestra: Sociedade do conhecimento" ficam em um mesmo grupo. Já WLCS agrupa as sessões de forma que a sessão s_b, com acesso a "Correio Eletrônico", fica em um grupo separado das demais sessões com acesso a "Correio Eletrônico". Isto porque, em WLCS, a sessão s_b não tem similaridade com nenhuma outra sessão, conforme ilustra a Figura 30-B.

6.4.5.2 Nível Conceitual de Interesse 1

Os passos (a) e (b) resultam na matriz de similaridade ilustrada na Figura 32-A. Os campos em branco na matriz de similaridade significam similaridade zero.

(A) sim_WLCS(x, y, 0.6)

	s_a	s_b	s_c	s_d	s_e	s_f	s_g
s_a	1	0,42			0,36	0,28	
s_b	0,42	1			0,68		
s_c			1	0,36	0,29	0,17	0,19
s_d			0,36	1	0,13	0,21	0,60
s_e	0,36	0,68	0,29	0,13	1	0,15	0,13
s_f	0,28		0,17	0,21	0,15	1	0,21
s_g			0,19	0,60	0,13	0,21	1

(B) WLCS(x, y)

	s_a	s_b	s_c	s_d	s_e	s_f	s_g
s_a	1	0,42			0,36	0,28	
s_b	0,42	1			0,68		
s_c			1	0,36	0,29		0,19
s_d			0,36	1	0,13		0,60
s_e	0,36	0,68	0,29	0,13	1		0,13
s_f	0,28					1	
s_g			0,19	0,60	0,13		1

Figura 32 – Comparação da matriz de similaridade (nci=1)

Pode-se observar que sim_WLCS encontra similaridade entre as sessões s_f e s_c, s_f e s_d, s_f e s_e, s_f e s_g, como ilustram os campos sombreados na Figura 32-A, pois estas têm conceitos em comum que atingem o limite de similaridade definido para formar a subseqüência em comum. Já o WLCS não consegue encontrar nenhum grau de similaridade entre estas sessões, como ilustra a Figura 32-B.

O passo (c) é realizado e as sessões são agrupadas com base na matriz de similaridade obtida neste nível conceitual de interesse. A Figura 33 ilustra a formação dos grupos considerando o sim_WLCS e o WLCS. Em negrito estão salientadas as diferenças entre os grupos.

sim_WLCS	WLCS
Material, Utilidades, Introdução, Fórum e Dimensões em EAD Grupo 0: S_c, S_d, S_f , S_g	Material, Utilidades e Introdução Grupo 0: S_c, S_d, S_g
Fórum, Correio e Introdução Grupo 1: S_a, S_b, S_e	Fórum, Correio, Introdução e Dimensões em EAD Grupo 1: S_a, S_b, S_e, S_f

Figura 33 – Comparação do resultado do agrupamento (nci=1)

Em sim_WLCS a sessão s_f ficou fora do grupo 1, pois s_f apresenta grau de similaridade com todas as demais sessões, exceto com a sessão s_b. Ou seja, neste exemplo mesmo tendo uma similaridade alta com as sessões s_a e s_e (0,28 e 0,15 respectivamente) a repulsão com a sessão s_b levou a sessão s_f ir para o grupo 0. Já em WLCS o agrupamento ficou melhor caracterizado, pois a sessão s_f ficou no grupo 1 dado que s_f não apresenta similaridade com nenhuma outra sessão além de s_a.

6.4.5.3 Nível Conceitual de Interesse 2 e 3

A matriz de similaridade calculada por sim_WLCS para os níveis conceituais de interesse 2 e 3 (nci=2 e nci=3) são idênticas. Além disso, não existe diferença entre as matrizes de similaridade calculadas por sim_WLCS e por WLCS nestes níveis conceituais de interesse, como ilustra a Figura 34. Os campos em branco na matriz de similaridade significam similaridade zero.

(A) sim_WLCS(x, y, 0.6)

	S_a	S_b	S_c	S_d	S_e	S_f	S_g
S_a	1	0,21			0,18	0,20	
S_b	0,21	1			0,68	0,77	
S_c			1	0,36	0,29	0,17	0,19
S_d			0,36	1	0,13	0,21	0,60
S_e	0,18	0,68	0,29	0,13	1	0,73	0,13
S_f	0,20	0,77	0,17	0,21	0,73	1	0,21
S_g			0,19	0,60	0,13	0,21	1

(B) WLCS(x, y)

	S_a	S_b	S_c	S_d	S_e	S_f	S_g
S_a	1	0,21			0,18	0,20	
S_b	0,21	1			0,68	0,77	
S_c			1	0,36	0,29	0,17	0,19
S_d			0,36	1	0,13	0,21	0,60
S_e	0,18	0,68	0,29	0,13	1	0,73	0,13
S_f	0,20	0,77	0,17	0,21	0,73	1	0,21
S_g			0,19	0,60	0,13	0,21	1

Figura 34 – Comparação da matriz de similaridade (nci=2 e nci=3)

Conseqüentemente, os grupos resultantes de sim_WLCS e de WLCS são idênticos, como ilustra a Figura 35. É importante notar que é possível que, em um determinado nível conceitual de interesse, não exista diferença entre sim_WLCS e WLCS devido à abstração dos conceitos para um nível mais generalizado na hierarquia conceitual.

sim_WLCS	WLCS
Utilidades, Temas e Ambiente do Curso Grupo 0: S_c, S_d, S_g	Utilidades, Temas e Ambiente do Curso Grupo 0: S_c, S_d, S_g
Ferramentas de Comunicação, Atividades e Temas Grupo 1: S_a, S_b, S_e, S_f	Ferramentas de Comunicação, Atividades e Temas Grupo 1: S_a, S_b, S_e, S_f

Figura 35 – Comparação do resultado do agrupamento (nci=2 e nci=3)

6.4.5.4 Nível Conceitual de Interesse 4

Os passos (a) e (b) resultam na matriz de similaridade ilustrada na Figura 36-A. Os campos em branco na matriz de similaridade significam similaridade zero.

(A) sim_WLCS(x, y, 0.6)

	S_a	S_b	S_c	S_d	S_e	S_f	S_g
S_a	1	0.21			0.18	0.20	
S_b	0.21	1			0.68	0.77	
S_c			1	0.49	0.40	0.15	0.49
S_d			0.49	1	0.50	0.30	1
S_e	0.18	0.68	0.40	0.50	1	0.73	0.50
S_f	0.20	0.77	0.15	0.30	0.73	1	0.30
S_g			0.49	1	0.50	0.30	1

(B) WLCS(x, y)

	S_a	S_b	S_c	S_d	S_e	S_f	S_g
S_a	1	0.21			0.18	0.20	
S_b	0.21	1			0.68	0.77	
S_c			1	0.48	0.24	0.15	0.49
S_d			0.49	1	0.50	0.30	1
S_e	0.18	0.68	0.40	0.50	1	0.73	0.50
S_f	0.20	0.77	0.15	0.30	0.73	1	0.30
S_g			0.49	1	0.50	0.30	1

Figura 36 – Comparação da matriz de similaridade (nci=4)

A matriz de similaridade calculada por sim_WLCS neste nível conceitual de interesse é idêntica à matriz de similaridade calculada por WLCS (Figura 36-B), como ilustra a Figura 36. Conseqüentemente, os grupos resultantes de sim_WLCS e de WLCS são idênticos, como ilustra a Figura 37.

sim_WLCS	WLCS
Utilidades e Ambiente do Curso Grupo 0: S_c, S_d, S_g	Utilidades e Ambiente do Curso Grupo 0: S_c, S_d, S_g
Ferramentas de Comunicação, Ambiente do Curso e Atividades Grupo 1: S_a, S_b, S_e, S_f	Ferramentas de Comunicação, Ambiente do Curso e Atividades Grupo 1: S_a, S_b, S_e, S_f

Figura 37 – Comparação do resultado do agrupamento (nci=4)

6.4.6 Conclusão

Considerar a similaridade quando computando a subseqüência em comum melhorou a qualidade do agrupamento para as sessões representadas no nível conceitual mais especializado da hierarquia do domínio. A melhora da qualidade do agrupamento foi medida nesta análise comparativa pelas características dos grupos formados considerando o sim_WLCS em comparação ao WLCS.

Outro fato observado está relacionado com o erro introduzido no agrupamento das sessões dependendo do nível de abstração das sessões. Pois, quanto mais elevado o nível conceitual de interesse em que as sessões são representadas na hierarquia conceitual do domínio, mais as sessões tendem a serem similares umas às outras. Entretanto, deve-se observar que este erro pode acontecer até mesmo nos trabalhos de agrupamento existentes, descritos no capítulo 4, quando o analista precisa voltar à etapa inicial da MUW para gerar manualmente uma nova perspectiva dos dados.

Observou-se também que quando as sessões têm similaridade com muitas outras sessões o agrupamento tende a ser definido por suas diferenças e não pelas suas características em comum. Assim, a definição de um limite de similaridade entre conceitos muito baixo (próximo de zero) pode resultar em sessões com alto grau de semelhança entre si, tendendo a formar grupos de má qualidade. A má qualidade dos grupos dificulta, por conseqüência, a interpretação e caracterização das sessões de aprendizado.

7 MECANISMO DE INTERPRETAÇÃO

Este capítulo apresenta o mecanismo de interpretação proposto, o qual representa visualmente os padrões resultantes do agrupamento e oferece facilidades para interpretação dos grupos considerando os diferentes níveis de abstração dos conceitos no domínio da aplicação.

O mecanismo de interpretação proposto permite representar os grupos resultantes de maneira condizente com os objetivos da mineração, inspecioná-los de acordo com o nível conceitual de interesse desejado, bem como interpretá-los em termos da abstração das sessões que compõem o grupo. Estes dois últimos aspectos são complementares à geração dos grupos por nível conceitual de interesse. Esta abordagem foi definida visando complementar as carências apresentadas pelos trabalhos relatados no Capítulo 4, e está fundamentada em três componentes principais: visualização, inspeção dos grupos, e interpretação dinâmica.

O componente de visualização permite representar os grupos visualmente, com base no objetivo do agrupamento. O componente de inspeção dos grupos permite escolher o grupo e o nível conceitual de interesse segundo os quais as sessões foram agrupadas durante a etapa de descoberta de padrões. O componente de interpretação dinâmica permite que as sessões que caracterizam um grupo previamente selecionado sejam abstraídas para outros níveis de abstração na hierarquia.

Detalhes sobre cada um dos componentes utilizados na abordagem proposta pelo mecanismo de interpretação são descritos no restante deste capítulo.

7.1 Tipos de Visualização

7.1.1 Visualização do Agrupamento de Interesse

O mecanismo de interpretação utiliza o perfil agregado, conforme discutido na seção 4.1.2, para representar visualmente as características comuns das sessões que compõem cada grupo. Esta forma de visualização visa facilitar a interpretação dos grupos resultantes do agrupamento de interesse, uma vez que o perfil agregado descreve cada grupo pela média consolidada dos conceitos das sessões pertencentes ao grupo. Além disso, o mecanismo de interpretação oferece as seguintes operações sobre os conceitos do perfil agregado: a) filtro de importância, e b) ordenação pelo peso.

- Filtro de importância: filtra os conceitos do perfil agregado com base em um valor mínimo de peso estipulado pelo analista. Caso contrário o conceito não é

visualizado no perfil agregado. Se o objetivo do agrupamento utiliza o peso binário, então o valor mínimo de peso representa a percentagem mínima de sessões que o conceito deve aparecer para permanecer no perfil agregado. Se o objetivo do agrupamento utiliza o peso pelo tempo de acesso, então o valor mínimo de peso representa o mínimo de tempo médio de acesso que o conceito deve atingir para permanecer no perfil agregado;

- Ordem pelo peso: ordena os conceitos pertencentes ao perfil agregado considerando o peso atribuído à cada conceito.

As operações de filtragem e ordenação realizadas sobre os conceitos do perfil consolidado visam, respectivamente, reduzir o número de conceitos e ordenar os mesmos, com base na importância associada ao peso, de modo a facilitar a caracterização do grupo. Por exemplo, considerando o grupo 1 contendo as sessões s_1, s_2 e s_3 representadas no nível conceitual básico e formatadas para o agrupamento de interesse, como ilustra a Figura 38-A, um filtro de importância de 15 segundos, e ordem pelo peso, a visualização oferecida pelo mecanismo de interpretação para o grupo 1 é ilustrado na Figura 38-B. Neste exemplo, os conceitos do perfil agregado estão ordenados pela ordem decrescente do peso, e o conceito c_{121} foi removido do perfil agregado pois seu peso não atinge o valor mínimo estabelecido pelo filtro de importância. A Figura 38-C ilustra o perfil consolidado considerando o peso binário e um limite de importância de 50%.

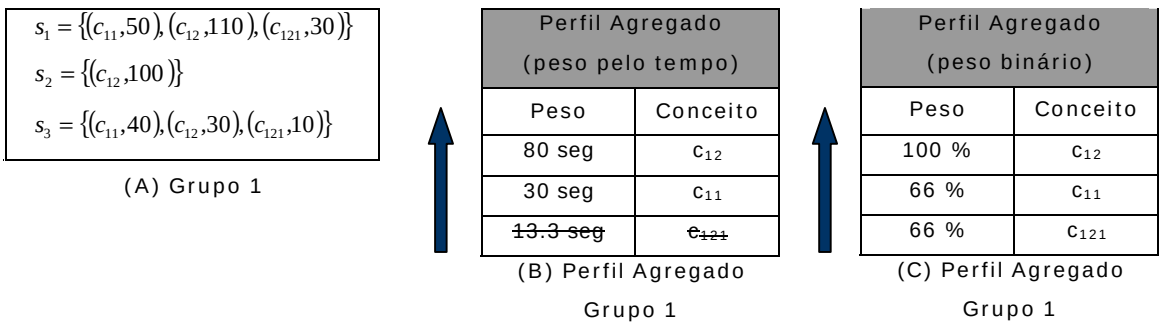

Figura 38 – Exemplo de visualização do agrupamento de interesse

7.1.2 Visualização do Agrupamento de Trajetória

O mecanismo de interpretação utiliza de forma integrada a árvore agregada apresentada na seção 4.2.2 e o perfil agregado como uma forma de salientar, além do caminho de navegação, os interesses em comum que caracterizam as sessões pertencentes ao grupo. Embora a árvore agregada ofereça ao analista dados sobre a trajetória do grupo, dependendo do número de ramificações, a interpretação desta

pode ficar muito complexa. Por exemplo, considerando o grupo 1 contendo as sessões s_1, s_2 e s_3 representadas no nível conceitual básico sem redução do caminho de navegação, como ilustra a Figura 39-A, um filtro de importância de 10 segundos, e ordem pelo peso, a visualização do agrupamento de trajetória oferecida pelo mecanismo de interpretação é ilustrada pela Figura 39-B. Neste exemplo, a importância do conceito c_{12} observada de forma distribuída na árvore agregada (pelas ocorrências e pesos associados) é refletida de forma consolidada no perfil agregado que caracteriza o grupo. Todos os conceitos acessados nas sessões do grupo 1 estão presentes no perfil agregado, pois todos atingem o valor mínimo estabelecido pelo filtro de importância.

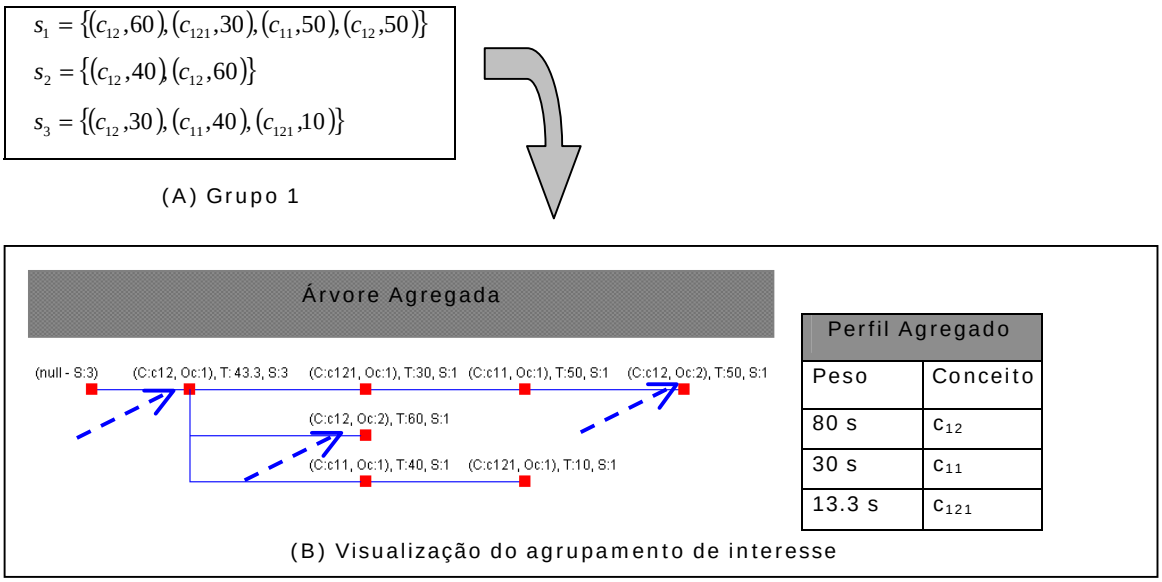

Figura 39 – Exemplo de visualização do agrupamento de trajetória

7.2 Inspeção dos Grupos

O mecanismo de agrupamento traduz as sessões, gera matrizes de similaridade e realiza o agrupamento para cada nível conceitual de interesse. Portanto existe um conjunto de grupos próprios, disponibilizados pelo mecanismo de agrupamento, para cada nível conceitual de interesse. Resta então ao mecanismo de interpretação possibilitar a seleção, visualização e interpretação dos grupos formados para cada nível de abstração.

