Geraldo Ramos Falci Júnior

Metodologia de Mineração de Dados para Ambientes Educacionais Online

Dissertação de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: Engenharia de Computação.

Orientador: Ivan Luiz Marques Ricarte

Campinas, SP
2010

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

F181m

Falci Júnior, Geraldo Ramos
Metodologia de mineração de dados para ambientes educacionais online / Geraldo Ramos Falci Júnior. -- Campinas, SP: [s.n.], 2010.

Orientador: Ivan Luiz Marques Ricarte.
Dissertação de Mestrado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Mineração de dados (Computação). 2. Ambiente educacional. 3. Educação à distância. I. Ricarte, Ivan Luiz Marques. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: Data mining methodology for online educational environments
Palavras-chave em Inglês: Data mining (Computer), Educational environment, Distance education
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora: Juan Manuel Adán Coello, Romis Ribeiro de Faissol Attux
Data da defesa: 21/12/2010
Programa de Pós Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: Geraldo Ramos Falci Júnior

Data da Defesa: 21 de dezembro de 2010

Título da Tese: "Metodologia de Mineração de Dados para Ambientes Educacionais Online"

Prof. Dr. Ivan Luiz Marques Ricarte (Presidente): _____

Prof. Dr. Juan Manuel Adan Coello: _____

Prof. Dr. Romis Ribeiro de Faissol Attux: _____

Resumo

Educação a distância populariza-se como meio prático de ensino com a expansão de recursos computacionais e da Internet. Apesar disto, ela traz dificuldades ao educador para compreender as necessidades de suas classes. A análise do uso desses Sistemas de Gerência de Aprendizado a distância por meio de técnicas de mineração de dados é uma forma de obter informações relevantes que permitam ao educador observar essas necessidades e modificar seus cursos de acordo. O objetivo deste trabalho é elaborar uma metodologia de trabalho que permita abordar problemas dessa natureza de forma objetiva e flexível, facilitando identificar potenciais problemas na análise e pontos de retorno adequados para correção e retomada do processo. Um conjunto de etapas é elaborado para compor esta metodologia e em seguida colocado à prova com um conjunto de dados reais obtidos através da instância do TIDIA-Ae utilizada pela UNICAMP como auxiliar às aulas presenciais. Os resultados mostram a eficácia do método proposto e permitiram a observação de diversos problemas devido à maneira de utilização do sistema por alunos e professores.

Palavras-chave: Mineração de Dados, Sistemas de Gerência de Aprendizado, Ambientes Educacionais Online, TIDIA-Ae, SAKAI.

Abstract

Computer-based distance education is becoming popular as computational resources and the Internet expand. Nevertheless, educators may have difficulties to understand the necessities of his classes and therefore improve their courses. Usage analysis of these distance Learning Management Systems through data mining techniques is a way of obtaining relevant information that allow the educator to observe some of the classes' needs and modify his courses accordingly. The goal of the work described in this thesis is to elaborate a methodology to allow tackling problems of this nature in an objective and flexible way, easing the identification of potential problems in the analysis and adequate points of feedback to correct and retake the process. A sequence of steps is elaborated to constitute this methodology and test it with real data obtained from the instance of TIDIA-Ae used by UNICAMP as an auxiliary to classes in campus. The results show the efficiency of the proposed method, though some problems surfaced on these results originated from the way the system is employed by students and teachers.

Keywords: Data Mining, Learning Management Systems, Online Educational Environments, TIDIA-Ae, SAKAI.

Agradecimentos

Ao meu orientador Prof. Ivan L. M. Ricarte, pela paciência, enorme disponibilidade e orientação.

À minha família pelo apoio incondicional e irrestrito. Meu pais Geraldo e Sônia pelo bom humor, apoio financeiro, cobrança e constantes tentativas de me levantar o ânimo, sempre bem vindas. Minhas irmãs pela inspiração através de suas constantes conquistas acadêmicas.

À Regina Mitsue Azuma, por estar lá por mim, sempre.

Aos amigos que me acompanharam desde os tempos do Projeto Harpia e que me acompanham agora no LaRCom, partilhando alegrias, tristezas e muito bom humor.

Aos amigos do Mangue, tanto aqueles que foram embora quanto os que ficaram neste casarão que todos chamamos de lar e partilhamos bons e maus momentos.

Ao Prof. Leonardo Mendes, pelas oportunidades oferecidas e apoio financeiro.

Aos desenvolvedores do ambiente TIDIA-Ae do Laboratório e-labora, da Unicamp, e dos analistas do Centro de Computação da Unicamp, pela disponibilização dos dados que permitiram a realização do estudo de caso.

Aos meus pais, Geraldo e Sônia.

Sumário

Lista de Figuras

Glossário

FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo

Hibernate - Arcabouço em Java que visa simplificar o acesso a bases de dados relacionais

JPEG ou JPG - *Joint Photographic Experts Group*, adotado como nome para um popular formato de compressão de imagens

LMS - *Learning Management Systems*, ou sistemas de gerência de aprendizado

MySQL - Sistema gerenciador de base de dados relacional, disponível livremente

PDF - *Portable Document Format*, formato usual para representar documentos para impressão e visualização

SAKAI - Sistema de Gerência de Aprendizado no qual o TIDIA-Ae é baseado

SOM - *Self-Organizing Map* ou Mapa Auto-Organizável

TIDIA-Ae - Tecnologias da Informação para Desenvolvimento da Internet Avançada - Aprendizado Eletrônico

Univesp - Universidade Virtual do Estado de São Paulo

Web Logs - Registros de acesso mantidos por servidores de aplicação

Trabalhos do Autor

1. G. R. Falci Jr., I. L. M. Ricarte. “Metodologias de Mineração de Dados aplicadas a Ambientes Educacionais Online”. Anais do II EADCA, Campinas, São Paulo, Brasil, pg. 89 - 92, Março de 2009.

2. G. R. Falci Jr., I. L. M. Ricarte. “Uma Metodologia para Realização de Mineração de Dados sobre Ambientes Educacionais Online”. Aceito para publicação como artigo completo no CISTI’2010 — 5ª Conferencia Ibérica de Sistemas y Tecnologías de Información, Santiago de Compostela, Espanha, Junho de 2010. (Não publicado por falta de recursos para participação no evento.)

3. Ivan Luiz Marques Ricarte, Geraldo Ramos Falci Junior. “ A Methodology for Mining Data from Computer-Supported Learning Environments”. Submetido à Informática na Educação: Teoria & Prática, número especial sobre e-Learning e Mineração de Dados.

Capítulo 1

Introdução

Em toda forma de ensino acadêmico é um desafio do educador entender as percepções e reações dos alunos ao curso ministrado. A partir desse entendimento, o professor consegue ampliar e aperfeiçoar o curso, revisando o material repassado aos alunos e a melhor forma de apresentá-lo. Na educação tradicional em sala de aula, muito desse entendimento vem do contato direto com os alunos. Mas no ensino a distância e nas interações assíncronas, esse contato em geral se perde. É necessário obter tais informações de outra maneira.

No caso específico de cursos *online*, parte das informações necessárias ao professor pode estar ocultas em meio a uma enormidade de dados de acesso e registros de atividades dos alunos. Técnicas de mineração de dados podem ser aplicadas para localizar, filtrar e organizar informações a partir desses dados, permitindo assim prover, de uma forma mais simples e direta, um retorno ao professor ou autor de conteúdos sobre o uso que está sendo feito do material disponibilizado.

Romero et al.[24] observam que "Milhares de cursos *online* foram lançados nos últimos anos. Entretanto, a maioria deles é meramente uma rede de páginas Web estáticas. Isto levou a problemas de orientação e compreensão para os estudantes uma vez que a navegação nestes cursos é completamente irrestrita." Uma possível consequência é a possibilidade de haver porções do conteúdo que dificilmente sejam visitadas pelos alunos, necessitando uma revisão na organização destes cursos para que tais conteúdos sejam mais facilmente alcançáveis.

1.1 Motivação

O aumento significativo do acesso a computadores e à Internet nos últimos anos tornou a educação a distância uma opção muito mais atraente e funcional para o ensino. Com o passar dos anos, diversos sistemas diferentes foram criados para servir de base operacional aos inúmeros cursos que surgem todo ano. Variando de simples conjuntos de páginas estáticas a complexos conjuntos de ferra-

mentas educacionais, torna-se progressivamente mais difícil para o educador manter-se atento a todas as atividades realizadas pelos alunos no ambiente, assim como garantir que eles sigam os padrões estipulados de acesso ao conteúdo para garantir melhores resultados. Embora muitos desses sistemas não se preocupem em manter registros específicos do uso de suas ferramentas pelos usuários, administradores, professores e alunos, a maioria permite que esses dados fiquem registrados de alguma maneira, como por meio dos registros de acessos realizados a servidores associados ao ambiente de apoio ao ensino, como servidores de aplicação ou repositórios de objetos educacionais.