O mecanismo de interpretação viabiliza a inspeção dos grupos para cada nível de abstração através de dois operadores gráficos: a) seletor do nível conceitual de interesse, e b) seletor do grupo. A combinação destes operadores possibilita a visualização do grupo selecionado no nível conceitual de interesse desejado, segundo o componente de visualização apresentado na seção 7.1.

- Seletor do nível conceitual de interesse: este operador permite que o analista escolha o nível conceitual de interesse segundo o qual as sessões foram agrupadas. A mudança neste operador implica na visualização de um diferente resultado de agrupamento, pois a formação os grupos pode variar de acordo com a similaridade entre as sessões no nível de interesse em questão, conforme discutido na 6.3;
- Seletor do grupo: este operador permite que o analista escolha, dentre os grupos pertencentes ao nível conceitual de interesse selecionado, qual o grupo deseja visualizar.

Os seletores de grupo e do nível conceitual de interesse viabilizam a inspeção do agrupamento, ou seja, através de interações com estes operadores o analista pode escolher qual grupo deseja visualizar em termos do nível de abstração que as sessões foram agrupadas pelo mecanismo de agrupamento. Por exemplo, considerando três possíveis níveis conceituais de interesse (nci), cinco sessões (s_1, s_2, s_3, s_4 e s_5), e três grupos, a Figura 40 ilustra as diferentes formação dos grupos, de acordo com a similaridade entre as sessões, disponibilizadas pelo mecanismo de agrupamento para cada nível conceitual de interesse.

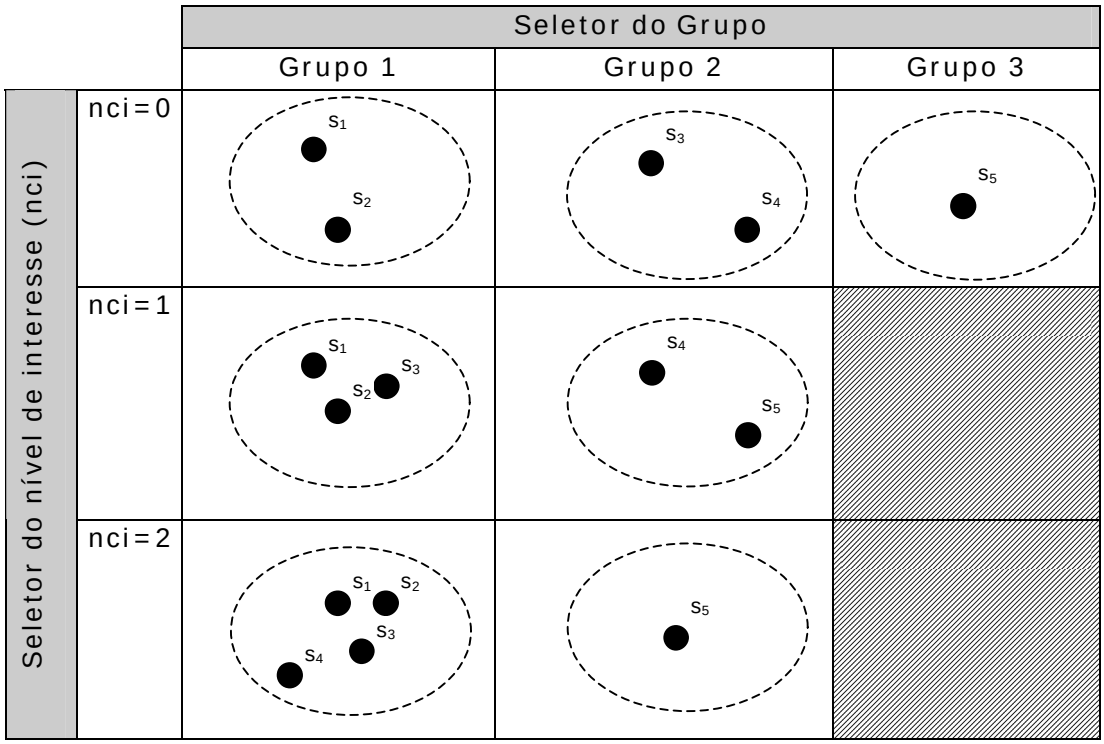

Figura 40 – Exemplo do conjunto de grupos disponibilizados pelo mecanismo agrupamento

7.3 Interpretação Dinâmica

A interpretação dinâmica permite que o analista interprete o grupo através do enriquecimento dinâmico das sessões que compõem o grupo. O mecanismo de interpretação viabiliza a interpretação dinâmica através de dois operadores gráficos: a) roll-up, e b) drill-down. As operações de roll-up e drill-down foram definidas em analogia às operações de mesmo nome propostas pelo trabalho de [VAN04a] em termos da abstração dos conceitos das sessões que caracterizam cada grupo. O trabalho de [VAN04a] aplica estas operações para substituição de um ou mais conceitos dentro de um padrão seqüencial visando obter um padrão conceitual abstrato que sumarize todos os padrões seqüenciais específicos relacionados. Já o mecanismo de interpretação proposto permite que estes operadores sejam aplicados, independentemente do nível conceitual de interesse no qual as sessões foram agrupadas, visando apenas facilitar a interpretação dos grupos por pessoas leigas. É importante salientar que os operadores de roll-up e drill-down não alteram os padrões resultantes do agrupamento, uma vez que estes são aplicados sobre as sessões já pertencentes aos grupos.

7.3.1 Operador de Roll-up

Este operador permite que todas as sessões que caracterizam o grupo sejam abstraídas por seus respectivos conceitos ascendentes na hierarquia conceitual. A operação de roll-up pode ser aplicada recursivamente ao conjunto de conceitos das sessões que caracterizam o grupo até chegar ao nível de abstração desejado, obedecendo o nível máximo de abstração de cada conceito. O operador de roll-up é desabilitado graficamente quando não existe mais abstração possível para o conjunto de conceitos que caracterizam o grupo.

A operação de roll-up pode ser empregada em ambos os tipos de visualização dos grupos discutidos na seção 7.1. Quando aplicada na visualização do agrupamento de interesse então ocorre o roll-up do perfil agregado. Caso contrário, quando aplicada na visualização do agrupamento de trajetória, ocorre o roll-up do perfil agregado e da árvore agregada ao mesmo tempo.

- Roll-up do perfil agregado: substitui cada conceito pertencente ao perfil agregado pelo seu ascendente direto na hierarquia conceitual, unificando os conceitos repetidos. Se o peso pelo tempo de acesso foi escolhido, então os tempos são somados para os conceitos unificados no perfil agregado abstrato. Se o peso binário foi escolhido os conceitos repetidos são simplesmente unificados e seu peso no perfil agregado recalculado para gerar o perfil

agregado abstrato. Quando aplicado a um perfil agregado representado no nível conceitual base, a operação de roll-up tem como resultado um perfil agregado abstrato. Os operadores de visualização (filtro de importância e ordem pelo peso) continuam disponíveis no perfil agregado abstrato. Por exemplo, considerando o perfil agregado representado no nível conceitual base, como ilustra a Figura 41-A, e o peso pelo tempo, a operação de roll-up (simbolizada pelo sinal de "+") abstrai os três conceitos c_{12}, c_{11} e c_{121} pelos seus respectivos ascendentes na hierarquia conceitual c_1, c_1 e c_{12}, sendo que os conceitos repetidos c_1 são unificados e seus tempos somados, gerando assim um perfil agregado abstrato, como ilustra a Figura 41-B. Uma nova operação de roll-up aplicada sobre o perfil agregado abstrato da Figura 41-B abstrai os conceitos para mais um nível de ascendente na hierarquia, como ilustra a Figura 41-C. Na segunda operação de roll-up não houve a unificação dos conceitos uma vez que os mesmos não se repetiram neste nível de abstração. As Figuras 45-D, 45-E, 45-F ilustram o mesmo processo para o peso binário;

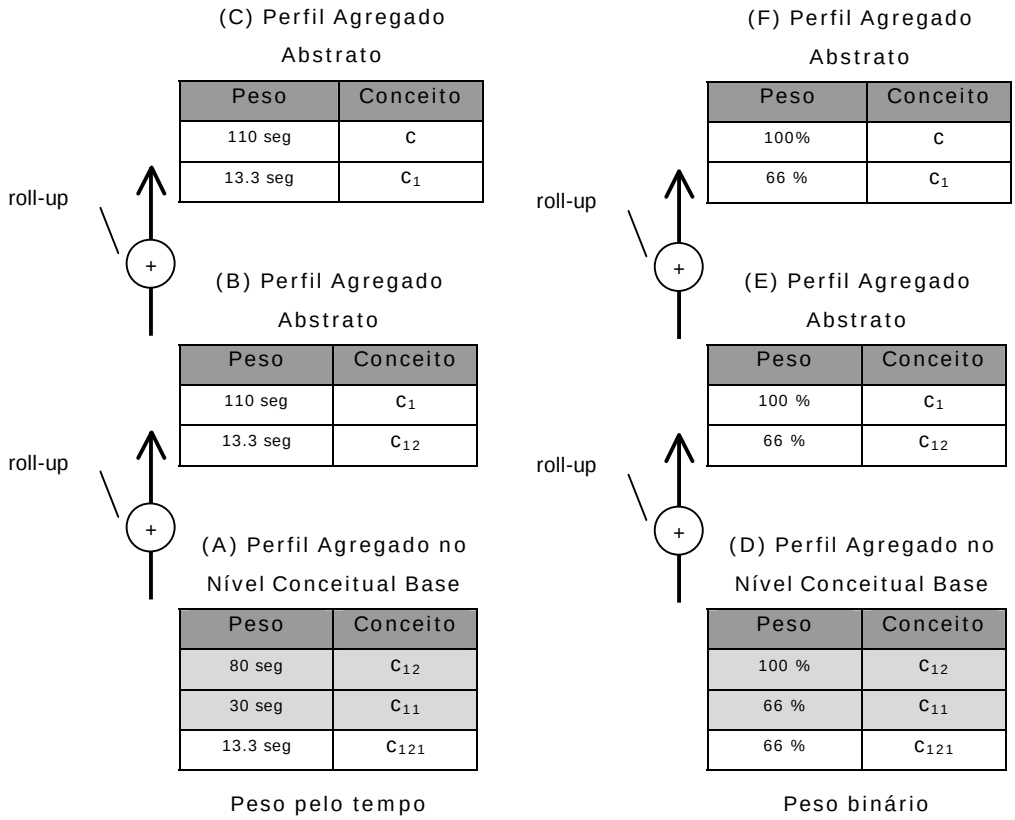

Figura 41 – Exemplo da operação de roll-up no perfil agregado

- Roll-up da árvore agregada: substitui os conceitos no caminho de navegação que identifica o grupo pelo seu ascendente direto na hierarquia conceitual. Quando aplicado a uma árvore agregada representada no nível conceitual base, a operação de roll-up tem como resultado uma árvore agregada abstrata. Por exemplo, considerando a árvore agregada representada no nível conceitual base, como ilustra a Figura 42-A, peso pelo tempo e redução do caminho de navegação, a operação de roll-up (simbolizada pelo sinal de "+") abstrai os conceitos das sessões s_1, s_2 e s_3 pelos seus respectivos ascendentes na hierarquia conceitual unificando conceitos contíguos, gerando assim uma árvore agregada abstrata, como ilustra a Figura 42-B, que representa trajetórias mais simplificadas. Uma nova operação de roll-up aplicada sobre uma árvore agregada abstrata, ilustrada na Figura 42-B, generaliza os conceitos um nível de ascendente na hierarquia, como ilustra a Figura 42-C.

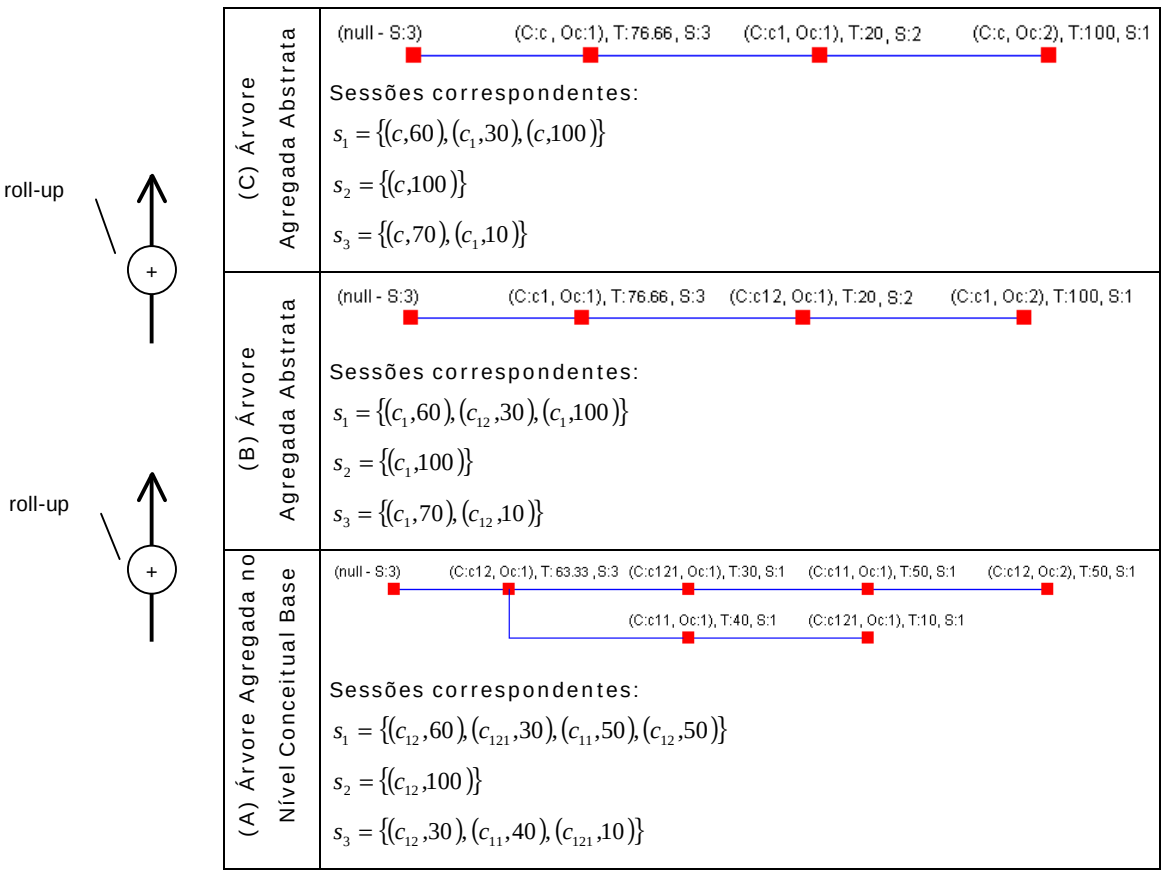

Figura 42 – Exemplo da operação de roll-up na árvore agregada

7.3.2 Operador de Drill-down

Este operador permite que a visualização do grupo retorne um nível de abstração na hierarquia. A operação de drill-down pode ser aplicada recursivamente

sobre a visualização de um perfil agregado abstrato e de uma árvore agregada abstrata, até que estes cheguem respectivamente no nível conceitual base de representação do grupo. Ou seja, o operador de drill-down é desabilitado graficamente quando o grupo está representado no nível conceitual mais especializado do grupo.

- Drill-down do perfil agregado: retorna um nível de abstração na visualização do perfil agregado. Os operadores de visualização (filtro de importância e ordem pelo peso) continuam disponíveis;
- Drill-down da árvore agregada: retorna um nível de abstração na visualização do caminho de navegação que identifica o grupo representado pela árvore agregada.

7.3.3 Complementariedade dos Operadores

Na prática os operadores de drill-up e drill-down funcionam em complementação um ao outro, independente do nível conceitual de interesse no qual as sessões foram agrupadas. Por exemplo, considerando o mesmo grupo 1 contendo as sessões s_1, s_2 e s_3, representadas no nível conceitual básico, ilustradas na Figura 43-A, um filtro de importância de 10 segundos, ordem pelo peso e redução no caminho de navegação, os operadores de roll-up e drill-down na visualização do agrupamento de trajetória são ilustrados pela Figura 43-B. Neste exemplo, cada vez que a operação de roll-up é aplicada, os conceitos das sessões que representam respectivamente o perfil agregado e a árvore agregada do grupo são abstraídos por seus ascendentes, gerando assim um perfil agregado abstrato e uma árvore agregada abstrata. Já quando a operação de drill-down é aplicada, os conceitos das sessões representados no perfil agregado abstrato e na árvore agregada abstrata voltam para o nível de abstração anterior.

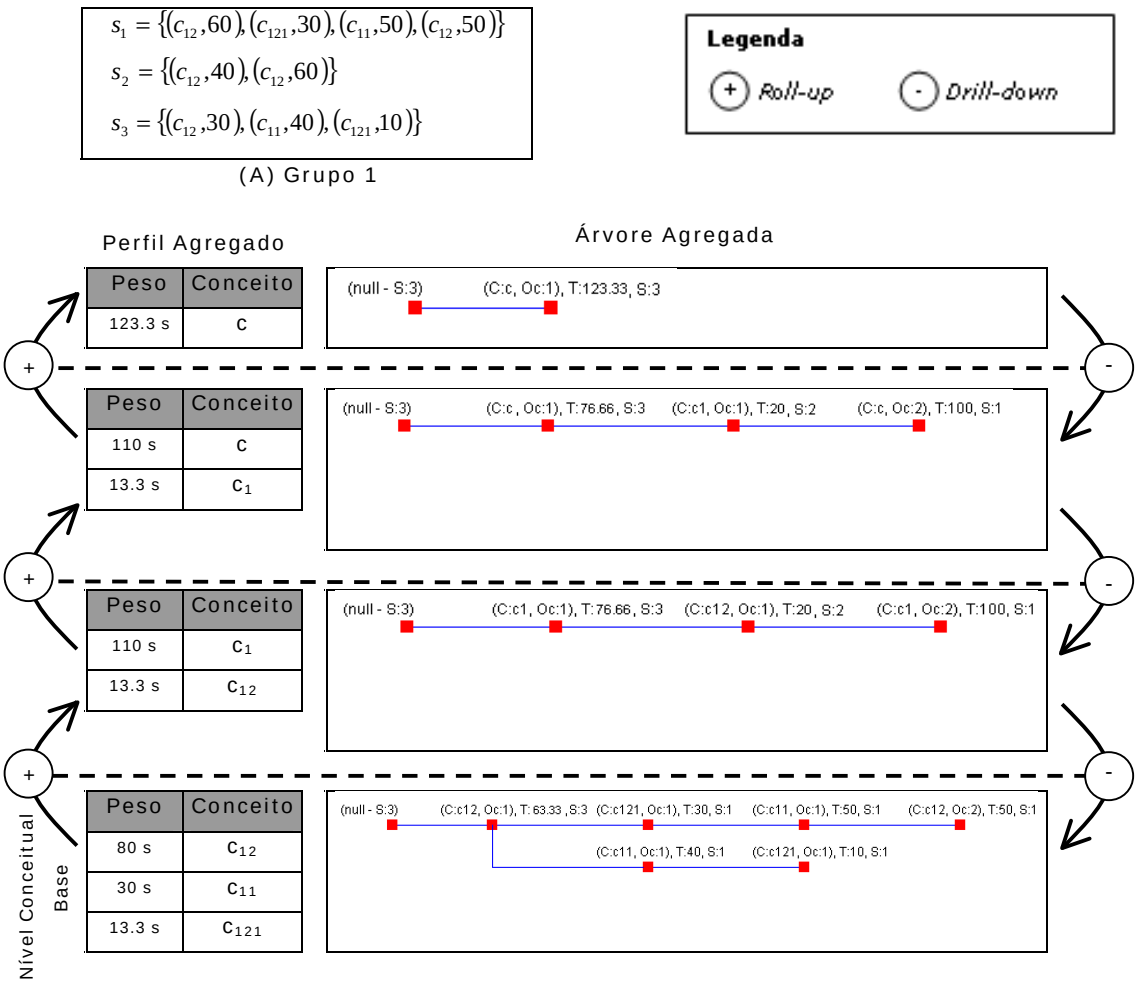

Figura 43 – Exemplo da combinação das operações de roll-up e drill-down

8 AMBIENTE DE APOIO À CARACTERIZAÇÃO DE SESSÕES

Este capítulo descreve o ambiente de apoio à execução das fases da MUW que incorpora os mecanismos de agrupamento e interpretação, bem como o protótipo desenvolvido que implementa estes mecanismos.

O ambiente apresentado neste capítulo, chamado de ambiente de apoio à caracterização de sessões de aprendizado, incorpora os mecanismos de agrupamento e interpretação discutidas nos capítulos 6 e 7, respectivamente, e serve como apoio à execução das fases do processo de MUW, a saber: preparação dos dados (pré-processamento), agrupamento (descoberta de padrões) e interpretação dos grupos (análise de padrões).

Um protótipo que implementa os mecanismos de agrupamento e interpretação foi desenvolvido. A construção deste protótipo visa a avaliação a execução e eficiência dos mecanismos de agrupamento e interpretação.

Uma visão geral do ambiente proposto, sua arquitetura, suas funcionalidades, bem como o protótipo que implementa os mecanismos de agrupamento e interpretação são descritos nas próximas seções.

8.1 Arquitetura do Ambiente

A arquitetura proposta para o ambiente de apoio à caracterização de sessões de aprendizado, ilustrada na Figura 45, representa como as entradas necessárias e os módulos desta arquitetura estão distribuídos pelas fases da MUW de modo a atender as funcionalidades dos mecanismos de agrupamento e interpretação.

O ambiente é dividido em 3 módulos principais: preparação dos dados, agrupamento e interpretação. Cada módulo suporta funcionalidades específicas, disponibilizadas em diferentes áreas da interface gráfica da aplicação, de acordo com a fase da MUW.

8.1.1 Módulo de Preparação dos Dados

Este módulo é responsável, respectivamente, pela execução de tarefas clássicas de pré-processamento [COO99] e de transformação de sessões [BAN01, FU00, MOB01, MOB04] visando obter grupos de melhor qualidade.

Como parte das tarefas típicas de pré-processamento os conceitos são classificados de acordo com o tempo de acesso (auxiliares e conteúdo) e as sessões, ou transações, são classificadas em [COO99]: sessões de conteúdo ou sessões

auxiliar-conteúdo. O agrupamento de interesse utiliza sessões de conteúdo e o agrupamento de trajetória utiliza sessões do tipo auxiliar-conteúdo.

8.1.2 Módulo de Agrupamento

O módulo de agrupamento é responsável pelo cálculo de similaridade entre as sessões, geração da matriz de similaridade e agrupamento das sessões para cada nível conceitual de interesse.

Este módulo considera a semântica do domínio, representada pela hierarquia conceitual do domínio, para computar a similaridade entre os conceitos em termos de sua localização na hierarquia e construir uma matriz de similaridade para cada nível conceitual de interesse que as sessões podem ser traduzidas. O conjunto de padrões resultante do agrupamento, considerando cada nível conceitual de interesse, é disponibilizado para o módulo de interpretação.

8.1.3 Módulo de Interpretação

Este módulo permite visualizar os grupos resultantes, considerando cada nível conceitual de interesse no qual as sessões foram agrupadas, bem como interpretar o grupo através do enriquecimento dinâmico das sessões que compõem o grupo. O perfil agregado é utilizado para visualizar o agrupamento de interesse e a árvore agregada mais o perfil agregado são utilizados para visualizar o agrupamento de trajetória.

8.1.4 Entradas Externas do Ambiente

8.1.4.1 Arquivo de Acesso

Os arquivos de acesso armazenam todas as requisições HTTP geradas durante a navegação no site Web. Os arquivos de acesso podem utilizar diversos formatos, como por exemplo, Extended Common Log Format (ECLF) [W3C05], e geralmente são provenientes dos servidores Web (log), conforme já mencionado na seção 3.1.1. O arquivo de acesso deve conter no mínimo informações sobre a identificação do usuário, data e hora de acesso, método de acesso (GET, POST ou HEAD), a URL acessada, bem como o código de resposta do servidor Web.

8.1.4.2 Base de Conhecimento

A base de conhecimento armazena informações da hierarquia do domínio em termos dos relacionamentos de generalização/especialização entre os conteúdos e

serviços disponibilizados pelo curso Web, bem como do mapeamento das URLs representadas no nível físico para os respectivos conceitos na hierarquia conceitual. A base de conhecimento foi implementada como um banco de dados Microsoft Access. O esquema da base de conhecimento está representado na Figura 44.

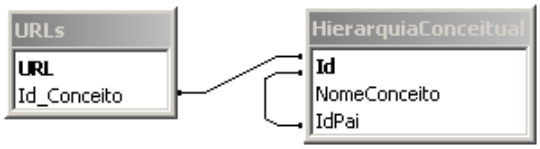

Figura 44 – Esquema da base de conhecimento

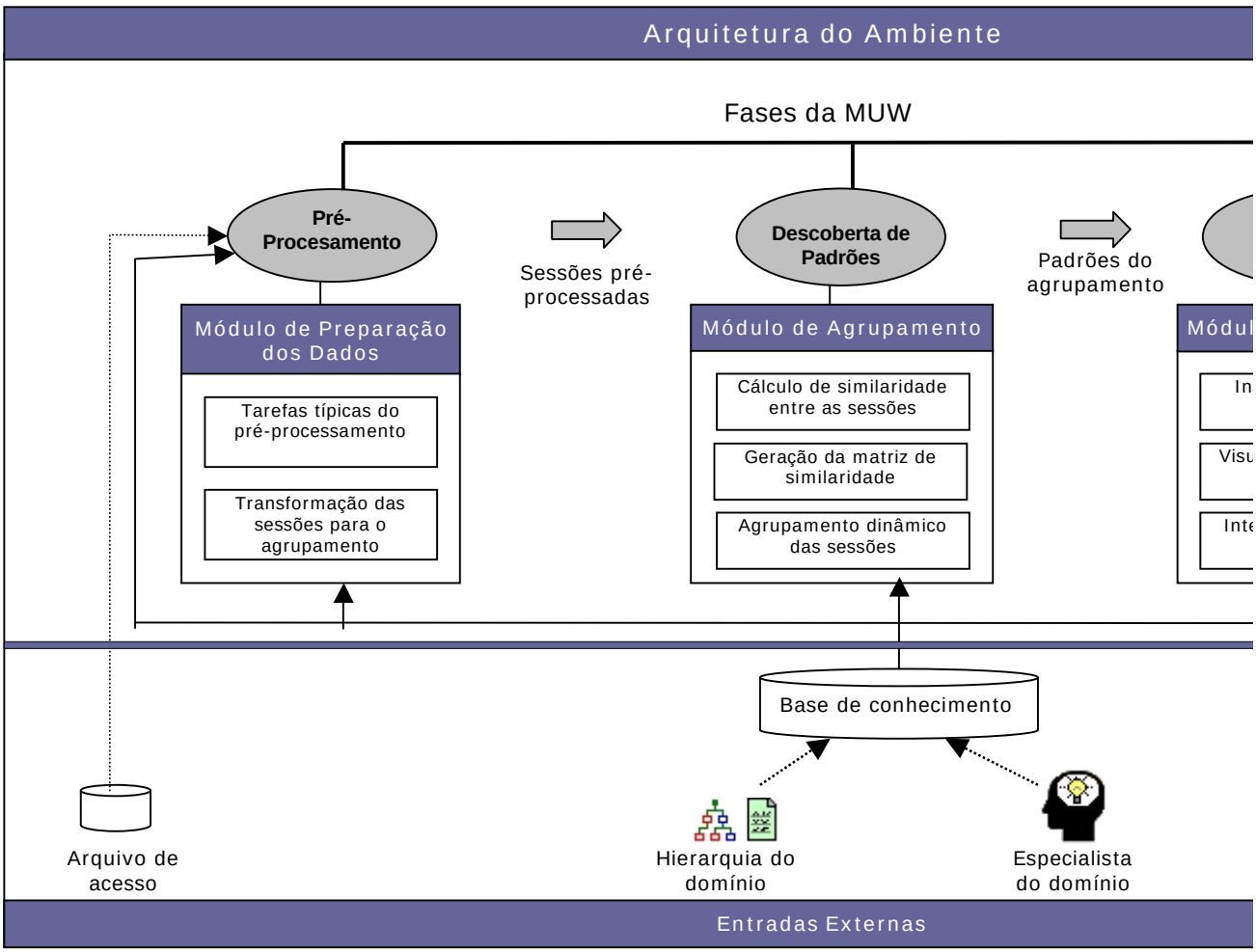

Figura 45 – Arquitetura do ambiente

8.2 Implementação

Esta seção descreve os detalhes relacionados aos aspectos práticos de implementação do ambiente de apoio à caracterização de sessões de aprendizado, bem como as funcionalidades expostas pela interface gráfica da aplicação.

Os mecanismos de agrupamento e interpretação propostos neste trabalho foram implementados em um protótipo chamado "Agrupamento para Caracterização de Sessões de Aprendizado" (ACSA). ACSA foi desenvolvido utilizando a linguagem de programação Java, e oferece funcionalidades gráficas que implementam os mecanismos de agrupamento e interpretação propostos.

Por uma questão de praticidade as tarefas de cálculo de similaridade entre as sessões e geração da matriz de similaridade, originalmente destinadas ao Módulo de Agrupamento, foram implementadas no Módulo de Preparação dos Dados.

8.2.1 Módulo de Preparação dos Dados

A Figura 46 esquematiza as entradas e saídas do Módulo de Preparação dos Dados implementado como uma extensão de LogPrep [MAR04b] (apresentado na seção 4.3). A ferramenta LogPrep foi estendida como forma de viabilizar a implementação do ambiente de apoio em termos das tarefas relacionadas à etapa de preparação dos dados para o agrupamento de sessões. LogPrep foi escolhido pela facilidade em adicionar novas tarefas e operadores.

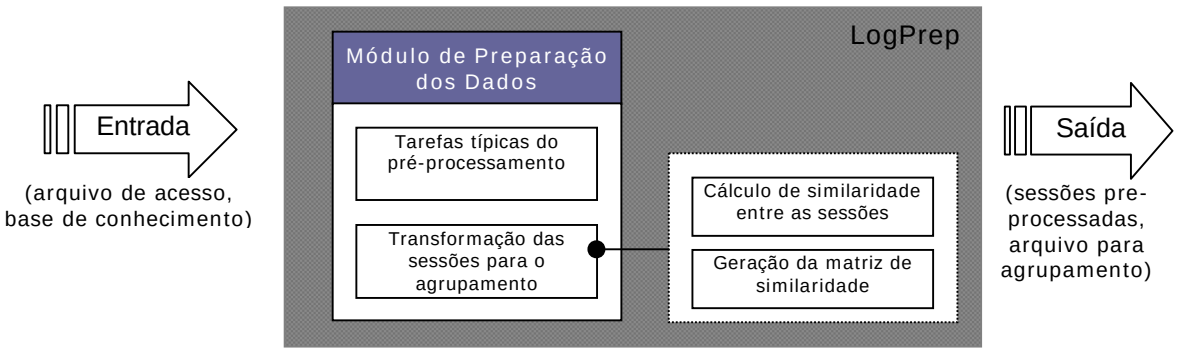

Figura 46 – Entradas e saídas de LogPrep

8.2.1.1 Arquivo de Entrada para Agrupamento

O Módulo de Preparação dos Dados, implementado através de LogPrep, além das sessões pré-processadas, gera um conjunto de informações que serve como entrada para ACSA. Este conjunto de informações contém, para cada nível conceitual de interesse, as sessões enriquecidas dinamicamente (formatadas de acordo com o objetivo do agrupamento) e a matriz de similaridade. Além disso, contém informações

relativas à preparação dos dados, a saber: objetivo do agrupamento, número de sessões identificadas, peso escolhido, similaridade entre os conceitos e máximo de abstração. O conteúdo das sessões identificadas deve conter, para cada nível conceitual de interesse: informações sobre o identificador da sessão do usuário, tempo de acesso, mapeamento da URL para o respectivo conceito no nível conceitual de acordo com a tabela HierarquiaConceitual (nome do conceito e ID). O conjunto de informações resultante do pré-procesamento é organizado em um arquivo XML, como ilustra a Figura 47, que é uma instância do XML Schema esquema denominado agrupamento.xsd (Anexo A deste volume). A Tabela 6 descreve cada um dos elementos do arquivo XML

```
<?xml version="1.0" encoding="UTF-8" ?>
- <dataset xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:noNamespaceSchemaLocation="file:///./agrupamento.xsd">
  <agrupamento>"Trajetoria"</agrupamento>
  <tipo_peso>"Tempo"</peso>
  <num_sessoes>7</num_sessoes>
  <similaridade>0,6</similaridade>
  <max_abstracao>2</max_abstracao>
- <nivel nci="0">
    <matriz>matriz-trajetoria_0.txt</matriz>
    <rclass>trajetoria_0.rclass</rclass>
    <clabel>trajetoria_0.clabel</clabel>
    <sessoes_AA>sessoes_AA_0.txt</sessoes_AA>
    <sessoes_PA>sessoes_PA_0.txt</sessoes_PA>
  </nivel>
+ <nivel id="1">
+ <nivel id="2">
</dataset>
```

Figura 47 – Exemplo do conjunto de informações para ACSA

Tabela 6 – Detalhes do arquivo XML

Elemento	Descrição
<dataset>	Elemento que identifica o conjunto de informações.
<agrupamento>	Tipo de agrupamento: Trajetória ou Interesse
<tipo_peso>	Tipo de peso: Binário ou Tempo
<num_sessoes>	Número de sessões identificadas e formatadas.
<similaridade>	Nível de similaridade definido entre as sessões (0,0 – 1,0).
<max_abstracao>	Máximo de abstração definido.
<nivel nci="X">	X é o Nível Conceitual de Interesse no qual as sessões foram traduzidas.
<matriz>	Nome do arquivo com a matriz de similaridade entre as sessões no nci.
<rclass>	Nome do arquivo com a identificação das sessões.
<clabel>	Nome do arquivo com a identificação dos conceitos existentes na hierarquia conceitual.
<sessoes_AA>	Nome do arquivo com as sessões formatadas para visualização com a Árvore Agregada.
<sessoes_PA>	Nome do arquivo com as sessões formatadas para visualização com o Perfil Agregado.