O que distingue os dados resultantes de atividades baseadas na Web em geral e em atividades de aprendizado *online* em particular, é a grande complexidade da informação e o volume dos dados coletados, assim como o fato de que a extração simples das informações não é possível [32], elas precisam ser deduzidas a partir dos dados disponíveis por meio de técnicas de mineração e visualização de dados adequadas. Todas essas informações, sejam elas intencionalmente guardadas ou apenas residuais, oferecem um potencial amplo para a melhoria desses sistemas educacionais que as geram. Por meio delas é possível não apenas orientar o educador a tomar decisões mais adequadas para melhorar seu curso, mas também criar novas ferramentas anexas ao sistema que possam oferecer um retorno também ao aluno, ajudando-o a ampliar a qualidade de sua experiência com o material disponibilizado. Técnicas de mineração de dados oferecem uma ampla gama de ferramentas para trabalhar com esses dados, permitindo explorar e identificar nos mesmos informações relevantes para construir essas repostas ao educador e aos alunos. Assim, torna-se possível automatizar grande parte do esforço de análise necessário ao educador para manter o curso *online*.

O problema abordado nesta dissertação: “Como oferecer a educadores e autores de material educacional *online* um retorno sobre o uso efetivo dos recursos que disponibilizam, sendo que os ambientes educacionais não oferecem dados adequados sobre esse uso?” já foi explorado por diferentes pesquisadores. As abordagens de alguns deles serão examinadas no capítulo seguinte. Técnicas de mineração de dados são frequentemente aplicadas para a solução deste problema e este trabalho estuda algumas das possibilidades oferecidas por estas técnicas e formaliza uma metodologia que visa introduzir o tema de forma simples e compreensiva.

1.2 Contribuição

O trabalho apresentado nesta dissertação tem por objetivo oferecer uma opção de aprendizado e compreensão das técnicas necessárias à realização da mineração de dados sobre sistemas de ensino a distância por meio de uma Metodologia desenvolvida com esta finalidade. Esta é baseada num estudo de diversos trabalhos previamente realizados na área buscando compreender melhor os problemas e vantagens inerentes às técnicas atualmente utilizadas.

A metodologia foi elaborada em passos correspondentes a atividades bem definidas e razoavelmente autônomas. Também foi elaborado um modelo de registro de dados suficientemente genérico para que possa ser adotado, com mínimas adaptações, por uma gama suficientemente ampla de sistemas de aprendizado *online* diferentes.

Um estudo de caso foi realizado para testar e melhor explorar a metodologia proposta, assim como destacar suas vantagens. Foram analisadas também as limitações da aplicação da metodologia ao caso estudado, explicitando os problemas oriundos de suas particularidades.

1.3 Organização

Os demais capítulos desta dissertação estão organizados de acordo com a estrutura descrita a seguir.

O Capítulo 2 examina as possibilidades já exploradas por outros pesquisadores na área de mineração de dados sobre sistemas de educação a distância focadas nos registros de uso comparando estas iniciativas com o trabalho aqui realizado. Também observa limitações nas técnicas geralmente aplicadas para mineração de dados nesses sistemas de educação a distância e destaca o que se pode esperar da análise de dados de uso. Como um elemento central nesses trabalhos é a técnica de agrupamento de dados, o capítulo também apresenta os princípios teóricos de alguns algoritmos de agrupamento utilizados em diversos trabalhos da área, assim como neste trabalho.

O Capítulo 3 apresenta um modelo de registro de informações relacionadas a ambientes educacionais *online* simples e portável. Uma metodologia de análise dos dados é proposta a partir deste modelo, visando eliminar algumas das limitações apresentadas. A metodologia é formalizada em cinco passos distintos e cada um é apresentado individualmente esclarecendo atores e possibilidades envolvidos.

O Capítulo 4 elabora mais a metodologia trabalhando-a sobre um estudo de caso. O sistema escolhido para o estudo é apresentado e seus registros de uso examinados dentro dos limites da proposta de trabalho. Toda a estrutura do trabalho desenvolvido é discutida, assim como as decisões de análise tomadas. Apresenta também os resultados obtidos com o estudo de caso. Os problemas encontrados e as razões de suas ocorrências são examinados, assim como a forma pela qual as características específicas do sistema estudado influenciaram nos resultados obtidos.

O Capítulo 5 conclui o trabalho com uma análise crítica da metodologia e dos resultados obtidos com sua aplicação. Possibilidades de estudos futuros a serem desenvolvidos aplicando a metodologia proposta são apresentadas, bem como cuidados a serem tomados para evitar que alguns dos problemas observados ocorram novamente.

Capítulo 2

Análise de Registros de Uso de Sistemas de Gerência de Aprendizado

Este capítulo apresenta um estudo dos registros de uso gerados por sistemas de gerência de aprendizado e as possíveis aplicações dos mesmos na obtenção de informações relevantes que possam auxiliar e orientar os professores na melhoria dos cursos disponibilizados por meio desses sistemas. Ele oferece uma revisão de trabalhos relevantes realizados na área por outros pesquisadores e demonstra as alternativas de exploração para análise. São examinadas as limitações das análises vistas e são apresentadas as expectativas da análise desses dados.

2.1 Registros de Acesso de Usuários

Sistemas de Gerência de Aprendizado, SGA, mantêm costumeiramente um registro de atividades dos usuários gravado e constantemente atualizado conforme o uso. Este registro é feito automaticamente em alguns sistemas, como por exemplo o SAKAI[15] e o Moodle[31], por meio de alguma ferramenta ou funcionalidade específica própria. O formato em que é mantido, assim como as informações contidas no mesmo, variam de sistema para sistema. Romero et al.[25] afirmam que "Sistemas SGA podem gravar quaisquer atividades em que os estudantes estejam envolvidos, tais quais leitura, escrita, realização de testes, realização de tarefas variadas e até comunicação com colegas. Eles provêem também uma base de dados que mantém toda a informação do sistema contendo desde aquelas pessoais dos usuários (perfil) até os dados de interação dos usuários."

O nível de detalhamento desse registro, assim como a fragmentação das informações através das várias tabelas da base de dados, variam muito de acordo com o sistema. Também é possível que um sistema não mantenha registro de uso algum (ou permita esta configuração por parte dos administradores) apoiando-se em sistemas externos para manter esses registros quando necessários.

Sistemas dependentes de Repositórios de Objetos Educacionais, tais como o Fedora [14], podem ter um registro de uso limitado em quantidade e qualidade de informações também.

Para se trabalhar com esses registros, algumas medidas geralmente são necessárias em maior ou menor grau, dependendo de fatores como o tipo do sistema, o tipo do registro a ser utilizado assim como o formato em que ele está armazenado e o processamento proposto. Desta forma, mesmo registros bem organizados podem necessitar de limpeza e remoção de dados imprecisos ou falhos, ou de uma reorganização em novas tabelas e bases de dados, seja para facilitar a associação de informações, seja para evitar que um eventual acesso direto à base de dados do sistema prejudique seu funcionamento ou até mesmo para que se possa concentrar apenas as informações desejáveis para a análise a ser realizada.

Hsu et al. [17] trabalham com os *logs* de acesso de um sistema de aprendizado *online* combinados com uma base de dados com informações dos alunos usuários do sistema e bases de dados contendo informações dos cursos e respectivos testes. Por meio de técnicas de mineração de dados, os autores buscam construir um sistema de recomendação que permita aos educadores melhorar os cursos a partir de conceitos combinando os resultados dos testes com os tempos despendidos pelos alunos nos referidos cursos. O objetivo é construir um modelo que ajude o educador a descobrir material mal formulado que possa resultar em comportamentos indesejáveis por parte dos aprendizes, como tempo insuficiente estudando e resultados bons nos testes ou tempo suficiente estudando e resultados ruins nos testes. O resultado é um Modelo de Retroalimentação de Resultados de Testes e seu design segue um padrão de não-interferência e automatização e permite a construção a partir de partes modulares de material educacional de cursos personalizados para cada aluno, solucionando problemas de má elaboração e organização dos cursos.

Sheard et al. [27] buscam responder a perguntas sobre o comportamento dos alunos usuários de um sistema específico de aprendizado *online*. Entre as características de uso observadas estão a frequência de acesso, o tempo gasto, a sequência de recursos acessados assim como quais recursos foram acessados. Mantendo estes dados em uma base de dados com o gerenciador *Microsoft Access*, foi aplicada uma combinação de resultados obtidos através da aplicação de técnicas de mineração de dados com os resultados de enquetes indagando os alunos sobre aspectos diversos do sistema para conseguir delinear os comportamentos dos alunos no sistema. Esta combinação verificou com sucesso que as inferências obtidas pela mineração de dados eram condizentes com as opiniões expressas nas pesquisas. Sheard et al. concluíram que a análise do uso do sistema pelos alunos é uma forma não intrusiva eficiente de aprender sobre seus hábitos durante o aprendizado.

Kampff et al. [18] desenvolvem um trabalho de mineração de dados durante uma disciplina ministrada a distância para alunos de cursos de graduação presencial oriundos de três polos diferentes (unidades presenciais onde ocorrem a aula introdutória e as provas finais) usando dados parciais, isto

é, obtidos antes do final do curso. O trabalho visa preparar alunos e educadores para trabalhar com as novas ferramentas disponibilizadas pelos ambientes de ensino a distância *online* e, por meio de mineração de dados, descobrir quais alunos têm maiores possibilidades de abandono ou reprovação no curso, alertando o educador com antecedência suficiente para que sejam tomadas as medidas necessárias. O trabalho busca relacionar os comportamentos dos alunos registrados pelo sistema (frequência de acesso, materiais acessados, atividades realizadas e entregadas) às realidades demográficas dos mesmos, isto é, faixa etária, sexo, curso e polo de origem. Um sistema de árvores de decisão classifica os alunos em grupos específicos, usando critérios pré-determinados pelos pesquisadores.

Romero e Ventura em seu trabalho [23] aprofundaram-se nas diversas possibilidades oferecidas pela mineração de dados educacionais, examinando uma grande gama de trabalhos desenvolvidos na área desde 1995 até 2005. Eles apontam as diferenças nos processos de mineração de dados entre aplicações comerciais e aplicações educacionais, diferenças essas que englobam o domínio, os tipos de dados disponíveis, os objetivos e as técnicas de mineração usadas. Procuram responder questões como a quem podem ser direcionados os trabalhos de mineração (alunos, educadores ou os responsáveis por manter os sistemas de aprendizado) e também métodos de análise já explorados dentro não apenas de ambientes de ensino a distância *online*, mas também em outras formas de ensino a distância e no ensino em classe. Diversas técnicas e abordagens diferentes para a mineração de dados são analisadas, assim como é apresentada uma extensa lista de trabalhos realizados por diversos pesquisadores da área. Romero e Ventura ainda enfatizam o quanto a área é nova e pouco explorada, apontando a necessidade de pesquisas com o intuito de facilitar e padronizar as ferramentas de mineração de dados, permitindo seu uso pelos próprios atores (educadores, alunos e administradores) em contato com o sistema e também a portabilidade dessas ferramentas entre diversos sistemas distintos assim como sua integração com os mesmos.

Em um trabalho mais recente, Romero e Ventura, em conjunto com García [25], expandem sua abordagem concentrando-se em construir um tutorial abrangente sobre as aplicações das diversas técnicas de mineração de dados educacionais vistas em seu trabalho anterior [23], aplicando o sistema Moodle [31] como base para essa experimentação e apoiando todo o trabalho realizado em ferramentas disponíveis livremente na Internet. Eles explicam a necessidade da criação de ferramentas de mineração de dados adequadas ao uso conjunto a sistemas educacionais e também a necessidade de maior simplicidade das mesmas, com interfaces mais intuitivas e amigáveis, permitindo que educadores possam utilizá-las com facilidade.

Trabalhar com registros próprios de um sistema é trabalhar com suas restrições e depender da estrutura de armazenagem de dados do mesmo. Apesar desse fato, um sistema devidamente construído com a mineração de dados em mente oferecerá inúmeras possibilidades a serem exploradas. SGAs com registros mais limitados podem ter seu dados complementados com bases de dados externas ou

a interação direta com os alunos através de enquetes. Possivelmente, por esses tipos de registros requererem esforços muito específicos de análise e transformação para serem aplicados à mineração de dados, os trabalhos analisados em sua maioria não entram nestes detalhes, concentrando-se na mineração de dados e seus objetivos propriamente. Os que entram, como o de Romero et al. [25], concentram o trabalho quase inteiramente em apresentar o maior número de opções o possível, o que pode ser um tanto intimidador para um novo pesquisador buscando uma introdução ao tema.

Isto motivou o trabalho apresentado nesta dissertação, uma metodologia de mineração de dados educacionais para aplicação em sistemas de ensino a distância *online*, cujo objetivo é apresentar de forma simples e objetiva o tema, buscando detalhar ao máximo cada passo necessário. Para tal finalidade, um cenário de uso para esta metodologia é construído tendo como base a utilização de registros de uso oriundos de uma instância do TIDIA-Ae, um sistema educacional brasileiro baseado no SAKAI, um sistema educacional livre desenvolvido e utilizado em diversas instituições ao redor do mundo. O cenário é desenvolvido com a ideia de auxiliar o educador a avaliar o interesse de seus alunos no uso do sistema e também a auxiliá-lo a averiguar o interesse dos alunos nos documentos que ele disponibiliza através do sistema. Para facilitar os trabalhos, foram reconstruídas as informações extraídas do sistema em bases de dados externas devidamente adequadas às necessidades do experimento.

Os registros próprios de um sistema têm a vantagem de reduzir a necessidade de formatação e filtragem necessários para a utilização dos dados, uma vez que os mesmos costumam ser gravados já seguindo algum padrão de formatação. No entanto, este caminho pode limitar as opções e a versatilidade do trabalho desenvolvido, uma vez que ele se prenderá a um formato específico de informações. Com este aspecto em vista, entre as outras possíveis opções de fontes de informação para se trabalhar estão os *logs* gerados pelos servidores de aplicação que suportam os sistemas de aprendizado *online*, como visto a seguir.

2.2 Registros de Acesso do Servidor

Uma fonte alternativa de informações de uso bastante utilizada, ainda que limitada em quantidade e qualidade de informações, são os registros de acesso dos servidores de aplicação sobre os quais os SGA são executados, conhecidos como *Web Logs*. *Web Logs*, como especificado por Zaïane [32], "costumeiramente contêm: o nome do domínio (ou endereço IP) da requisição; o *login* do usuário que gerou a requisição (se aplicável); a data e a hora da requisição; o método da requisição (*GET* ou *POST*); a identificação do recurso solicitado; o resultado da requisição (sucesso, falha, erro); o tamanho dos dados retornados; a *URL* da página requerida; a identificação do agente cliente; e um *cookie*, uma parcela de dados gerada por uma aplicação e trocada entre o cliente e o servidor. Uma

entrada é gravada no *log* automaticamente cada vez que uma requisição por um recurso é enviada ao servidor". Infelizmente, esses registros estão num formato inaplicável à mineração de dados. Eles requerem um extenso trabalho de limpeza, formatação e eventual combinação de informações para que possam se tornar compreensíveis a uma ferramenta de mineração de dados.

Ainda no mesmo trabalho, Zaïane propôs a aplicação de técnicas de mineração de dados usadas em *sites* comerciais, como lojas *online*, a um sistema de aprendizado *online* para, a partir do comportamento dos alunos da classe, recomendar ações e conteúdos a um determinado aluno enquanto ele acessa o material de um curso, tal qual um *site* comercial recomenda produtos a consumidores a partir das compras de outros consumidores de perfil similar. Essa previsão de comportamento dos alunos é extrapolada a partir dos *Web Logs* do servidor de aplicação usado em conjunto com o sistema. Embora não apresente resultados práticos, ele cita estar testando a proposta em um curso *online* e em um sistema *online* usado por médicos novatos no hospital da Universidade de Alberta e, neste caso, será avaliada pela medida de tempo economizado pelos alunos que seguem as recomendações do sistema contraposta ao tempo despendido pelos que não as seguirem.

Em um trabalho anterior, Zaïane e Luo [33] elaboram, de maneira similar, a partir de técnicas de mineração de dados e *Web Logs*, um trabalho de descoberta e reconhecimento de padrões de aprendizado utilizados pelos alunos de um curso permitindo aos educadores observarem e identificarem potenciais problemas na forma como os alunos acessam o material disponível, além de poder acompanhá-los através de suas ações dentro do sistema e avaliar a efetividade da estrutura do material no sistema. O artigo é encerrado mencionando que o sistema construído está em processo de teste sobre dois sistemas educacionais diferentes e que os resultados até então encontrados são excelentes. Eles ainda citam estar trabalhando em conjunto com os usuários dos SGAs para melhorar a interface do programa, tornando-a mais intuitiva e amigável para o propósito de avaliar o processo de aprendizado.

Santos [11] realizou um trabalho similar com *logs*, extraindo padrões de acesso às páginas de um dado sistema durante a realização de atividades pelos alunos com o objetivo específico de ajudar o educador a avaliar o material do curso apresentado. Em um trabalho anterior [10], o autor efetivamente desenvolveu um trabalho similar ao estudo de caso desenvolvido neste trabalho, lidando com *Web Logs* de acesso originários de duas edições de um curso ministrado através do WebCT pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), buscando a frequência de acessos do aluno a áreas específicas do site (*e.g.* leituras, participações em *chat*, fórum pré-determinados), a ordem e sequência dos acessos realizados durante a navegação no ambiente de ensino para executar as tarefas e padrões de comportamento de navegação para refletir a maneira como o aluno ou o grupo de alunos virtual executou suas atividades por tipo de atividade solicitada.