8.2.1.2 Funcionalidades

A Tabela 7 descreve as funcionalidades específicas da etapa de pré-processamento disponibilizadas na ferramenta LogPrep.

Tabela 7 – Funcionalidades do Módulo de Preparação dos Dados

Funcionalidade	Descrição
(A) Realizar tarefas típicas de pré-processamento	O analista define as tarefas de pré-processamento serão executadas: coleta dos dados, limpeza, filtragem, identificação das sessões, classificação das páginas, classificação das sessões (conteúdo ou auxiliar-conteúdo), enriquecimento das sessões (mapeamento das URLs para conceitos na hierarquia), etc.
(B) Definir o objetivo do agrupamento	O analista define qual o tipo de agrupamento será realizado. O agrupamento de interesse requer que as sessões sejam classificadas em conteúdo, já o agrupamento de trajetória requer que as sessões sejam classificadas em auxiliar-conteúdo.
(C) Realizar tarefas de transformação das sessões	De acordo com o objetivo do agrupamento o analista define qual as tarefas de transformação das sessões serão realizadas: remoção de sessões, remoção de acessos das sessões, redução do caminho de navegação, escolha do tipo do peso (binário ou pelo tempo).
(D) Gerar e exportar conjunto de informações para o agrupamento	O analista exporta o conjunto de informações gerado durante a preparação dos dados. Este conjunto de informações é a entrada para ACSA.

A ferramenta LogPrep já suporta a funcionalidade (A) da Tabela 7. Assim, foi necessário implementar uma nova tarefa, chamada “Transformação”, na ferramenta LogPrep para abranger os operadores de transformação das sessões relacionados à definição do objetivo do agrupamento (B), preparação das sessões (C) e exportação das sessões (D) para o Módulo de Agrupamento. A Figura 48 ilustra como as funcionalidades (A), (B), (C) e (D) da Tabela 7 estão organizadas na interface gráfica de LogPrep, por exemplo, considerando a transformação das sessões para o agrupamento de trajetória (Figura 48-B).

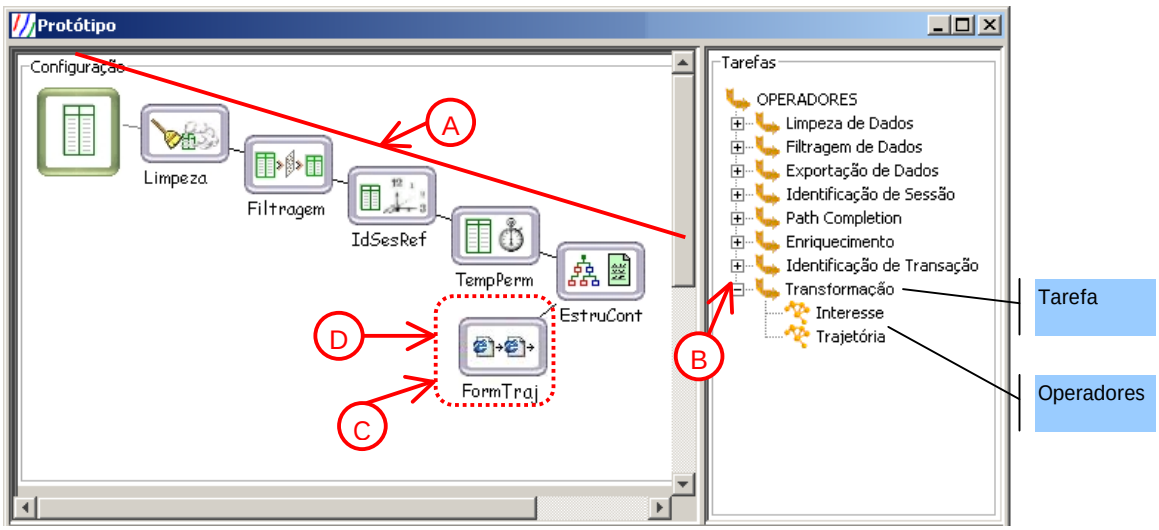

Figura 48 – Interface do módulo de preparação dos dados (LogPrep[MAR04b])

Para a tarefa de transformação das sessões foram criados dois operadores que permitem que o analista escolha entre o objetivo do agrupamento (interesse ou trajetória). A Figura 49 ilustra a interface gráfica do operador que implementa as tarefas de transformação das sessões necessárias ao agrupamento de interesse.

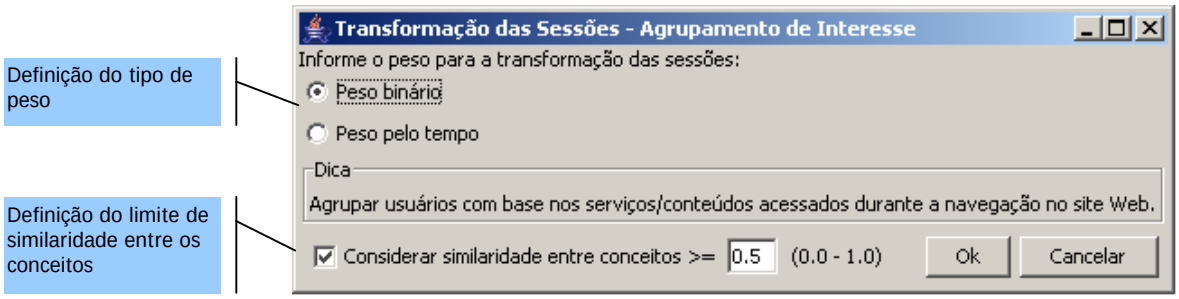

Figura 49 – Operador de transformação das sessões para o agrupamento de interesse

Este operador apresenta ao final do seu processamento o resultado da transformação das sessões, como ilustra a Figura 50. Neste ponto o operador permite que as demais atividades de transformação das sessões relacionadas à estatística do uso do agrupamento de interesse (remoção de conceitos e remoção de sessões) sejam executadas. Observa-se que as tarefas relacionadas à estatística de uso não encontram-se implementadas. O cálculo de similaridade entre as sessões e a geração da matriz de similaridade para cada nível conceitual de interesse é realizada através do botão “Exportar para agrupamento”. O botão “Exportar para agrupamento” abre uma janela de diálogo onde o analista informa o nome do arquivo XML que irá armazenar o conjunto de informações destinado ao protótipo ACSA.

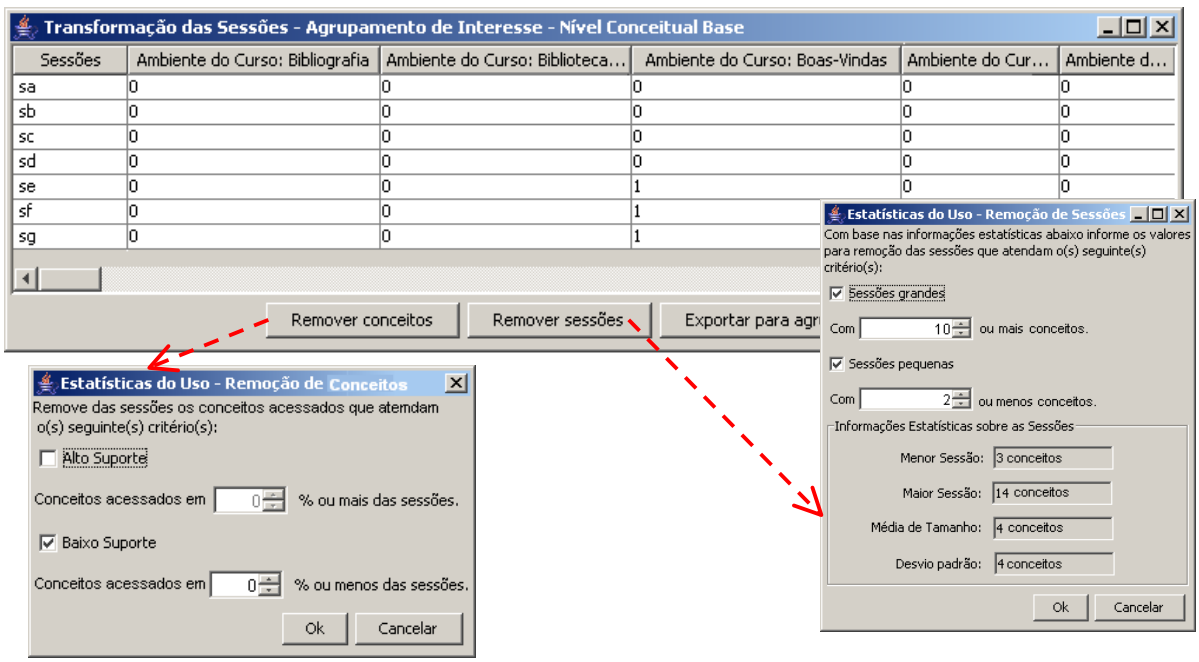

Figura 50 – Resultado da transformação das sessões para o agrupamento de interesse

A Figura 51 ilustra a interface gráfica do operador de trajetória que implementa as atividades de transformação das sessões necessárias ao agrupamento de trajetória.

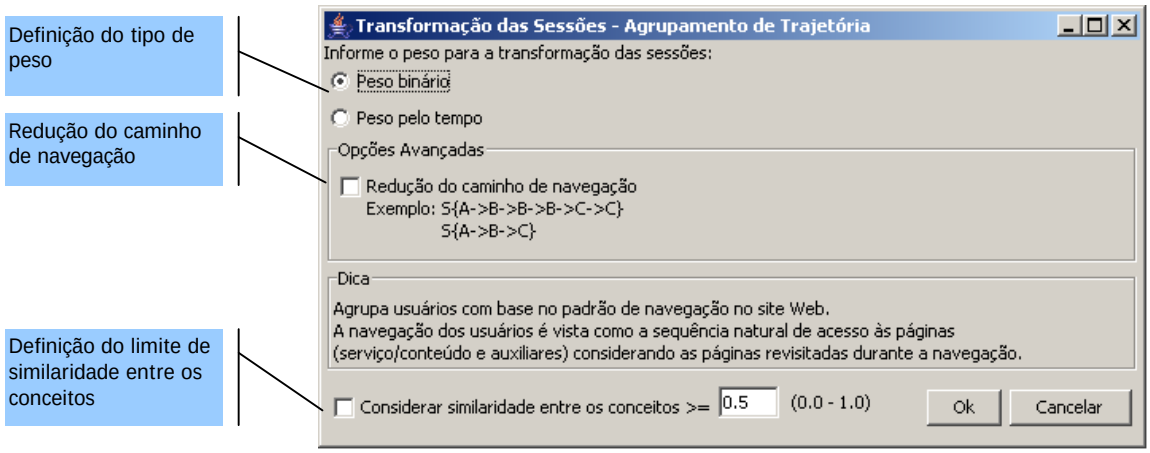

Figura 51 – Operador de transformação das sessões para o agrupamento de trajetória

Este operador apresenta ao final do seu processamento o resultado da transformação das sessões, ilustrado na Figura 52. Portanto, neste ponto o operador permite que as demais atividades de transformação das sessões relacionadas à estatística do uso do agrupamento de trajetória (remoção de sessões) sejam executadas e as matrizes de similaridade para cada nível conceitual de interesse sejam criadas.

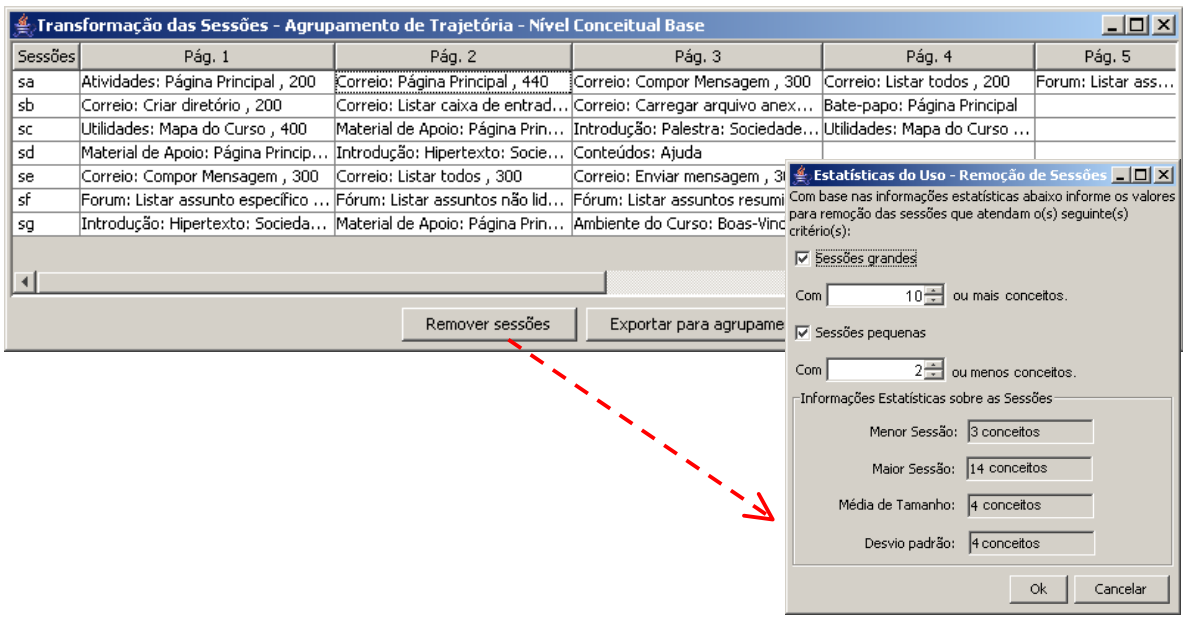

Figura 52 – Resultado da transformação das sessões para o agrupamento de trajetória

8.2.1.3 Detalhes da Implementação

A ferramenta LogPrep foi estendida para incluir a nova tarefa chamada de "Transformação" e seus operadores de interesse e trajetória da seguinte maneira:

- Componentes Java: foram criados dois componentes que implementam, respectivamente, os novos operadores de transformação das sessões para o agrupamento de trajetória e de interesse. A matriz de similaridade gerada, referente à cada nível conceitual de interesse, é representada por uma matriz densa e é armazenada em um arquivo texto;
- Atualização do arquivo de configuração: foi atualizado o arquivo de configuração de LogPrep (configOperadores.txt) para incluir a nova tarefa de transformação, os novos operadores, bem como as restrições de execução destes operadores.