O trabalho desenvolvido por Castro et al. [5] concentra-se em descobrir que características dos

alunos são relevantes para determinar padrões de uso que permitam a formação de grupos personalizados de aprendizado através de informações de uso de um sistema pelos mesmos, como quantidade de sessões de uso, quantidade média de seleções em tela realizadas pelo usuário por seção, tempo entre a disponibilização de um documento pelo educador e o acesso, entre outras informações extraídas de *Web Logs* do servidor do sistema educacional estudado. Ao fim, ele se utiliza de uma técnica derivativa de Mapas Auto-Organizáveis para determinar estes grupos e quais dos tipos de informações se mostraram relevantes ou não para o processo.

Para o cenário desenvolvido neste trabalho, chegamos a considerar e experimentar este meio de obtenção de informações. O ambiente educacional que escolhemos, no entanto, por opção da equipe de desenvolvimento responsável (membros do Laboratório e-labora, da Unicamp) não registra as informações necessárias nos *Web Logs*. Esta opção foi feita para reduzir a quantidade de informações produzidas nos arquivos de *Web Logs* e facilitar a localização e identificação de problemas relacionados à execução do sistema. Como evidenciado pelos parágrafos acima e na seção anterior, o desenvolvimento do trabalho após a identificação, extração e adequação dos dados é basicamente o mesmo de quando se trabalha com registros próprios dos sistemas. De fato, chegamos a considerar algumas das métricas aplicadas por Castro et al.[5] no cenário desenvolvido neste trabalho.

Apesar das limitações grandes impostas pela reduzida quantidade de informações dispostas e pela complexidade natural do formato dos *logs* dos servidores, que dificulta bastante a localização e o tratamento dos dados, trabalhos baseados nessa fonte de dados têm a possibilidade de um escopo maior, uma vez que não são limitados pelo sistema. Ou seja, embora menos possa ser obtido a partir destas fontes de dados, os trabalhos sempre têm maior adaptabilidade entre sistemas diversos que os desenvolvidos sobre os registros próprios dos sistemas. Em teoria, quaisquer sistemas de aprendizado *online* que se apoiem num dado modelo de servidor seriam adaptáveis ao que for desenvolvido com relativa facilidade e propostas específicas de um desenvolvedor podem se tornar bastante expansíveis.

2.3 Limitações nas Análises Vistas

Como anteriormente mencionado na seção 2.1, uma das maiores limitações do uso dos *logs* próprios dos sistemas educacionais *online* é a falta de um padrão específico a ser seguido e a variabilidade das informações disponíveis em cada um. Isto dificulta e, na maioria dos casos, impossibilita a portabilidade de soluções entre sistemas, uma vez que formatos de armazenagem dos dados, assim como os próprios dados armazenados, diferem de um sistema para outro não apenas em tipos de informações armazenadas como também em qualidade de registro e armazenagem destas informações.

De fato, ao iniciar trabalhos de análise e mineração de dados em um sistema sem preparo prévio, é bastante comum que o analista conceba e implante no sistema seu próprio esquema de captura e

armazenagem de informações de uso; alguns trabalhos consistem inclusive no próprio sistema de captura, como o de Cardieri [4]. Desta maneira, o pesquisador consegue obter informações mais precisas e também focar o registro naquilo que for relevante para o trabalho de análise planejado, ao custo de talvez um semestre ou mais experimentando e gravando registros de uso antes de iniciar o trabalho que visa beneficiar o educador e o aluno.

Por conta disto, muitos pesquisadores direcionam seus trabalhos a lidar com *logs* gerados pelos servidores, uma fonte de dados rústica e limitada e que exige um amplo esforço de formatação e limpeza dos dados para prepará-los para a mineração, além do sempre presente risco de que parte da informação possa se perder em meio às massas de textos geradas pelos servidores em seus *logs*.

Outro aspecto negativo, menos controlável, dos sistemas educacionais *online* é a susceptibilidade dos mesmos a problemas inerentes a sistemas dependentes de redes de computadores, particularmente problemas de conexão, tais como quedas ou lentidão excessiva, que podem causar uma grande quantidade de ruído e informações incompletas no registro do mesmo. A queda da conexão atrapalha diretamente o curso de navegação do usuário, podendo alterar informações de caminho percorrido pelo mesmo através do sistema forçando-o a reentrar no mesmo, por exemplo. Lentidão, por outro lado, pode incitar o usuário a múltiplos pedidos de atualização de uma página tentando terminar de carregá-la e, ao mesmo tempo, criando uma grande quantidade de registros de acesso consecutivos à mesma página na base de dados do sistema. Problemas desta natureza foram verificados durante o estudo de caso realizado nesta dissertação, apresentado detalhadamente no capítulo 4.

2.4 O que Esperar de uma Análise dos *Logs*

Análises de uso oferecem oportunidades práticas de observar diversos aspectos do comportamento e do aprendizado de cada aluno assim como da classe em geral. Os registros evidenciam quais materiais os alunos acessam e em que ordem eles os acessam, podem registrar o tempo que eles dedicam a cada material (embora este tempo possa não ser tão significativo, uma vez que o aluno pode abandonar o computador com o material aberto, por exemplo) e podem apontar quais são os materiais mais acessados e em que momentos do aprendizado estes materiais são revisitados.

Com essas informações, é possível traçar os caminhos percorridos pelos alunos ao acessar as páginas que compõem determinados materiais disponibilizados pelos professores, permitindo a eles ter uma boa noção de como a classe estuda o material disponível e encontrar pontos onde o material pode não ser tão claro ou estar organizado da maneira mais adequada para a compreensão dos alunos. Com a análise adequada, é possível identificar páginas mais frequentemente acessadas e investigar a fundo se as mesmas representam um ponto difícil para a compreensão do aluno ou se informações importantes estão inadequadamente concentradas em um único ponto do material, por exemplo.

É possível identificar alunos com um desvio significativo no padrão de acesso do material do restante da classe, o que pode ajudar a explicar um resultado ruim por parte destes alunos ou mesmo uma forma mais eficiente de estudo que possa estar tornando os resultados destes alunos melhores. De fato, comparar os padrões de estudo dos alunos de melhor desempenho com aqueles de médio ou baixo desempenho torna-se uma possibilidade viável através da análise de *logs*.

2.5 Algoritmos para o Processamento dos Dados

Fayyad et al. [13] citam seis diferentes objetivos possíveis para a mineração de dados. Eles consistem em classificação, que visa mapear através do aprendizado de uma função um dado item para uma de várias classes pré-definidas; regressão, que procura mapear um cada item dos dados a variáveis de predição e também descobrir relações funcionais entre essas variáveis; agrupamento, cujo objetivo é identificar um conjunto finito de grupos ou categorias que descrevam os dados; sumarização, a qual visa encontrar formas compactas de descrever subgrupos de dados; modelagem de dependências, que procura encontrar modelos descritivos que estabeleçam dependências entre variáveis; e finalmente detecção de mudanças e desvios que foca em encontrar mudanças significativas nos dados em relação a tomadas de medidas anteriores ou a valores normativos.

Romero et al. [25] citam, de forma similar, sete objetivos possíveis para a mineração de dados aplicada a ambientes educacionais online: determinar estatísticas de uso do sistema, aplicação de técnicas de visualização de dados, agrupamento, classificação, mineração de regras de associação (que se relaciona diretamente com a sumarização descrita por Fayyad et al.), mineração de padrões sequenciais, mineração de textos (focada em agrupar documentos por assunto, por exemplo).

O trabalho realizado nesta dissertação focou-se principalmente na mineração de dados de uso por ser um objetivo comum e bastante útil, sendo aplicável a quase qualquer ambiente educacional existente. Para a mineração de dados efetiva preferiu-se a aplicação de algoritmos de agrupamento devido à restrições de tempo e disponibilidade de dados apropriados para o estudo de caso.

Apresenta-se nesta seção os algoritmos de mineração de dados empregados ao longo do estudo de caso desenvolvido nesta dissertação. Estes são algoritmos bastante conhecidos e utilizados. Foram escolhidos em parte porque são bem documentados e têm ferramentas e implementações diversas bastante acessíveis o que facilita a replicação dos experimentos conduzidos aqui. Devido à baixa complexidade dos dados analisados no estudo de caso, também faria pouco sentido a aplicação de algoritmos mais sofisticados.