8.2.2 ACSA

ACSA emprega o algoritmo baseado em grafo chamado “graph-partitioning” disponibilizado por SCLUSTER para realizar o agrupamento com base na matriz de similaridade calculada para cada nível conceitual de interesse. SCLUSTER é uma aplicação em linha de comando oferecida por CLUTO [CLU06] e foi escolhida por se tratar de uma aplicação simples de utilizar, que recebe como entrada um arquivo representando a matriz de similaridade e retorna como saída um outro arquivo contendo o resultado do agrupamento. ACSA utiliza a ferramenta WebPath [TRI04] como base para implementar a visualização da árvore agregada, estendendo a classe de visualização de WebPath para considerar o peso pelo tempo de acesso. WebPath foi escolhida por já implementar as funcionalidades de visualização da árvore agregada. Os demais funcionalidades foram implementadas em Java. A Figura 53 esquematiza as entradas e saídas de ACSA.

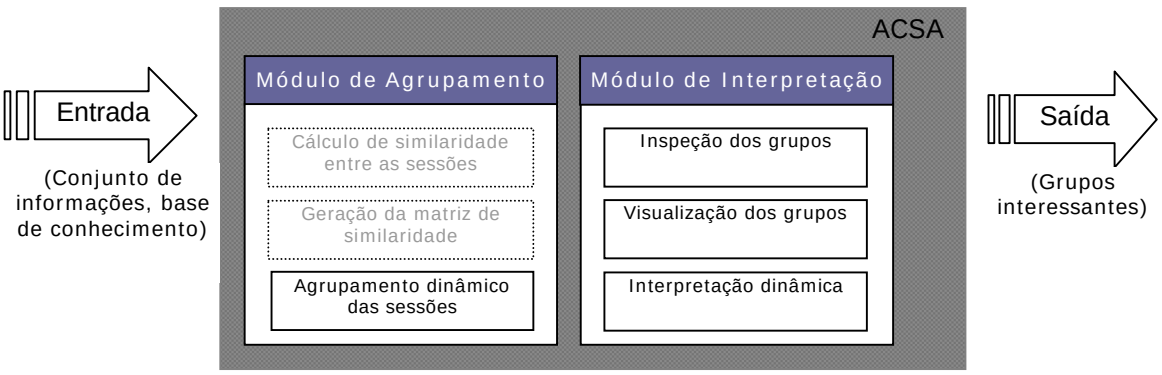

Figura 53 – Entradas e saídas de ACSA

8.2.2.1 Funcionalidades

A Tabela 8 descreve as funcionalidades específicas de ACSA em termos das etapas de descoberta de padrões e análise dos padrões.

Tabela 8 – Funcionalidades de ACSA

Funcionalidade	Descrição
(E) Importar conjunto de informações	O analista importa o conjunto de informações gerado na etapa de preparação dos dados.
(F) Definir parâmetros do agrupamento	O analista define os seguintes parâmetros de agrupamento: número de grupos e localização da ferramenta de agrupamento (SCLUSTER).
(G) Executar o agrupamento	O analista executa o agrupamento das sessões.
(H) Definir o grupo e o nível conceitual de interesse	O analista escolhe o grupo e o nível conceitual de interesse para visualizar.
(I) Visualizar o perfil agregado que caracteriza o grupo	O analista visualiza o perfil agregado que representa o grupo selecionado.
(J) Visualizar a árvore agregada que caracteriza o grupo	O analista visualiza a árvore agregada do grupo selecionado.
(L) Definir filtro de importância	O analista define o filtro de interesse para visualizar o perfil agregado.
(M) Definir ordem dos pesos no perfil agregado	O analista define a ordem em que os conceitos do perfil agregado serão mostrados de acordo com os pesos associados.
(N) Executar roll-up	O analista executa operação de roll-up na representação do perfil agregado ou na árvore agregada.
(O) Executar drill-down	O analista executa operação de drill-down na representação do perfil agregado ou na árvore agregada.

8.2.2.2 Módulo de Agrupamento

A Figura 54 ilustra a primeira aba de ACSA, chamada “Agrupamento”, que suporta as funcionalidades (E), (F) e (G) da Tabela 8. Estas funcionalidades estão disponibilizadas na interface de ACSA respectivamente pelas áreas de “Arquivo de entrada”, “Parâmetros do agrupamento”, e pelo botão “Executar agrupamento”, como ilustra a Figura 54.

8.2.2.2.1 Área de Arquivo de Entrada

Esta área, ilustrada pela Figura 54-E, permite a importação do arquivo que contém o conjunto de informações necessárias para realizar o agrupamento, produzido pelas adaptações em LogPrep, como discutido na seção 8.2.1. Com base nas informações deste arquivo, esta área apresenta informações importantes ao agrupamento, a saber, o objetivo do agrupamento, número de sessões identificadas, peso escolhido e o limite de similaridade entre os conceitos. O número de sessões é destinado a auxiliar o analista na definição do número de grupos. Já o objetivo do agrupamento, o peso e o máximo de abstração são utilizados por ACSA para montar corretamente a segunda aba referente ao Módulo de Interpretação.

8.2.2.2.2 Área de Parâmetros do Agrupamento

Esta área, ilustrada pela Figura 54-F, permite que o analista informe os parâmetros para executar o agrupamento (número de grupos e caminho da ferramenta SCLSUTER). Depois de importado o arquivo de entrada e definidos os parâmetros do agrupamento, o botão "Executar agrupamento" (Figura 54-G) é habilitado, permitindo que o mecanismo de agrupamento seja executado. O algoritmo "graph-partitioning" implementado por SCLUSTER tem como base as premissas dos algoritmos Metis e hMetis. SCLUSTER recebe como parâmetro de entrada uma matriz de similaridade e modela os dados recebidos utilizando um grafo do tipo nearest-neighbor, onde cada sessão é representada por um vértice e está conectada somente às sessões mais similares. SCLUSTER permite ainda especificar parâmetros para a construção do grafo (ex: número máximo e mínimo de vizinhos para cada vértice, tipo de ligações entre os vértices, valor limite para eliminação de ligações, etc). Este trabalho assume valores padrão para estes parâmetros, definidos pelo próprio SCLSUTER, de modo que o analista não precisa se preocupar com a definição destes. Após construir o grafo, SCLUSTER particiona o grafo usando o algoritmo min-cut considerando o número de grupos definido pelo analista. A aba "Interpretação" é habilitada ao final da execução do agrupamento.

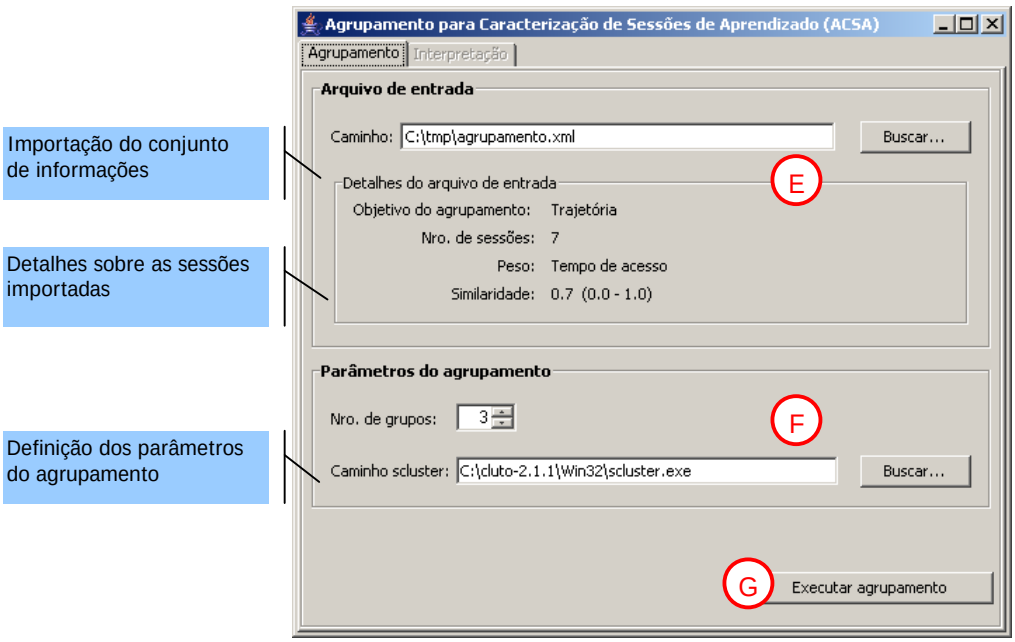

Figura 54 – Módulo de Agrupamento

8.2.2.3 Módulo de Interpretação

A segunda aba de ACSA, chamada de "Interpretação", suporta as funcionalidades (H), (I), (J), (L), (M), (N) e (O) da Tabela 8. Estas funcionalidades

estão disponibilizadas na interface de ACSA respectivamente pelas áreas de: “Inspeção do grupo”, “Perfil agregado” e “Árvore agregada”, como ilustra a Figura 55.

A Figura 55 ilustra a visualização do perfil agregado e da árvore agregada que representam o grupo 0 no nível conceitual de interesse 0, considerando o agrupamento de trajetória e o peso pelo tempo de acesso.

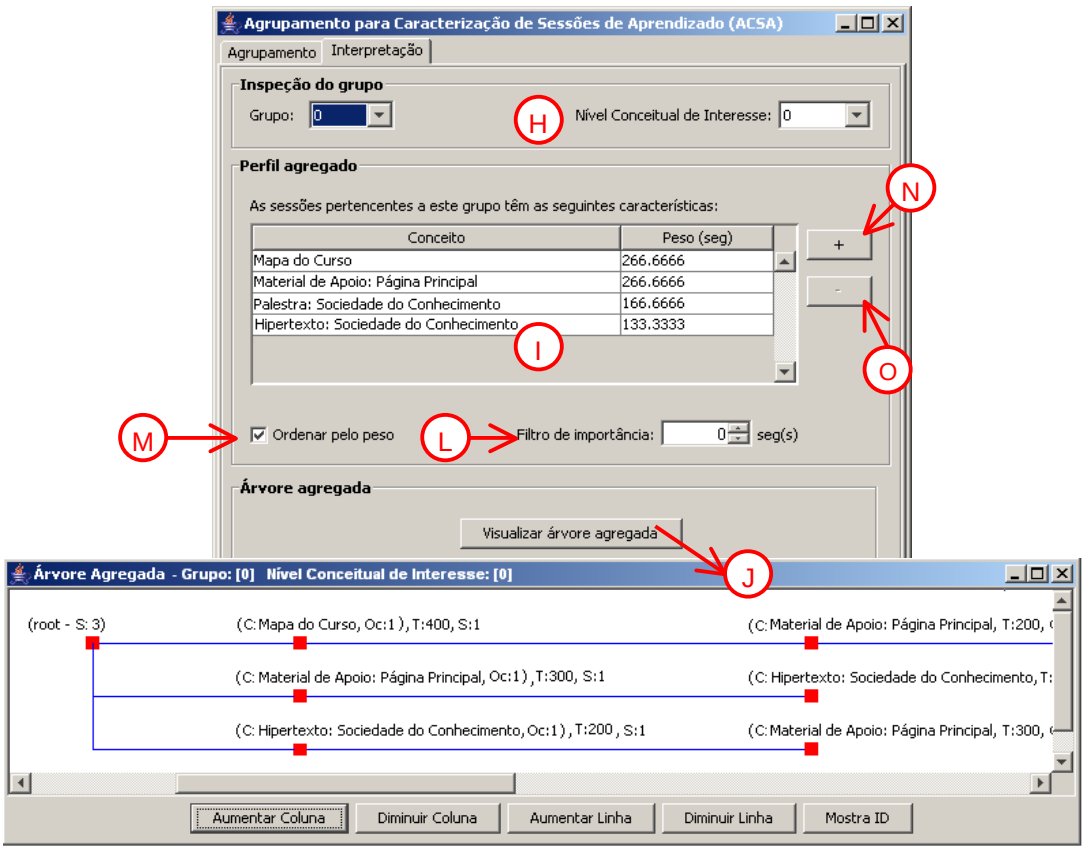

Figura 55 – Módulo de Interpretação

8.2.2.3.1 Área de Inspeção do Grupo

Esta área, ilustrada pela Figura 55-H, permite que o analista selecione o grupo e o nível conceitual de interesse a visualizar. É importante notar que cada vez que o seletor do grupo e/ou do nível conceitual de interesse são alterados, a visualização do grupo no perfil agregado (Figura 55-I) volta ao nível conceitual base. Ou seja, não importando em qual nível de abstração a operação de roll-up/drill-down estava, quando um dos seletores na área de “Inspeção do grupo” é modificado a visualização do grupo na área do “Perfil agregado” retorna à representação no nível de conceitual base.

8.2.2.3.2 Área de Perfil Agregado

Esta área, ilustrada pela Figura 55-I, permite visualizar o perfil agregado do grupo selecionado no nível conceitual de interesse selecionado, definir o filtro de importância que restringe os conceitos apresentados no perfil agregado, ordenar os conceitos no perfil agregado, bem como executar os operadores de roll-up e drill-down. No presente exemplo, a unidade apresentada nos conceitos no perfil agregado e no filtro de interesse é em segundos, pois o peso utilizado é o peso pelo tempo de acesso. Se o peso binário fosse escolhido como objetivo do agrupamento, então o sinal gráfico da percentagem ("%") seria apresentado ao lado do título da coluna do peso e do filtro de interesse. Caso seja interessante para o analista reduzir o número de conceitos apresentados no perfil agregado, basta alterar o valor especificado no filtro de importância (Figura 55-L). Caso o analista deseje ordenar os conceitos dentro do perfil agregado pelo peso basta marcar a opção "ordenar pelo peso" (Figura 55-M).

Os operadores de roll-up e drill-down no perfil agregado são disponibilizados respectivamente pelos botões "+" (Figura 55-N) e "-" (Figura 55-O). No presente exemplo, o perfil agregado do grupo 0 no nível conceitual de interesse 0 está abstraído para alguns níveis de generalização na hierarquia conceitual. A implementação dos operadores de roll-up e drill-down não foi finalizada para o perfil agregado com peso binário.

8.2.2.3.3 Visualização da Árvore Agregada

Esta área, ilustrada pela Figura 55-J, está disponível somente para o agrupamento de trajetória. O botão "Visualizar Árvore Agregada" permite visualizar em uma outra janela a árvore agregada referente ao grupo e nível conceitual de interesse selecionado. A implementação dos operadores de roll-up e drill-down não foi finalizada para a árvore agregada.

8.3 Considerações

O cálculo de similaridade entre as sessões e a geração da matriz de similaridade, originalmente destinadas ao Módulo de Agrupamento, foram implementadas no Módulo de Preparação dos Dados. Esta escolha afetou de forma negativa a facilidade esperada na aplicação do mecanismo de agrupamento proposto, pois a necessidade de obter agrupamentos mais significativos considerando a redefinição do valor mínimo de similaridade entre os conceitos ou a escolha de outro valor para o máximo de abstração das URLs, utilizado durante o enriquecimento dinâmico das sessões, implica no retorno à etapa inicial de preparação dos dados.

9 ESTUDO DE UM CASO EM UM AMBIENTE DE ENSINO A DISTÂNCIA

Este capítulo descreve o estudo de caso realizado no contexto da Educação a Distância para avaliar os mecanismos de agrupamento e interpretação de padrões propostos.

Este estudo de caso tem como objetivo avaliar como o ambiente de apoio à caracterização de sessões de aprendizado proposto auxilia o analista durante as fases de Descobrimento de Padrões e Análise de Padrões para a caracterização das sessões de aprendizado. Para isso, é descrito um cenário de uso do protótipo ACSA (Agrupamento para Caracterização de Sessões de Aprendizado) com dados reais obtidos do departamento de ensino a distância da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS). A unidade de EAD da PUCRS Virtual foi escolhida para a realização deste estudo de caso deste trabalho pela viabilidade de contactar os instrutores, uma vez que o presente trabalho faz parte da unidade de Informática dentro da PUCRS, e pela facilidade na obtenção dos arquivos de logs do curso Web.

Nas sessões seguintes, é descrito com maiores detalhes o ambiente de cursos de EAD da PUCRS Virtual, o processo de preparação dos dados do curso escolhido, bem como o estudo de caso realizado para a caracterização das sessões de aprendizado utilizando o protótipo ACSA.

9.1 Ambiente de Ensino da EAD da PUCRS Virtual

O ambiente de ensino da PUCRS Virtual oferece inúmeros cursos de EAD. Cada curso tem suas características próprias e necessita de uma infra-estrutura de conteúdos e serviços diferenciados. A ferramenta WebCT é utilizada para criar e gerenciar todos os cursos de EAD instalados na PUCRS Virtual.

WebCT é uma plataforma composta de um conjunto de ferramentas de suporte para a criação e manutenção de ambientes de ensino baseados na Web, bem como controle de acesso Web e autenticação dos alunos e instrutores. Além disso, o WebCT disponibiliza ferramantas educacionais tais como sistema de vídeo conferência, bate-papo, correio eletrônico, fórum, calendário do curso, glossário, acompanhamento do progresso do aluno, distribuição e controle de notas, geração automática de índices de pesquisa, etc. A escolha de quais ferramentas educacionais farão parte do curso é realizada pelo projetista do cruso no momento de criação do mesmo de acordo com as necessidades de cada curso.

O acesso ao ambiente de EAD na PUCRS Virtual é dado através de um servidor Web central onde o WebCT é instalado. Este servidor registra os acessos dos alunos e

instrutores e armazena estes dados em um arquivo de log. O formato dos arquivos de log gerados pelo WebCT é o CLF [W3C05], onde cada entrada indica qual a página ou serviço foi requisitado, quando e de onde partiu a requisição, o status de resposta do servidor Web, e pode ainda conter a identificação do usuário caso este tenha se autenticado.

A Figura 56 ilustra uma amostra do arquivo de log gerado pelo servidor Web da PUCRS Virtual para os cursos gerenciados pelo WebCT. Pode-se notar que as URLs acessadas no servidor Web onde o ambiente do WebCT está instalado apresenta em sua grande maioria chamadas de scripts com inúmeros parâmetros que tornam o evento representado pela URL acessada irreconhecível mesmo para um conhecedor do domínio. Ou seja, analisando cada entrada dentro do arquivo de log bruto puro é muito difícil reconhecer qual o evento do domínio associado. Isso é uma característica da organização dos conteúdos e serviços oferecidos pelo WebCT, que descreve por exemplo, cada evento de serviço por um script armazenado em um diretório /script/ dentro da estrutura física de cada curso, e utiliza diversos parâmetros em cada script para identificar usuários e ações específicas. Ou seja, no ambiente da PUCRS Virtual um mesmo evento do domínio, por exemplo, “Compor mensagem”, da ferramenta de correio eletrônico, apresenta uma URL específica no arquivo de acesso de acordo com o curso em questão, dado que a localização física do curso faz parte da URL acessada.

```
200.176.25.11 - aluno1 [2002/01/10 00:00:06.0 -0200] "GET /ESP_SE_01130/competencia/07_01/conselhos.doc" 200 2345
200.176.8.249 - -      [2002/01/10 00:10:17.0 -0200] "GET /" 200 1245
200.176.8.249 - -      [2002/01/10 00:10:18.0 -0200] "GET /webct/public/home.pl" 200 4566
200.176.8.249 - -      [2002/01/10 00:10:23.0 -0200] "GET /webct/homearea/homearea" 401 899
200.248.5.164 - aluno3 [2002/01/10 00:10:39.0 -0200] "GET /webct/homearea/homearea" 200 7677
200.176.8.249 - -      [2002/01/10 00:11:20.0 -0200] "GET /SCRIPT/CursoCCD/scripts/student/dropbox_stud_home.pl" 401 899
200.248.5.164 - -      [2002/01/10 00:11:21.0 -0200] "GET /webct/homearea/homearea" 401 899
200.176.8.249 - aluno2 [2002/01/10 00:11:25.0 -0200] "GET /SCRIPT/CursoCCD/scripts/student/dropbox_stud_home.pl" 200 5676
200.248.5.164 - aluno3 [2002/01/10 00:11:32.0 -0200] "GET /webct/homearea/homearea',200 7677
200.248.5.164 - -      [2002/01/10 00:11:58.0 -0200] "GET /SCRIPT/CursoCCD/scripts/serve_home" 401 899
200.248.5.164 - aluno3 [2002/01/10 00:12:02.0 -0200] "GET /SCRIPT/CursoCCD/scripts/serve_home" 200 5674
200.176.8.249 - aluno2 [2002/01/10 00:11:30.0 -0200] "GET /SCRIPT/CursoCCD/scripts/student/dropbox_stud_home.pl?START+++" 200 345
200.248.5.164 - aluno3 [2002/01/10 00:13:49.0 -0200] "GET /SCRIPT/CursoCCD/scripts/student/serve_home?1010412547+view" 200 344
```

Figura 56 – Amostra do arquivo de log do ambiente da PUCRS Vitual

9.2 Estudo de Caso

Foi escolhido um curso de capacitação docente da PUCRS Virtual como fonte de dados para este estudo de caso. O curso escolhido, doravante denominado simplesmente CursoCCD, foi ministrado em Janeiro do ano de 2002 com um total de 9 dias de aulas interativas na Web. Este trabalho utiliza os três primeiros dias do CursoCCD (dias 10, 11 e 12) para realizar o estudo de caso para caracterização de sessões de aprendizado. Os três dias de curso escolhidos juntos têm um total de 33.500 entradas e 27 diferentes alunos. Dentre os alunos que acessaram o CursoCCD

estavam, além dos alunos do CursoCCD, os administradores do sistema, os instrutores, alunos de outros cursos, bem como usuários anônimos. Além disso, outros cursos estavam sendo ministrados na mesma época, por esta razão nem todas as entradas dos arquivos de log correspondiam ao CursoCCD.

9.2.1 Preparação dos Dados

Os dados contidos nos arquivos de log dos três dias de curso escolhidos passaram pelo Módulo de Preparação de Dados implementado como uma extensão de LogPrep [MAR04b]. Assim, a ferramenta de pré-processamento LogPrep foi utilizada para realizar as seguintes tarefas relacionadas à etapa de preparação dos dados para o agrupamento de sessões:

- Limpeza dos Dados: foram removidas as entradas desnecessárias e irrelevantes com base na extensão dos arquivos acessados: figuras (.gif, .jpg, .bmp, .jpeg), estilos (.css, .js, .swf) e ícones (.ico). Além disso, foram mantidas somente as entradas referentes à páginas requisitadas e atendidas com sucesso pelo servidor Web, ou seja, as requisições com código de resposta (status) diferente de 200 foram removidas;
- Identificação do Usuário: os usuários foram identificados através do campo de autenticação do arquivo de log (Auth). Somente as entradas relativas aos usuários pertencentes ao CursoCCD foram mantidas. As demais entradas, relacionadas às visitas de usuários anônimos, alunos de outros cursos, administradores do sistema e instrutores foram removidas. Para manter a privacidade dos alunos do curso de capacitação docente da PUCRS Virtual, os nomes dos usuários do CursoCCD foram trocados por nomes fictícios;
- Identificação de Sessões: as sessões dos usuários foram quebradas pelo tempo máximo de visita entre às páginas de 30 minutos e pelo intervalo máximo de 60 minutos para acessos a páginas em cada sessão;
- Identificação de Visão de Páginas: URLs que contribuíram para a formação e visualização de uma página no navegador do usuário foram mapeadas por um mesmo conceito na hierarquia conceitual. Assim, quando aplicada a redução do caminho de navegação as URLs que compõem cada visão de página na sessão do usuário são reduzidas respectivamente à um conceito acessado;
- Enriquecimento: páginas com tempo de acesso de 1 segundo ou menos foram identificadas como auxiliares e cada URL acessada no arquivo de log foi mapeada pelo conceito correspondente na hierarquia do domínio do CursoCCD;

- Representação das Sessões: foi escolhido o objetivo de agrupamento de trajetória e o peso pelo tempo de acesso;
- Transformação das Sessões: o filtro de importância removeu os acessos às páginas auxiliares e a redução do caminho de navegação foi aplicada onde os tempos de acesso foram somados para as páginas contíguas unificadas nas sessões dos usuários.

As tarefas clássicas de complemento dos caminhos de navegação e identificação de transações não foram realizadas neste estudo de caso, pois LogPrep não implementa as mesmas. Além disso, as operações de normalização dos tempos de acesso e estatísticas do uso relacionadas à tarefa de transformação das sessões para o agrupamento não foram aplicadas, pois as mesmas ainda não estavam implementadas no Módulo de Preparação de Dados no período deste trabalho.

Ao final da etapa de pré-processamento dos dados obteve-se 4.130 entradas válidas, 17 alunos do CursoCCD e 165 sessões.

O limite de similaridade (m) entre os conceitos utilizado para determinar a subseqüência em comum entre duas sessões foi 0.6. O nível máximo de abstração escolhido foi 2 níveis abaixo da raiz da hierarquia conceitual pois, com base na semântica dos eventos da aplicação, era importante limitar a generalização dos conceitos desconsiderando os conceitos dos dois níveis do topo da hierarquia conceitual (“Eventos”, “Serviços” e “Conteúdos”). Assim, as 165 sessões identificadas foram traduzidas considerando 4 níveis conceituais de interesse. O número de níveis conceituais de interesse utilizado foi obtido pela altura da hierarquia conceitual (6 níveis) menos o nível máximo de abstração definido (2 níveis do topo).

9.2.2 Hierarquia Conceitual

A hierarquia do domínio que representa o CursoCCD foi criada manualmente e avaliada por um professor especialista do domínio da EAD da PUCRS Virtual. Vale ressaltar que o presente trabalho não tem por objetivo avaliar o método utilizado na construção da hierarquia conceitual, mas sim explorar a utilização desta como suporte aos mecanismos propostos.

Cada evento de domínio do CursoCCD foi acessado e o padrão da URL correspondente gerado foi mapeado para um conceito na hierarquia conceitual. Observou-se nos arquivos de logs coletados que diferentes URLs correspondiam a mesmos eventos do domínio. Por exemplo, as URLs ilustradas na Figura 57, com o padrão marcado em negrito, apesar de serem URLs diferentes, correspondem ao mesmo evento e foram mapeadas para o conceito “Sala de Entrega de Atividades”.

```
/SCRIPT/CursoCCD/scripts/student/dropbox_stud_home.pl
/SCRIPT/CursoCCD/scripts/student/dropbox_stud_home.pl?
/SCRIPT/CursoCCD/scripts/student/dropbox_stud_home.pl?START+++
```

Figura 57 – Exemplo de mapeamento de URLs para conceitos na hierarquia conceitual

Assim, a hierarquia do domínio foi sendo construída à medida que as URLs eram acessadas. Para finalizar, a hierarquia do domínio criada sofreu o último refinamento de acordo com o conhecimento do especialista do domínio da PUCRS Virtual. No total, a hierarquia conceitual criada é formada de 183 conceitos e uma altura de 6 níveis conceituais de interesse.

9.2.3 Protótipo ACSA: Cenário de Uso

Esta seção descreve um cenário de uso do protótipo ACSA que disponibiliza os mecanismos de agrupamento e interpretação de padrões utilizados pelo analista para identificar e melhor caracterizar as sessões de aprendizado. Este cenário de uso considera que as atividades de preparação dos dados, descritas na seção 9.2.1, já foram desenvolvidas anteriormente.

9.2.3.1 Iniciando o Protótipo ACSA

Primeiramente, o analista inicia o protótipo ACSA (duplo clique no arquivo ACSA.jar). O protótipo ACSA é carregado e sua tela principal é mostrada inicialmente com somente a aba “Agrupamento” habilitada.

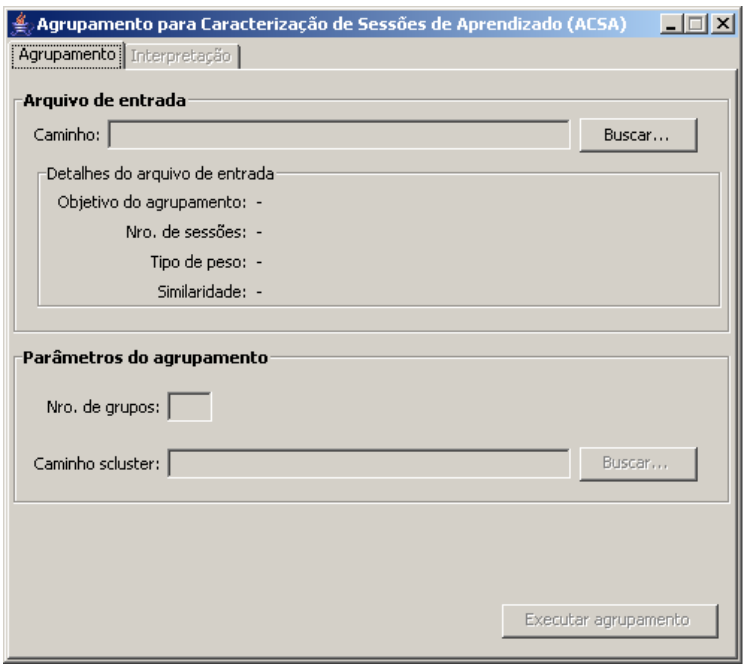

Figura 58 – Tela principal do protótipo ACSA

9.2.3.2 Importando os Dados

O analista interage com a área do arquivo de “Arquivo de entrada”, onde através do botão “Buscar...” informa o caminho do arquivo de entrada preparado pelo Módulo de Preparação de Dados. O botão “Buscar...” abre a janela padrão do Windows para busca de arquivos com extensão “.xml”, como ilustra a Figura 59.

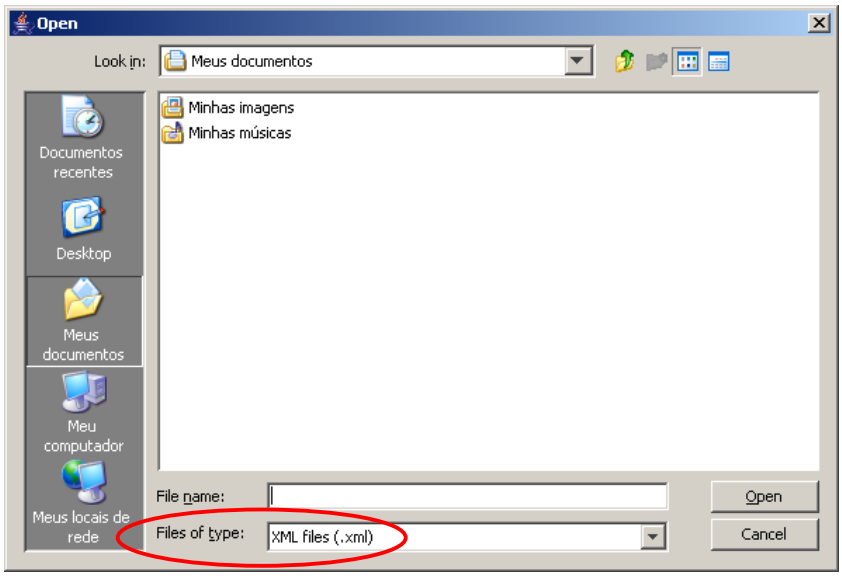

Figura 59 – Buscar arquivo de dados (.xml)

Importados os dados, os detalhes sobre as sessões pré-processadas são apresentados ao analista na área “Detalhes do arquivo de entrada”, conforme ilustra a Figura 60. O analista pode então importar novo arquivo de entrada (clicando no botão “Buscar...” novamente) ou começar o agrupamento das sessões importadas. Neste exemplo, o analista escolhe ir adiante com o agrupamento das sessões.

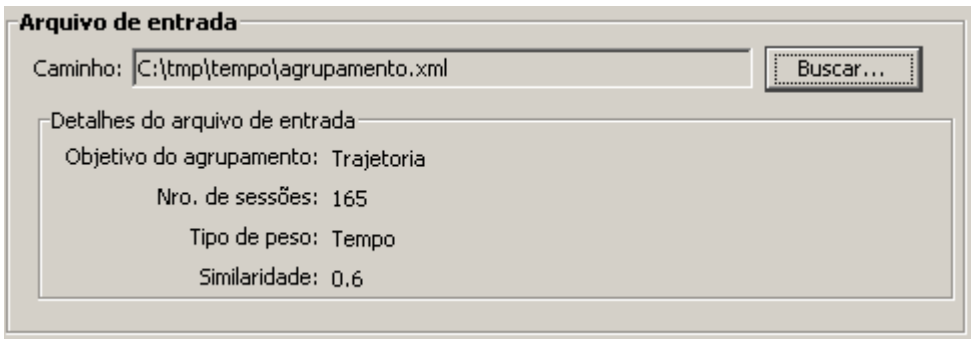

Figura 60 – Detalhes do arquivo de entrada importado

9.2.3.3 Realizando o Agrupamento das Sessões

Na área de “Parâmetros do agrupamento” (que implementa o Módulo de Agrupamento descrito na seção 8.2.2.2), o analista informa primeiramente o caminho

da ferramenta SCLUSTER através do botão “Buscar...”. Este botão abre a janela padrão do Windows para busca do arquivo “scluster.exe”, como ilustra a Figura 61.