2.5.1 K-Means

O K-Means é um método de agrupamento de dados capaz de dividir e agrupar um conjunto de n dados em k grupos, ou *clusters*, de forma a maximizar a similaridade dos dados dentro de um grupo e minimizar a similaridade entre grupos distintos. A divisão visa atender de forma otimizada um critério de partição denominado função de similaridade ou função-objetivo, que pode ser representado por vários modelos matemáticos distintos dependendo da necessidade da aplicação. Entre estes modelos possíveis destacam-se a distância euclidiana entre os objetos do conjunto de dados e o erro total para a quantidade k de grupos definida. A quantidade máxima de grupos para um dado conjunto de dados é naturalmente igual à quantidade de dados distintos presentes no conjunto inicial. Cada dado n_i é composto de uma ou mais dimensões, geralmente um vetor de informações numéricas. Cada agrupamento é definido por um centróide $\mathbf{m}_i$ e tem como membros os n dados mais próximos a ele.

O algoritmo parte da quantidade k de grupos definida pelo usuário e aponta aleatoriamente objetos do conjunto de dados como centróides para cada grupo. As sucessivas iterações associam cada objeto ao grupo mais próximo e ao final da iteração, os centróides de cada grupo são recalculados para o novo conjunto de dados pertencente a eles.

Tendo como função-objetivo o erro total, tem-se a equação:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |\mathbf{x} - \mathbf{m}_i|^2 \tag{2.1}$$

que define o erro total E, onde x é o ponto no espaço que representa um dado do conjunto inicial e m_i o centróide do *cluster* C_i, sendo ambos multidimensionais, o objetivo final do algoritmo é tornar os k grupos resultantes o mais compactos e separados possível. O algoritmo visa minimizar o erro total, geralmente chegando a um ótimo local [16]. O algoritmo geralmente tem como critério de parada um número específico de iterações consecutivas onde não ocorram mudanças na composição dos grupos ou quando o erro total atingir um valor estipulado suficientemente pequeno.

A necessidade de definir inicialmente uma quantidade específica k de grupos é vista como uma das desvantagens do algoritmo. Sua aplicação é limitada a problemas onde uma média é definível, tornando-o inadequado para problemas onde os dados tenham atributos categóricos [16]. O algoritmo também se mostra inadequado para problemas que envolvam grupos de formas não convexas ou de tamanhos muito diferentes. Outro problema significativo é a definição inicial dos centróides que pode levar o algoritmo a atingir um critério de parada tendo encontrado um agrupamento menos que ótimo dos dados.

2.5.2 Mapas Auto-Organizáveis

Mapas Auto-Organizáveis (MAOs), também conhecidos por Mapas ou Redes de Kohonen, consistem em um método de visualização e análise de dados de alta dimensionalidade, especialmente dados empíricos [20]. Os MAOs são um tipo de rede neural aritificial capaz de projetar um conjunto de dados de dimensionalidade elevada sobre um espaço de menor dimensão, tipicamente uma rede uni ou bidimensional de neurônios. Ao realizar esta projeção não-linear, o algoritmo tenta preservar ao máximo a topologia do espaço original, ou seja, procura fazer com que os neurônios vizinhos no arranjo apresentem vetores de pesos que retratem as relações de vizinhança entre os dados [35]. Desta forma, os vetores competem entre si para representar cada dado, o vencedor é aquele que representa melhor o dado e, portanto, ele tem seus vetores de pesos ajustados na direção do dado. Esta redução de dimensionalidade com preservação topológica permite ampliar a capacidade de análise de agrupamentos dos dados pertencentes a espaços de elevada dimensão [35].

O aprendizado competitivo representado pela rede neural do MAO se dá inicialmente através da apresentação, de forma aleatória e repetitiva, de um conjunto de dados representados por vetores num espaço $\Re^D$ a uma rede organizada de neurônios, onde cada neurônio possui um vetor de pesos no $\Re^D$.

O próximo passo é chamado **estágio competitivo**. Nele, a cada dado apresentado a rede, haverá uma competição entre neurônios pelo direito de representá-lo. Aquele que possuir o vetor de pesos mais próximo do dado, será o vencedor. A métrica que define esta proximidade é definida previamente pelo pesquisador. Esta estrutura competitiva é chamada *winner-takes-all*, o vencedor leva tudo.

No terceiro passo, **estágio cooperativo**, o neurônio vencedor é adaptado alterando seu vetor de pesos para aproximá-lo ainda mais do dado apresentado. A seguir, os neurônios vizinhos ao vencedor segundo o arranjo utilizado sofrem uma adaptação similar, porém de menor intensidade, para viabilizar a similaridade de vizinhança e permitir que dados similares ao apresentado sejam capturados pelos neurônios vizinhos ao vencedor, preservando, dadas as devidas proporções, a topologia dos dados no espaço original.

Seja um conjunto de dados de entrada $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$,$\mathbf{V} \in \Re^D$, de vetores $\mathbf{v}_n = [v_{n1}, \ldots, v_{nD}]^T \in \Re^D$ e $n = 1, \ldots, N$, onde cada $\mathbf{v}_n$ representa um dado no espaço D-dimensional [35]. O MAO, por sua vez, é definido por um conjunto de neurônios i com $i = 1, \ldots, Q$, dispostos em um arranjo tal que define a vizinhança de cada neurônio.

Cada neurônio i é representado por um vetor de pesos sinápticos $\mathbf{m}_i = [m_{i1}, \ldots, m_{iD}]^T \in \Re^D$, sendo todos eles conectados à entrada de dados, isto é, recebem o mesmo dado ao mesmo tempo. Assim, a distância entre cada vetor $\mathbf{v}_n$ e o vetor de pesos $\mathbf{m}_i$ de cada neurônio é calculada segundo uma métrica e o vencedor será aquele que tiver menor distância de $\mathbf{v}_n$. Um exemplo de métrica é a distância euclidiana, dada por:

$$d(\mathbf{m}_i, \mathbf{v}_n) = \|\mathbf{m}_i - \mathbf{v}_n\| = \sqrt{\sum_{j=1}^{D} |m_{ij} - v_{nj}|^2} \tag{2.2}$$

O neurônio vencedor tem, então, seu vetor de pesos adaptado para aproximá-lo do vetor entrada de modo que ele represente-o melhor. Também são adaptados os neurônios vizinhos estabelecendo uma interação local entre eles. Ao longo do aprendizado, é esta interação que promove a organização geral do mapa. O novo valor do i-ésimo neurônio no instante de tempo $(t+1)$ é definido pela equação de adaptação:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t).h_{ci}(t)[\mathbf{m}_i(t) - \mathbf{v}_n(t)] \tag{2.3}$$

onde t é um número inteiro positivo representando a coordenada discreta de tempo e $\alpha(t)$ é um escalar de valor decrescente a cada iteração que define a correção aplicada ao neurônio, ou seja, é a **taxa de aprendizado**. $h_{ci}(t)$ representa a **função de vizinhança** e é um tipo de *kernel* atenuador, seu valor é unitário quando $c = i$ e decresce conforme o neurônio vencedor m_c e o neurônio vizinho m_i divergem. A cada aumento do índice t a largura espacial do *kernel* no grid diminui também. Isto significa que $h_{ci}(t) \to 0$ quando $t \to \infty$. Traidicionalmente, $\alpha(t) \to 0$ nestas condições também.

Normalmente, $h_{ci}(t) = h(\|\mathbf{r}_c - \mathbf{r}_i\|, t)$, e $\mathbf{r}_c$ e $\mathbf{r}_i$ representam as posições dos vetores de índices c e i dentro do arranjo. Quando $\|\mathbf{r}_c - \mathbf{r}_i\|$ aumenta, $h_{ci}(t)$ sofre uma redução exponencial [35].

Os valores associados a esses índices e funções devem ser escolhidos de maneira cuidadosa e, portanto, não serão discutidos aqui. Para melhores referências quanto à determinação destes valores, assim como explicações mais detalhadas a respeito de mapas auto-organizáveis, veja [19] e [35].

2.6 Considerações Finais

Neste capítulo foram exploradas as possibilidades de dados para análise dentro de sistemas de aprendizado a distância *online* dentro do escopo das informações de rastro e registros gerados pelos próprios sistemas ou pelos servidores. Foram examinados alguns trabalhos realizados por outros pesquisadores sobre o assunto com origens similares de dados analisados. Foram observadas as limitações inerentes às análises estudadas e quais as expectativas de resultados para análises de registros de uso de sistemas educacionais *online*. Por fim, foram examinados os algoritmos de agrupamento aplicados no trabalho desenvolvido nesta dissertação.

Capítulo 3

Metodologia para Mineração de Dados

Este capítulo descreve e examina uma proposta de metodologia para a aplicação de técnicas de mineração de dados a ambientes educacionais *online*. A partir do exame de um modelo conceitual de *logs*, o capítulo descreve a elaboração de uma proposta seguida da construção efetiva da metodologia e seus passos. Ele descreve cada passo individualmente. Apresenta os atores responsáveis por cada passo, assim como também determina suas contribuições particulares para o processo. Neste capítulo, vêem-se ainda exemplos para cada passo da metodologia, oferecendo uma descrição dos dados envolvidos nos processos e as transformações sofridas pelos mesmos.

3.1 Modelo Conceitual dos Logs

É apresentado aqui um modelo conceitual, ilustrado pela figura 3.1, de um conjunto de classes que representa em um formato generalizado os dados de *logs* associados por um SGA, já contextualizados para a aplicação educacional. Tal modelo pode ser utilizado para a construção de uma base de dados relacional por meio de um mapeamento objeto-relacional. Para uso em linguagem Java, um arcabouço tal qual o *Hibernate*[22] provê as ferramentas necessárias para transformar este conjunto de classes em uma base de dados relacional (suportando diversos tipos de sistemas gerenciadores de bancos de dados) e permitir um diálogo direto entre ambos. A ideia é que o modelo implementado registre o bastante para que se obtenham os dados necessários para uma análise do uso do sistema, além de ser adaptável entre sistemas com esforço relativamente baixo.

A classe **Evento** representa os registros de atividades dos usuários mantidos pelo sistema. Cada instância dela, correspondente a uma entrada na base de dados, representa um registro específico. O registro agrega as informações referentes ao usuário que realizou a ação registrada, o curso que ele acessava e o conteúdo acessado pelo mesmo dentro do sistema. O tipo de ação realizada define o tipo de acesso ao documento realizado pelo usuário (leitura, escrita, alteração) e, embora este modelo

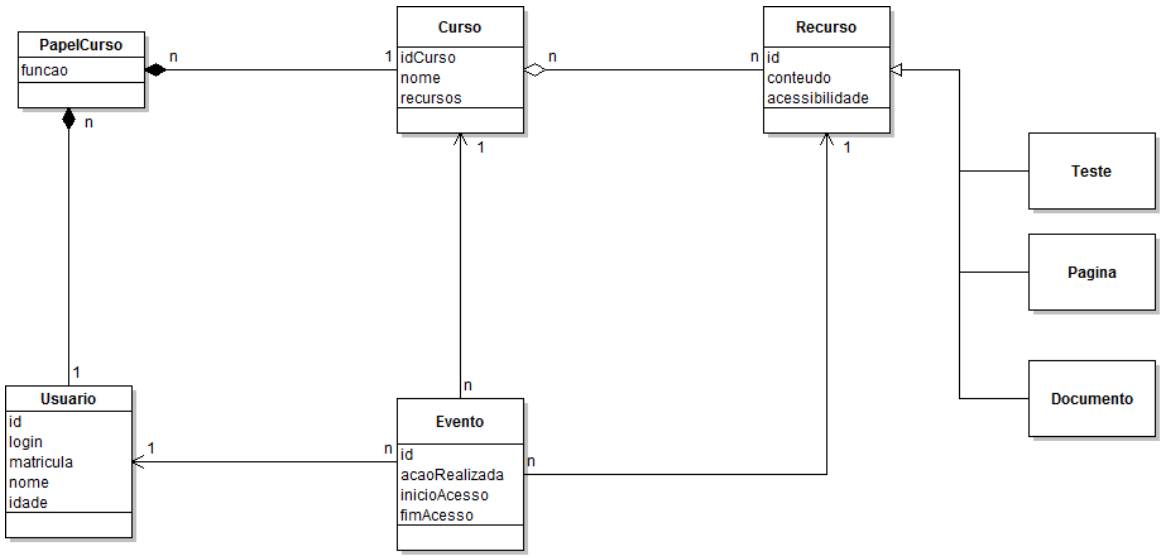

Fig. 3.1: Modelo Conceitual de Logs

preveja o registro do momento de início e de fim do acesso ao recurso, que pode ser obtido, entre outras opções, por meio de um pequeno aplicativo Java executando em conjunto com a ferramenta como demonstrado por Cardieri [4], o mais comum é que seja registrado apenas o momento do início do acesso.

Em conjunto com Evento, as classes **Usuario**, **Curso** e **Recurso** constituem a parte central do modelo e suas associações são o aspecto mais importante do mesmo. O conteúdo destas classes abrange atributos importantes a qualquer sistema SGA. **Usuario** define alguém com acesso ao sistema (um aluno, educador ou administrador, por exemplo), normalmente associado a algum dos Cursos disponíveis. É possível que um mesmo usuário tenha papéis distintos quando associado a mais de um curso (professor em um, aluno em outro, por exemplo) e, prevendo esta possibilidade, o modelo enquadra esta característica do usuário em uma classe própria, PapelCurso.

Os recursos compõem a parte interativa dos cursos, sendo que um mesmo recurso pode ser partilhado por diversos cursos diferentes. Variadas implementações da classe **Recurso** visam cobrir tipos diferentes dos mesmos, simbolicamente representados aqui por **Teste** (denominando um exercício ou avaliação do aluno para ser realizada dentro do sistema), **Pagina** (uma página qualquer de um curso) e **Documento** (um arquivo em PDF ou JPEG, por exemplo).

O modelo visa rápido e fácil acesso para qualquer *software* construído para análise dos dados registrados e a formatação precisa dos dados que ele mantém depende exclusivamente do sistema no qual ele for implementado. Diferentes sistemas podem especificar estes atributos, assim como as relações entre as diversas partes componentes do modelo de maneiras bastante particulares, e um esforço de adequação pode ser necessário ao implementar um sistema de *logs* baseado neste modelo.

Esta adequação poderá, no entanto, incorrer em uma maior complexidade no trato dos dados durante a análise dos mesmos. Dependendo do estágio em que o sistema estiver, particularmente estágios iniciais de planejamento e desenvolvimento, pode ser mais apropriado adaptá-lo ao modelo proposto.

3.2 Proposta

Tendo em mãos o formato dos registros de acesso armazenados, assim como os próprios registros feitos pelo sistema e as possíveis análises desejadas, é preciso traçar uma lista de tarefas, ou objetivos, a serem cumpridas para a obtenção dos resultados desejados.

É necessário um estudo do formato dos registros para estabelecer formas de analisá-los. Um formato similar ao apresentado na seção 3.1 permite uma ampla gama de análises além de apresentar informações em um formato bem organizado e de fácil exploração. No entanto, nem todos os sistemas apresentarão todos os dados disponíveis neste formato, e um estudo pode ser necessário para identificar as limitações dos registros do sistema e formas de contorná-las.

Tendo compreendido os dados e a análise desejada, deve-se determinar quais destes dados serão relevantes para o processo de análise e mineração escolhidos. Etapas de seleção, filtragem e formatação destes dados são realizadas neste ponto do processo. Selecionamos apenas os dados que são significativos para a análise pretendida. Também é aqui que eliminamos dados espúrios e registros incompletos ou errôneos que possam dificultar a análise ou até mesmo alterar seus resultados significativamente. A formatação visa adequar os dados para entrada nos algoritmos de mineração de dados escolhidos.

Por fim, os resultados do processamento dos dados pelos algoritmos escolhidos devem oferecer as respostas buscadas pelos educadores, ainda que possa ser necessária uma certa ordenação e adequação dos resultados para melhor apresentação e compreensão por parte do educador. É possível que não sejam encontrados bons resultados em uma primeira tentativa, o que requer um retorno à análise de todo o processo para identificação de pontos problemáticos que precisem ser alterados ou repensados para análise adequada.

3.3 A Metodologia

A Metodologia proposta nesta seção é inspirada em trabalhos anteriores de outros pesquisadores, sendo que muitos deles trabalham em escopos separados do ensino a distância. A proposta aqui realizada tem muito em comum em particular com a definição de Descoberta de Conhecimento em Bases de Dados (DCBD) de Fayyad et al. [13], descrita como "um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e, finalmente, compreensíveis em dados". O método

é dividido em passos de forma similar à Metodologia apresentada aqui e em seu artigo eles ainda examinam a mineração de dados em maior profundidade, suas características como parte do processo de DCBD, as possibilidades que ela oferece e também alguns dos desafios que podem surgir quando se trabalha com algoritmos de mineração de dados.

A Metodologia aqui apresentada foi idealizada em um conjunto de cinco passos a serem aplicados sobre um sistema de ensino a distância *online*. O objetivo é partir das estruturas de dados internas ou anexas ao sistema passíveis de conterem informações significativas e, por meio dos diversos processos envolvidos, determinar novas informações relevantes, encontrar e destacar, ou organizar, informações úteis que possam não estar tão óbvias ao educador dentro do sistema.

Os passos visam orientar e organizar em etapas um trabalho de análise desses dados, apontando quais caminhos podem ser seguidos a cada etapa e ajudando a entender o que se deve conseguir a cada etapa para que seja possível dar continuidade ao trabalho. O ideal dos passos é que eles sejam guias e não um caminho pré-determinado único.

Romero et al. [23] explicitaram as particularidades da aplicação da mineração de dados a um sistema de ensino a distância quando comparada à aplicação sobre sistemas de *e-commerce*. As diferenças abrangem o domínio, os dados disponíveis, os objetivos da mineração e as técnicas aplicadas.