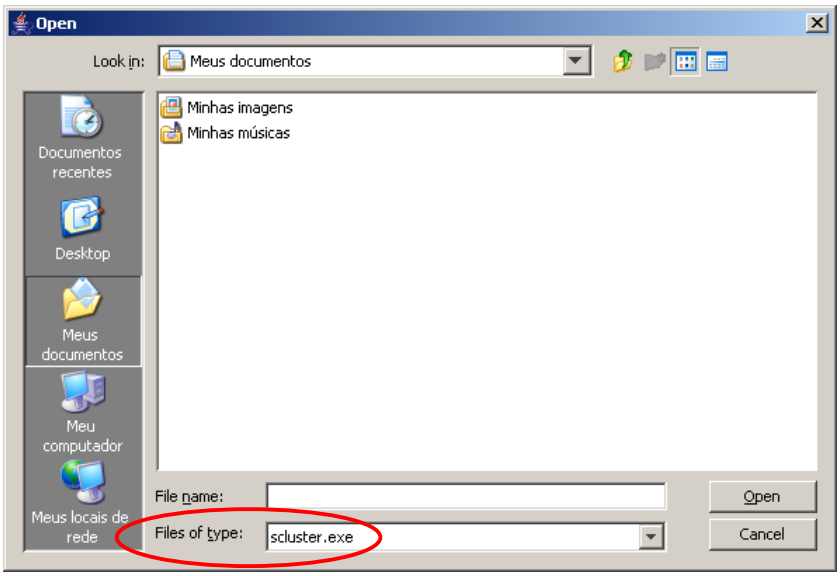

Figura 61 – Buscar arquivo scluster.exe

Após informar o caminho de SCLSUTER, o analista informa o número de grupos e clica em “Executar agrupamento” para realizar o agrupamento das sessões importadas, como ilustra a Figura 62.

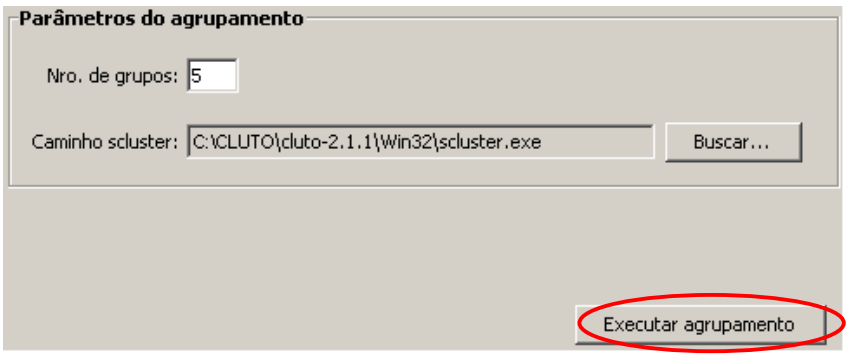

Figura 62 – Parâmetros para o agrupamento

Neste estudo de caso, foi escolhido agrupar as sessões em 5 grupos. Este número foi escolhido empiricamente embora existam métodos científicos para determinar tanto o número ideal de grupos [HEE02] quanto a qualidade dos grupos gerados [WAN02]. Pois, vale ressaltar que este trabalho não tem por objetivo avaliar método escolhido para definição do número de grupos.

O protótipo ACSA levou um total de 0.015 segundos para agrupar, considerando os 4 níveis conceituais de interesse, as 165 sessões em 5 grupos utilizando a ferramenta SCLUSTER, e 0.371 segundos para armazenar em memória as informações do agrupamento dinâmico realizado. Após realizar o agrupamento e

armazenar as informações dos grupos em memória, o protótipo ACSA então habilita a aba “Interpretação”.

9.2.3.4 Inspecionando e Interpretando os Grupos

O analista vai então para a aba “Interpretação” (que implementa o Módulo de Interpretação descrito na seção 8.2.2.3) onde é possível observar as características dos grupos formados, como ilustra a Figura 63.

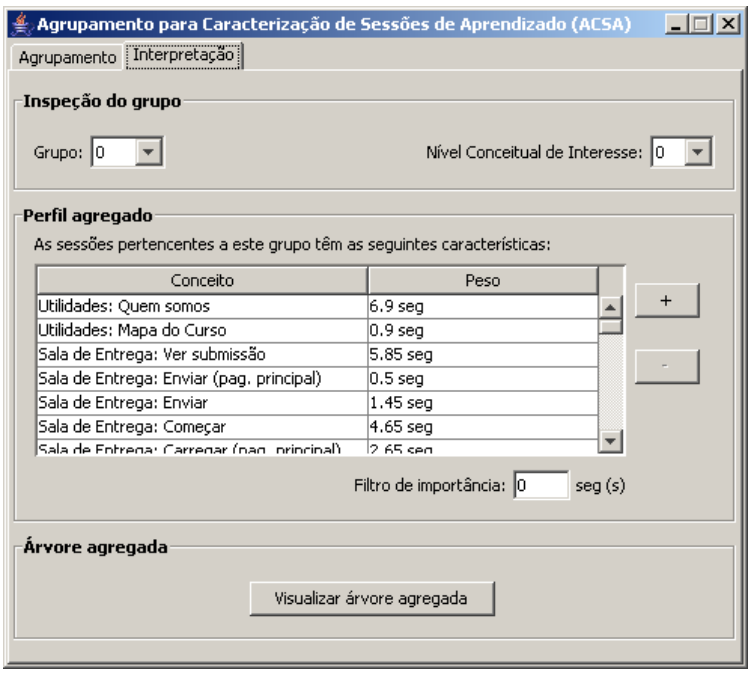

Figura 63 – Interpretação dos grupos

Inicialmente o analista explora o perfil agregado de cada grupo formado através da seleção do grupo e do nível conceitual na área de “Inspeção do grupo”. No exemplo ilustrado pela Figura 64-A, o analista visualiza o perfil agregado para o grupo 2 no nível conceitual 0 (nível conceitual base). O analista então informa o valor de 100 segundos para o filtro de importância e seleciona a ordem dos conceitos pelo peso. Então, o perfil agregado é automaticamente atualizado na área “Perfil agregado” para refletir o filtro e a ordenação estipulada (Figura 64-B). Como resultado, 40 conceitos que não atingiram o filtro de interesse foram descartados, permenacendo somente 17 conceitos para representar o perfil agregado do grupo 2.

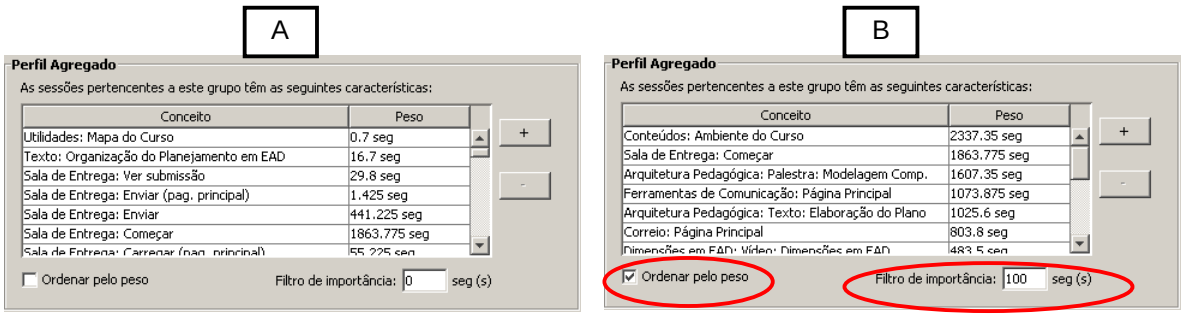

Figura 64 – Inspeção do perfil agregado

Como forma de aprofundar a caracterização do grupo 2, o analista clica no botão "Visualizar árvore agregada" e uma nova janela é aberta (Figura 65) com o caminho de navegação que representa as sessões pertencentes ao grupo 2. Neste exemplo, a árvore agregada representa a navegação das 40 sessões pertencentes ao grupo 2. Cabe ressaltar que a árvore gerada no exemplo em questão é de difícil interpretação, pois existem muitas ramificações e algumas sessões são muito extensas. Neste estudo de caso, o protótipo ACSA levou 0.016 segundos para exibir a árvore agregada independente do grupo e do nível conceitual de interesse selecionado.

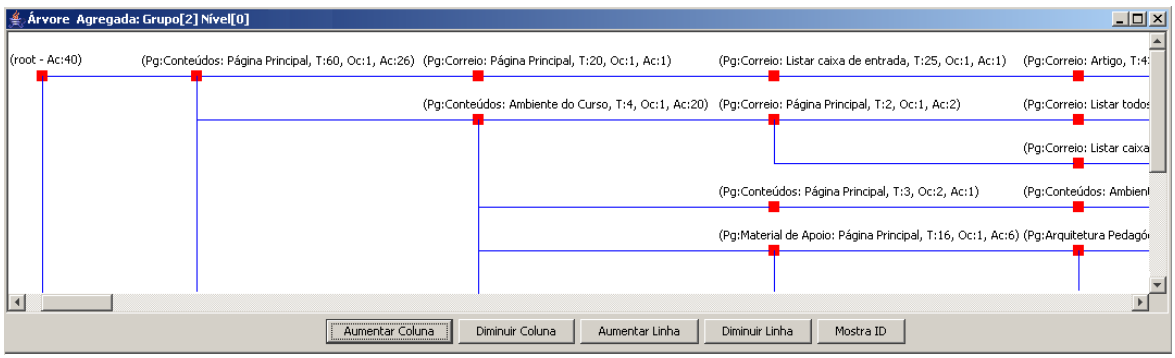

Figura 65 – Inspeção da árvore agregada

Mesmo com a aplicação do filtro de interesse, ordenação por peso, e visualização da árvore agregada, o analista continua com dificuldades em interpretar os grupo 2. Pois, neste nível conceitual de interesse, as URLs acessadas foram mapeadas para eventos mais especializados na hierarquia conceitual do domínio (ex: "Hipertexto: Sociedade do Conhecimento" e "Palestra: Dimensões em EAD" que são, respectivamente, conteúdos dos temas "Introdução" e "Dimensões em EAD").

Assim, com o objetivo de melhor compreender os conceitos que caracterizam as sessões pertencentes ao grupo 2 o analista clica no botão de roll-up (botão "+"), ilustrado na Figura 66-A, o qual realiza a interpretação dinâmica do grupo em questão para 1 nível acima de abstração. O analista entende que a operação de roll-up não desfaz a formação do grupo 2 selecionado, nem altera o nível conceitual em que as

sessões foram agrupadas originalmente. O analista descobre que os conceitos que caracterizam o grupo 2 são abstraídos para um nível acima de abstração a cada vez que o botão de roll-up (botão "+") é clicado. Neste exemplo, somente quando o analista realiza o roll-up do perfil agregado do grupo selecionado ele descobre que os eventos "Listar todas tarefas", "Listar todos assuntos" e "Ver texto" são recursos da ferramenta de fórum e que este, por sua vez, é uma ferramenta de comunicação. Para voltar ao perfil agregado original, o analista clica no botão de drill-dow (botão "-"), conforme ilustra a Figura 66-B.

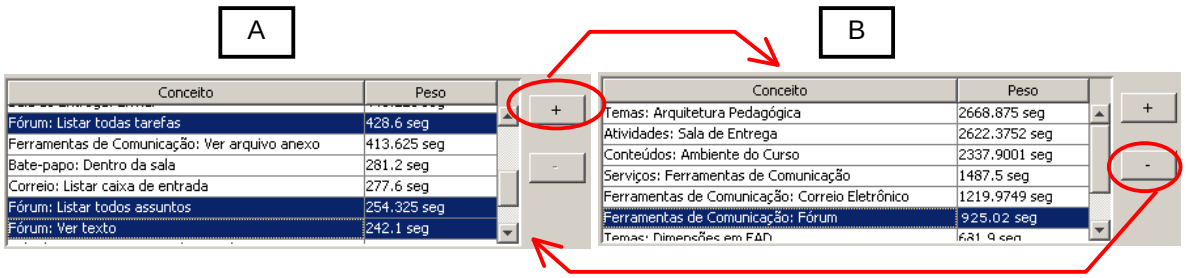

Figura 66 – Operação de roll-up e drill-down

Com base nas informações do perfil agregado foi possível determinar as características das sessões de aprendizado para cada um dos 5 grupos deste estudo de caso, conforme ilustra a Figura 67:

Grupo 0	
Características:	Ambiente do Curso
Filtro de interesse:	1000 segundos

Grupo 3	
Características:	Ambiente do Curso Informações Gerais do Curso
Filtro de interesse:	500 segundos

Grupo 1	
Características:	Sala de Entrega: Enviar Ambiente do Curso Descrição das Atividades
Filtro de interesse:	1000 segundos

Grupo 4	
Características:	Ambiente do Curso Sala de Entrega: enviar
Filtro de interesse:	1000 segundos

Grupo 2	
Características:	Arquitetura Pedagógica Ferram. de Comunicação Sala de Entrega: Começar Ambiente do Curso
Filtro de interesse:	1000 segundos

Figura 67 – Características das sessões de aprendizado

Pode-se também verificar como ficariam agrupadas as 165 sessões caso as URLs tivessem sido mapeadas para um nível conceitual de interesse mais abstrato. O

analista então seleciona o nível conceitual de interesse 1 na área de “Inspeção do grupo” (Figura 68) e o perfil agregado é atualizado para refletir a mudança do agrupamento. É importante notar que nos trabalhos de agrupamento existentes o enriquecimento das sessões para outra dimensão de interesse requer a volta à etapa inicial da MUW.

Figura 68 – Mudança do nível conceitual de interesse

O analista explora então os grupos formados no nível conceitual de interesse 1, selecionando os diferentes grupos na área de “Inspeção do grupo”, e visualizando o perfil agregado e a árvore agregado dos mesmos.

Neste ponto da interpretação, começam a surgir hipóteses sobre a formação dos grupos e a possibilidade de caracterização das sessões de aprendizado considerando um número diferente de grupos. O analista retorna a aba “Agrupamento” e informa, por exemplo, o valor 10 como número de grupos e clica no botão “Executar agrupamento”. Após o novo agrupamento das sessões ser executado, o analista volta à aba “Interpretação” para analisar os 10 grupos formados. O analista utiliza os mesmos recursos disponibilizados pelo protótipo ACSA para caracterizar os novos grupos de sessões de aprendizado.

10 CONCLUSÕES

O presente trabalho apresenta mecanismos para o agrupamento e interpretação de sessões de aprendizado no domínio da EAD. Os mecanismos propostos têm como objetivo facilitar a aplicação da técnica de agrupamento e a análise dos grupos por pessoas leigas, visando auxiliar na caracterização das sessões de aprendizado em um ambiente de EAD. Estes mecanismos fazem uso de uma taxonomia como forma de agregar semântica aos eventos do domínio, reduzindo assim a necessidade de retorno à etapa de pré-processamento.

O mecanismo de agrupamento proposto contribui para a aplicação da técnica de agrupamento, por: a) facilitar a preparação das sessões de acordo com o objetivo do agrupamento, b) não exigir do analista um conhecimento detalhado dos eventos domínio, uma vez que estes estão representados em uma taxonomia construída por um especialista do domínio, c) considerar a similaridade entre os conceitos durante o cálculo de similaridade entre as sessões, d) permitir lidar de forma homogênea tanto com o agrupamento de interesse quanto com o agrupamento de trajetória, e e) realizar o agrupamento dinâmico das sessões considerando todos os possíveis níveis conceituais de interesse que as sessões podem ser traduzidas. Uma análise comparativa do mecanismo de agrupamento em relação à técnica WLCS foi apresentada na seção 6.4.

O mecanismo de interpretação proposto contribui para facilitar a caracterização das sessões de aprendizado, por: a) representar os grupos de maneira condizente com os objetivos da mineração, b) permitir a inspeção dos grupos de acordo com o nível conceitual de interesse em que as sessões foram agrupadas, e c) permitir a interpretação dos grupos em termos da abstração das sessões que compõem o grupo.

Foi definido um ambiente de apoio à execução das fases da MUW que incorporou os mecanismos propostos, para o qual foi desenvolvido um protótipo. A construção deste protótipo foi de suma importância para a avaliação da viabilidade e correta execução dos mecanismos propostos. Um estudo de caso utilizando este protótipo e uma quantidade significativa de sessões obtidas de um curso de EAD da PUCRS Virtual foi apresentado no capítulo 9.

Observou-se que a implementação do cálculo de similaridade entre as sessões e geração da matriz de similaridade no Módulo de Preparação dos Dados afetou de modo negativo a facilidade esperada na aplicação do mecanismo de agrupamento proposto. Pois, a mudança no valor de similaridade para tentar gerar agrupamentos mais significativos implica no retorno à etapa de pré-processamento.

Dentre as principais limitações deste trabalho estão: a) a falta de uma análise mais profunda sobre a qualidade dos agrupamentos obtidos de acordo com o objetivo da mineração, e b) a falta de uma análise sobre as facilidades da aplicação dos mecanismos propostos por pessoas leigas. Outras limitações referem-se: à definição do limite de similaridade entre os conceitos por não considerar a relatividade deste valor em relação à profundidade dos conceitos na hierarquia, e à definição do número ideal de grupos que não é considerado por este trabalho.