Estas diferenças podem ser descritas de maneira mais específicas como: o foco em guiar o aluno ao aprendizado, as múltiplas fontes de dados disponíveis, o objetivo subjetivo de aprimorar o aprendizado e as características especiais do ensino a distância que requerem adaptações nas técnicas de mineração de dados para o seu uso. Tendo-se uma compreensão adequada destas diferenças, é possível adaptar a metodologia proposta para o uso em sistemas mais gerais, em princío, os dedicados ao *e-commerce*. Esta possibilidade, no entanto, não será abordada neste trabalho.

A Figura 3.2 ilustra todo o processo.

A seguir, os passos são detalhados individualmente.

3.3.1 Primeiro Passo — Compreensão do Sistema de Gerência de Aprendizado

Inicialmente, é necessário um estudo do sistema educacional e das ferramentas e ambientes funcionando anexos a ele. É preciso entender como e onde esses programas captam e armazenam seus dados, assim como quais são os dados mantidos pelos mesmos. A partir dessa compreensão, faz-se a pré-seleção das fontes de dados mais promissoras a serem exploradas nos passos seguintes. Essa escolha deve levar em conta os tipos de dados presentes e o potencial que eles têm para descrever os comportamentos e hábitos dos alunos em relação aos diversos aspectos do sistema educacional.

Considerando apenas sistemas educacionais disponibilizados pela Web, uma primeira fonte de

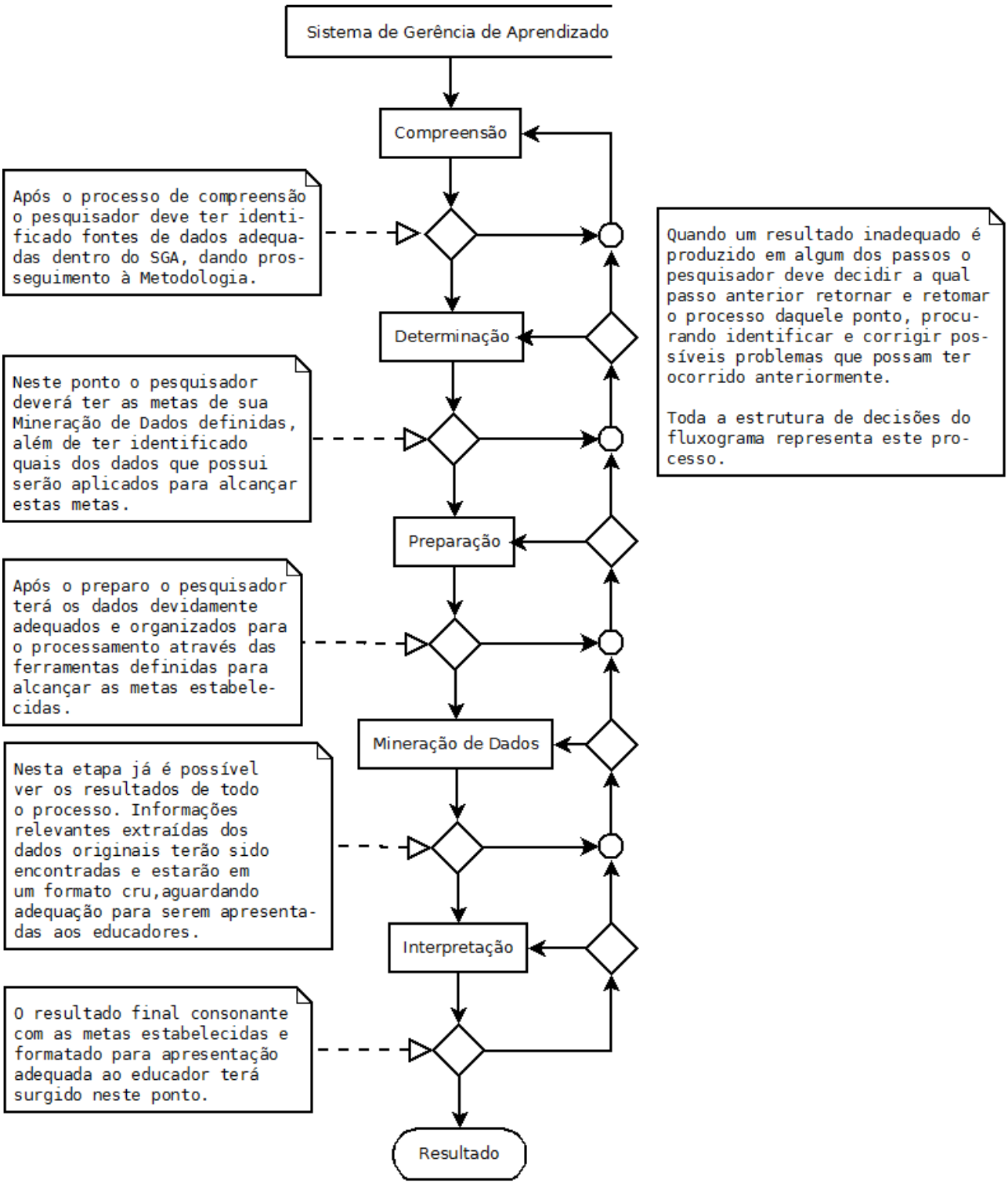

Fig. 3.2: A Metodologia representada graficamente.

dados a ser considerada é o conjunto de registros de acesso mantido pelo servidor de aplicação junto ao qual o ambiente educacional foi disponibilizado. O Apache Tomcat [7], por exemplo, utiliza-se do arcabouço Commons Logging[30] para controlar seu sistema de *log*. Isto permite ao desenvolvedor optar entre sistemas distintos de *log*, embora use primariamente uma versão adaptada dos pacotes próprios da linguagem Java (o pacote *java.util.logging*). Esta adaptação permite que os *logs* sejam gerados de forma particular para cada aplicação *web* rodando no servidor, ao invés de limitar a um único formato por Máquina Virtual. Os *logs* são configurados em níveis hierárquicos diversos aplicáveis a cada pacote da aplicação separadamente. Quanto mais alto o nível configurado para a geração de *logs*, maior será o volume de dados gerado e gravado pelo servidor durante a operação. Estas configurações são *SEVERE, WARNING, INFO, CONFIG, FINE, FINER, FINEST* e *ALL*. Elas ajustam os registros para gravar desde apenas erros severos até todo e qualquer tipo de mensagem informativa gerada pelo servidor ou aplicativos rodando nele.

Em algumas implementações mais antigas, os sistemas educacionais são organizados de forma monolítica, com todas as funcionalidades centralizadas em uma única aplicação associada a um servidor de aplicação. Em alguns desses casos, o ambiente mantém alguns dados adicionais sobre usuários e uso de suas ferramentas, geralmente por meio de um sistema gerenciador de banco de dados associado ao ambiente. Nesses casos, essas bases de dados constituem uma rica fonte de dados para o processo de mineração.

O Teleduc é um ambiente para a criação, participação e administração de cursos na Web[1]. Em sua terceira versão (o ambiente encontra-se atualmente em sua versão 4.2.1), o Teleduc era estruturado num grupo de ferramentas ligadas a uma ferramenta central, **Atividades**. Souza[29], em seu trabalho, descreve as principais ferramentas desta versão do ambiente com as quais os alunos interagem diretamente. A ferramenta **Acessos** registra os acessos dos participantes e quais ferramentas eles usam. Embora o acesso à mesma seja livre a todos os participantes, os responsáveis pelo sistema podem configurá-la para limitar este acesso. Souza cita que esse acesso livre ajuda a aumentar a interatividade entre os alunos, que podem detectar através da ferramenta outros colegas *online*, mas critica o formato de seus relatórios, sobrecarregados com números. O conteúdo armazenado e disposto por essa ferramenta é uma ótima fonte de informações de uso para o desenvolvimento de trabalhos de mineração de dados justamente por esta grande quantidade de dados numéricos usada para descrever os acessos, facilitando a aplicação de algoritmos diversos de mineração de dados.

Assim como o Teleduc 3, o Moodle, que é um sistema educacional construído de forma modular, possui uma ferramenta que permite ao educador ou administrador do sistema gerar um relatório completo, ainda que em um formato rudimentar, de todas as atividades realizadas no sistema pelos alunos. O sistema mantém todos os registros em uma única base de dados relacional, sendo MySQL e PostgreSQL as melhor suportadas, ainda que não as únicas [25].

A tendência mais recente é ter esses ambientes organizados por meio de arquiteturas de *software* que permitem uma maior flexibilidade de configurações de ferramentas. Nesses casos, o ambiente atua como uma plataforma à qual ferramentas externas podem ser anexadas, e assim operar no contexto do ambiente. Um exemplo comum desse tipo de ferramenta externa são os repositórios de objetos educacionais, que podem ser utilizados para manter os recursos educacionais de forma independente de ambientes ou cursos. Em tais situações, é importante observar se essas ferramentas externas também não mantêm seus próprios registros.