A abordagem proposta assume que um especialista do domínio é necessário para: a) construir a taxonomia que representa os eventos do domínio, b) definir o mapeamento as URLs para o nível conceitual base, e c) definir o máximo de abstração dos conceitos na hierarquia.

Trabalhos futuros concentram-se em:

- análise mais profunda sobre a qualidade dos agrupamentos obtidos pelo mecanismo de agrupamento;
- validação do mecanismo de agrupamento proposto para o objetivo de agrupamento de interesse em relação às técnicas existentes;
- avaliar aplicabilidade dos mecanismos propostos em outros domínios de aplicação;
- avaliar a possibilidade de utilizar um limite de similaridade diferente em cada nível conceitual de interesse para a qual as sessões são traduzidas durante o agrupamento dinâmico;
- considerar a utilização de outras medidas de similaridade entre os conceitos;
- considerar outro algoritmo LCS para obter subseqüências mais significativas;
- definição do número ideal de grupos;
- integração com a Web Semântica.

REFERÊNCIAS

[ADR96] ADRIANS, P.; ZANTINGE, D. **Data Mining**. Inglaterra: Addison-Wesley, 1996.

[BAN00] BANERJEE A.; GHOSH, J. Concept-based Clustering of Clickstream Data. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY, 3., 2000. **Proceedings…** [S.l.: s.n.], 2000. p. 145-150.

[BAN01] BANERJEE, A.; GHOSH, J. Clickstream Clustering Using Weighted Longest Common Subsequences. In: WEB MINING WORKSHOP AT THE SIAM INTERNATIONAL CONFERENCE ON DATA MINING (SDM), 1., 2001. **Proceedings…** Chicago, USA: [s.n.], 2001.

[BEC03] BECKER, K.; VANZIN, M. Discovering Interesting Usage Patterns in Web-based Learning Environments. In: INTERNATIONAL WORKSHOP ON UTILITY, USABILITY AND COMPLEXITY OF INFORMATION SYSTEMS, 2003. **Proceedings…** Namur, Italy: [s.n.], 2003.

[BER00] BERENDT, B.; SPILIOPOULOU, M. Analysing navigation behaviour in Web sites integrating multiple information systems. **The VLDB Journal**, [S.l.:s.n.], v. 9, p. 56-75, 2000.

[BER01] BERENDT, B; MOBASHER, B.; SPILIOPOULOUS, M.; Wiltshire, J. Measuring the Accuracy of Sessionizers for Web usage analysis. In: WEB MINING WORKSHOP AT THE SIAM INTERNATIONAL CONFERENCE ON DATA MINING (SDM), 1., 2001. **Proceedings…** Chicago, USA: [s.n.], 2001.

[BER02] BERENDT, B; HOTHO, A.; STUMME, G. The semantic Web. In: International Semantic Web Conference (ISWC), 1., 2002. **Lecture Notes in Computer Science**. Heidelberg, German: Springer, 2002. v. 2342, p. 264-278.

[CLU06] CLUTO: CLUstering TOolkit. Disponível em: <http://www-users.cs.umn.edu/~karypis/cluto/>. Acesso em: 30 mar. 2006.

[COO00] COOLEY, R. **Web Usage Mining: Discovering and Application of Interesting Patterns from Web Data**. 2000. 181 f. Dissertation (Ph.d in Computer Science)-Department of Computer Science, University of Minnesota, Minneapolis, Minnesota, 2000.

[COO03] COOLEY, R. The Use of Web Structure and Content to Identify Subjectively Interesting Web Usage Patterns. In: ACM TRANSACTIONS ON INTERNET TECHNOLOGY (TOIT), 2003. New York, USA: ACM Press, 2003. v. 3, p. 93-116.

[COO97] COOLEY, R.; MOBASHER, B.; SRIVASTAVA, J. Web Mining: Information and Pattern Discovery on the World Wide Web. In: IEEE INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE (ICTAI), 9., 1997. **Proceedings…** Newport Beach, USA: [s.n.], 1997. p. 558-567.

[COO99] COOLEY, R.; MOBASHER, B.; SRIVASTAVA, J,. Data Preparation for Mining Word Wide Web Browsing Patterns. In: JOURNAL OF KNOWLEDGE AND INFORMATION SYSTEMS, 1999. [S.l.; s.n.], 2003. v. 1, p. 5-32.

[COR90] CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L. **Introduction to Algorithms**. [S.l.]: MIT Press, 1990.

[DAI02] Dai, H.; Mobasher, B. Using Ontologies to Discover Domain-Level Web Usage Profiles. In: SEMANTIC WEB MINING WORKSHOP AT 6TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES (PKDD), 2., 2002. Helsinki, Finland: [s.n.], 2002. p. 61-82.

[FAY96] FAYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMITH, P.; UTHURUSAMY, R. Advances in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI/MIT Press, 1996.

[FU00] FU, Y.; SANDHU, K.; SHIH, M. A Generalization-Based Approach to Clustering of Web Usage Sessions. In: WEB USAGE ANALYSIS AND USER PROFILING, 2000. **Lecture Notes in Artificial Intelligence**. Heidelberg, German: Springer, 2000. v. 1836, p. 21-38.

[GAN03] GANESAN, P.; GARCIA-MOLINA, H.; WIDOW, J.: Exploiting Hierarchical Domain Structure to Compute Similarity. In: ACM TRANSACTIONS ON INFORMATION SYSTEMS (TOIS), 2003. [S.l.: s.n], 2003. v. 21, n. 1, p.64-93.

[GUN03] GÜNDÜZ, S.; ÖZSU, M. T.: A Web Page Prediction Model Based on Click-stream Tree Representation of User Behavior. In: WEB MINING WORKSHOP AT THE ACM-SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY IN DATABASES (KDD), 2003. **Proceedings…** [S.l.: s.n.], 2003. p. 535-540.

[HAN00] HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. [S.l.]: Morgan Kaufmann Publishers, 2000.

[HEE01] HEER, J.; CHI, E. H. Identification of Web User Traffic Composition using Multi-Modal Clustering and Information Scent. In: WEB MINING WORKSHOP AT SIAM CONFERENCE ON DATA MINING, 2001. **Proceedings…** Chicago, EUA: SIAM Press, 2001. p. 51-58.

[HEE02] HEER, J.; CHI, E. H. Mining the Structure of User Activity using Cluster Stability. In: WORKSHOP ON WEB ANALYTICS AT SIAM CONFERENCE ON DATA MINING, 2002. **Proceedings…** Arlington, EUA:[s.n.], 2002. Não paginado.

[HIS77] HISCHBERG, D. S. Algorithms for the Longest Common Subsequences Problem. [S.l.]: J. ACM, 1977. v. 24, p. 664-675.

[JAI99] JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data Clustering: A Review. In: ACM COMPUTING SURVEYS, 1999. [S.l.: s.n.], 1999. v. 31, n. 3, p .264-323.

[MAC03a] MACHADO, L. **Mineração do Uso de Dados Aplicada à Educação a Distância: Propostas para a condução de um processo a partir de um estudo de caso**. 2003. Dissertação (Mestrado em Ciência da Computação)-Faculdade de Informática, PUCRS, Porto Alegre, RS, 2003.

[MAC03b] MACHADO, L.; BECKER, K. Distance Education: A Web Usage Mining Case Study for the Evaluation of Learning Sites. In: IEEE INTERNATIONAL CONFERENCE ON ADVANCED LEARNING TECHNOLOGIES (ICALT), 2003. Athens, Greece: IEEE Computer Society, 2003. p. 360-361.

[MAR04a] MARQUARDT, C. G. **Apoio ao Pré-Processamento de Dados da Mineração do Uso em Ambientes de Ensino Web**. 2004. Dissertação (Mestrado em Ciência da Computação)-Faculdade de Informática, PUCRS, Porto Alegre, RS, 2004.

[MAR04b] MARQUARDT, C. G.; BECKER, K.; RUIZ, D. A Pre-Processing Tool for Web Usage Mining in the Distance Education Domain. In: INTERNATIONAL SYMPOSIUM ON DATABASE ENGINEERING APPLICATIONS (IDEAS), 2004. **Proceedings…** Coimbra, Portugual: [s.n.], 2004. p. 78-87.

[MOB00a] MOBASHER, B.; COOLEY, R.; SRIVASTAVA, J. Automatic Personalization Based On Web Usage Mining. In: COMMUNICATION OF ACM, 2000. [S.l.: s.n.], 2000. v. 43, n. 8, p. 142-151.

[MOB00b] MOBASHER, B.; DAI, H.; LUO, T.; SUNG, Y.; ZHU, J. Integrating Web Usage and Content Mining for More Effective Personalization. In: INTERNATIONAL CONFERENCE ON E-COMMERCE AND WEB TECHNOLOGIES (ECWeb), 1., 2000. **Proceedings…** Greenwich, UK: [s.n.], 2000. p. 165-176.

[MOB00c] MOBASHER, B.; DAI, H.; LUO, T.; SUNG, Y.; NAKAGAWA, M.; WILTSIRE, J. Discovery of Aggregate Usage Profiles for Web Personalization. In: WEB MINING FOR E-COMMERCE WORKSHOP (WebKDD) AT THE CONFERENCE ON KNOWLEDGE DISCOVERY IN DATABASES (KDD), 2000. **Proceedings…** Boston, EUA: [s.n.], 2000. Não paginado.

[MOB01] MOBASHER, B.; DAI, H.; LUO, T.; NAKAGAWA, M. Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data. In: WORKSHOP ON INTELLIGENT TECHNIQUES FOR WEB PERSONALIZATION (ITWP) AT INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE (IJCAI). **Proceedings…** Seattle, EUA: [s.n.], 2001. p. 53-60.

[MOB02] MOBASHER, B.; DAI, H.; LUO, T.; NAKAGAWA, M. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. In: DATA MINING AND KNOWLEDGE DISCOVERY, 2002. [S.l.: s.n.], 2002. v. 6, n. 1, p. 61-82.

[MOB04] MOBASHER, B. Web Usage Mining and Personalization. In: Chapman Hall & CRC Press. **Practical Handbook of Internet Computing**. [S.l.]: Munindar P. Singh (ed.) and CRC Press, 2004. Não paginado.

[NIC04a] NICHELE, C. **Estudo das Técnicas de Segmentação de Sessões Web**. 2004. 57 f. Trabalho Individual parte I (apresentado como requisito parcial para obtenção do grau de mestre em ciência da computação)-Faculdade de Informática, PUCRS, Porto Alegre, RS, 2004.

[NIC04b] NICHELE, C. **Estudo das Técnicas de Segmentação de Sessões Web**. 2004. 93 f. Trabalho Individual parte II (apresentado como requisito parcial para obtenção do grau de mestre em ciência da computação)-Faculdade de Informática, PUCRS, Porto Alegre, RS, 2004.

[NIC06] NICHELE, C.; BECKER, K: Clustering Web Sessions by Levels of Page Similarity. In: PACIFIC ASIA CONFERENCE IN KNOWLEDGE DISCOVERY AND DATA MINING (PAKDD), 2006. Singapore, Singapore: [s.n.], 2006. p. 346-350.

[OBE03] OBERLE, D. et al.. Conceptual User Tracking. In: INTERNATIONAL ATLANTIC WEB INTELLIGENCE CONFERENCE, 2003. Madrid, Spain: Springer, 2003. p. 142-154.

[PIR99] PIROLLI, P; PITKOW, J. E. Distributions of Surfers' Paths Through the World Wide Web: Empirical Characterization. In: WORLD WIDE WEB, 1999. [S.l.: s.n.], 1999. v. 2, n. 1-2, p. 29-45.

[PUC06] PUCRS VIRTUAL. Unidade de Educação a Distância da Pontifícia Universidade Católica do Rio Grande do Sul. Disponível em: http://www.ead.pucrs.br. Acessado em: 20 jul. 2006.

[SPI99] SPILIOPOULOU, M; FAULSTICH, L. C.; WINKLER, K. A Data Miner analyzing the Navigational Behavior of Web Users. In: WORKSHOP ON MACHINE LEARNING IN USER MODELING AT INTERNATIONAL CONFERENCE (ACAI), 1999. **Proceedings…** Creta, Greece: [s.n.], 1999. Não paginado.

[SRI00] SRIVASTAVA, J. et al. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In: ACM SIGKDD Explor. Newsl., 2000. [S.l.: s.n.]: 2000. v. 1, n. 2, p. 12-23.

[STU02] STUMME, G.; BERENDT, B.; HOTHO, A. Usage Mining for and on the Semantic Web. In: NATIONAL SCIENCE FOUNDATION WORKSHOP ON NEXT GENERATION DATA MINING, 2002. **Proceedings...** Baltimore, USA: [s.n.], 2002. p. 77-86.

[TAN06] TAN, P. **Introduction to data mining**. Boston, EUA: Addison-Wesley, 2006.

[TRI04] TRISTÃO, C.; PEREIRA, V.; BECKER, K. WebPath: Uma ferramenta para a Mineração e Visualização de Padrões de Navegação do Uso da Web. In: SIMPÓSIO DA SBC EM SISTEMAS DE INFORMAÇÃO, 1., 2004. Porto Alegre, Brasil: [s.n], 2004. p. 103-110.

[VAN04a] VANZIN, M. **Mecanismos de Apoio a Interpretação e Recuperação de Padrões do uso da Web Baseados em Ontologia de Domínio**. 2004. 134 f. Dissertação (Mestrado em Ciência da Computação)-Faculdade de Informática, PUCRS, Porto Alegre, RS, 2004.

[VAN04b] VANZIN, M.; BECKER, K. Exploiting Knowledge Representation for Pattern Interpretation. In: WORKSHOP ON KNOWLEDGE DISCOVERY AND ONTOLOGIES, 2004. **Proceedings…** Pisa, Italy: [s.n.], 2004. Não paginado.

[VAN05] VANZIN, M.; BECKER, K.; RUIZ, D. Ontology-Based Filtering Mechanisms for Web Usage Patterns Retrieval. In: E-COMMERCE AND WEB TECHNOLOGIES (EC-Web), 2005. [S.l.: s.n.], 2005. p. 267-277.

[WAN02] WANG, W.; ZAIANE, O. Z. Clustering Web Sessions by Sequence Alignment. In: INTERNATIONAL WORKSHOP ON DATABASES AND EXPERT SYSTEMS APPLICATIONS (DEXA), 2002. [S.l.: s.n.], 2002. p. 394-398.

[W3C05] World Wide Web Consortium (W3C), Web Characterization Activity. Disponível em: http://www.w3.org. Accessado em: 30 mar. 2005.

[ZAN96] ZANG, T.; RAMAKRISHNAN, R.; LIVNY, M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 1996. **Proceedings…** Montreal, Canada: [s.n.], 1996. p. 103-114.

ANEXO A – ARQUIVO XML SCHEMA

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">

<xs:element name="id" type="xs:integer"/>
<xs:simpleType name="string-agrupamento">
  <xs:restriction base="xs:string">
    <xs:enumeration value="Trajetoria"/>
    <xs:enumeration value="Interesse"/>
  </xs:restriction>
</xs:simpleType>

<xs:simpleType name="string-peso">
  <xs:restriction base="xs:string">
    <xs:enumeration value="Binario"/>
    <xs:enumeration value="Tempo"/>
  </xs:restriction>
</xs:simpleType>

<xs:element name="matriz" type="xs:string"/>
<xs:element name="rclass" type="xs:string"/>
<xs:element name="clabel" type="xs:string"/>
<xs:element name="sessoes_PA" type="xs:string"/>
<xs:element name="sessoes_AA" type="xs:string"/>
<xs:element name="agrupamento" type="string-agrupamento"/>
<xs:element name="tipo_peso" type="xs:string-peso"/>
<xs:element name="num_sessoes" type="xs:integer"/>
<xs:element name="similaridade" type="xs:float"/>
<xs:element name="max_abstracao" type="xs:integer"/>

<xs:element name="nivel">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="matriz" use="required"/>
      <xs:element ref="rclass" use="required"/>
      <xs:element ref="clabel" use="required"/>
      <xs:element ref="sessoes" use="required"/>
      <xs:element ref="sessoes_PA"/>
      <xs:element ref="sessoes_AA"/>
    </xs:sequence>
    <xs:attribute ref="nci" use="required"/>
  </xs:complexType>
</xs:element>

<xs:element name="dataset">
  <xs:complexType>
    <xs:sequence>
    <xs:element ref="nivel" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute ref="agrupamento" use="required"/>
    <xs:attribute ref="peso" use="required"/>
    <xs:attribute ref="num_sessoes" use="required"/>
    <xs:attribute ref="similaridade" use="required"/>
    <xs:attribute ref="max_abstracao" use="required"/>
  </xs:complexType>
</xs:element>
</xs:schema>
```