Um bom exemplo de repositório de objetos digitais, isto é, objetos que englobam conteúdos digitais e métodos de acesso e modificação destes conteúdos(inclusive conteúdos educacionais), é o Fedora Commons[14] que é amplamente utilizado por universidades para manter bibliotecas digitais, além de centros médicos, museus e coorporações entre outras entidades[28]. Desenvolvido em um formato modular para facilitar a amplicação e modificação de suas funcionalidades, o Fedora Commons não possui um registro de acessos aos seus objetos. No entanto, possui funcionalidades de versionamento de objetos e um sistema de diários (*journaling*)[8] que mantém um registro de todas as modificações efetuadas no repositório com o objetivo de permitir a reconstrução passo a passo de um repositório sem a necessidade de se basear em *backups* e sem sofrer com as desatualizações dos mesmos. Estes diários, gravados em arquivos em formato **xml**, podem servir como uma forma rudimentar e limitada de extração de dados de uso do sistema uma vez que são configuráveis para manter informações extras a respeito dos usuários que desempenham as atividades registradas por ele.

O AGORA (do espanhol Ayuda para la Gestión de Objetos de Aprendizaje Reutilizables ou Ajuda para a Gestão de Objetos Educacionais Reutilizáveis) é um repositório de objetos digitais educacionais cujo foco principal é atender as necessidades de professores e *designers* no processo de construção de objetos educacionais de acordo com suas necessidades, partindo de recursos de aprendizado digitais e utilizando tecnologia avançada[9]. Prieto et al. [21] examinaram detalhadamente o AGORA, particularmente sua Arquitetura de Sistema de Recomendação, que visa prover o auxílio adequado para que o educador construa da melhor maneira possível os objetos educacionais que comporão os cursos ministrados por intermédio dele. O trabalho de Prieto foi desenvolvido sobre uma versão de testes do AGORA e prevê a disponibilização futura de uma versão *plugin* do AGORA para sistemas educacionais tais como o Moodle.

3.3.2 Segundo Passo — Determinação dos Objetivos da Mineração de Dados

Enquanto a compreensão do sistema é uma função geralmente desempenhada pelos especialistas em software responsáveis por mantê-lo, é comum que o educador estabeleça as metas a serem cumpridas. Estas metas podem envolver tentativas de prever comportamentos específicos dos alunos diante do curso ou dos testes ou algum aspecto do sistema, ou identificação de padrões de acesso às páginas

ou recursos disponíveis ou encontrar pontos do curso que precisem ser revisados e melhorados, entre outras possibilidades.

Embora consista num processo independente da compreensão do sistema, podendo inclusive ser desempenhado antes do mesmo, a determinação precisa que seus resultados, as metas estabelecidas, estejam em consonância com as possibilidades oferecidas pelo sistema. É necessária uma interação entre os atores, aquele que definiu as metas (presumivelmente o educador) e aquele que conhece e compreende o sistema (o especialista em software), para que seja possível reconstruir as metas dentro das possibilidades oferecidas pelas fontes de dados existentes dentro do sistema. Diversas características dos alunos podem ser determinadas por meio de dados comumente disponíveis nos registros dos sistemas de aprendizado a distância.

Registros de acesso podem levar a traçar comportamentos de grupos de alunos ou alunos individuais a partir da forma pela qual os alunos acessam o sistema, isto é, quais páginas eles visitam, qual a ordem em que as visitam, quanto tempo eles permanecem nas páginas, quais das ferramentas disponíveis eles aplicam em seu aprendizado. As similaridades entre a maioria dos alunos definirá o comportamento da turma, ao passo que as diferenças definirão quais os alunos que se desviam do grupo.

Resultados de testes e questões individuais podem identificar fraquezas de alunos ou grupos de alunos com determinado material ou inadequação deste material. Uma questão com alto índice de erros pode referenciar um material mal desenvolvido pelo educador, ou que esteja sendo ignorado pelos alunos por alguma razão, ou pode ter sido mal elaborada e ser inadequada para a avaliação do conhecimento dos alunos a partir do conteúdo oferecido pelo educador no curso. É possível determinar as fraquezas específicas dos alunos através de seus resultados também, redirecionando-os para um melhor estudo dos tópicos em que falharam, ou mesmo recomendando-nos a conteúdos mais detalhados sobre os assuntos.

A ordem de acesso às páginas do conteúdo pode estabelecer a necessidade de uma alteração ou simplificação na forma em que o conteúdo é apresentado, ou, possivelmente, ajudar a prever resultados de testes. Um aluno de bons ou maus resultados pode estar seguindo uma ordem de estudos do conteúdo não planejada ou esperada pelo educador. De forma similar, um grupo de alunos pode estar acompanhando o material disponível de uma maneira inadequada, denotando uma necessidade de alterar o formato do curso para tornar mais compreensível a ordem de estudos intencionada pelo educador.

Tendo estes objetivos definidos, é necessário identificar quais dados disponíveis estão associados às metas em maior ou menor grau. Em alguns casos, todo o trabalho desenvolvido pode ser focado em descobrir se certos dados descrevem ou se relacionam com algum aspecto específico do curso ou do sistema.

3.3.3 Terceiro Passo — Preparação

Definidas as metas e as informações que serão usadas para alcançá-las, é necessário preparar esses dados para o processamento, adequando-os aos formatos apropriados para aplicação nas ferramentas de pesquisa, eliminando ruídos e inconsistências que possam estar presentes nos dados. O processo de reorganizar os dados que serão usados em novas estruturas, geralmente para facilitar e agilizar o acesso aos mesmos, também é parte deste preparo. Este é um passo desenvolvido em conjunto pelo especialista responsável pelo sistema e o especialista em mineração de dados. O primeiro, por conhecer bem os dados extraídos do sistema e o segundo, por conhecer bem o formato necessário para a aplicação das técnicas de mineração aos dados.

O processo de preparo dos dados envolve limpeza, verificação de consistência, reorganização e sumarização dos mesmos. A limpeza dos dados é feita removendo dados desnecessários para o processo, deixando apenas aqueles que serão efetivamente analisados através da mineração. A consistência é garantida removendo ruídos oriundos de falhas de gravação ou uso do sistema, tais como repetidas requisições consecutivas devido a falhas de conexão, ou dados incorretamente gravados ou mesmo incompletos. A reorganização e a sumarização involvem procedimentos aplicados para adequar o formato em que os dados estão registrados à mineração, alterando entradas presentes nos registros mudando a ordem em que estão dispostas, a quantidade e os tipos de informações contidas por cada entrada, eliminando o desnecessário sem, no entanto, alterar a informação contida no registro original que será processada. Faz-se necessário um cuidado para que não se percam ou adulterem as informações contidas nos dados devido às transformações aplicadas.

Entre as muitas possibilidades para a realização dessas tarefas preparativas estão a construção e aplicação de softwares para leitura, formatação e armazenamento apropriado dos dados. Também se incluem o estabelecimento de uma estrutura adequada para a informação e a eliminação de ruídos, tentativas seguidas de autenticação ou acesso a páginas e documentos por falhas de rede, tempos de acesso extremamente elevados por abandono do computador, entre outras fontes de ruído possíveis.

3.3.4 Quarto Passo — Mineração de Dados

Com os dados adequados, chega o momento de processá-los em busca dos resultados estabelecidos pelas metas definidas no segundo passo. Este é o ponto crucial da contribuição do especialista em mineração de dados para o desenvolvimento da metodologia, uma vez que é desempenhado somente por ele.

Análises comportamentais geralmente envolvem a busca pela definição de grupos de comportamentos específicos assim como de indivíduos isolados dos grupos mais gerais através dos dados de uso do sistema. Desta forma, espera-se que os comportamentos descritos por estes grupos ajudem a

identificar problemas diversos no curso, como materiais que precisem ser revistos ou reorganizados ou complementados, alunos que necessitem de atenção especial do professor ou que estejam demonstrando desinteresse pelo curso, entre outras possibilidades. Para encontrar estes grupos, geralmente aplicam-se técnicas de agrupamento, supervisionado ou não-supervisionado. Um conhecimento prévio das características que constituem os grupos desejados permite a aplicação de técnicas de agrupamento supervisionado.

O trabalho pode ter um foco em organizar e filtrar informações úteis presentes em meio aos dados, não analisando e elaborando sobre as mesmas, mas destacando-as para que o educador possa ter fácil acesso a elas e possa traçar suas próprias conclusões. Um trabalho desenvolvido em maior proximidade com o educador pode, inclusive, permitir um sistema que suporte mais facilmente a obtenção dessas conclusões.

3.3.5 Passo Final — Interpretação

O último passo é verificar a adequação dos resultados às metas estabelecidas. Desta forma, para um trabalho focado em agrupamentos, deve-se realizar a análise dos grupos assim como da influência de cada tipo de dado utilizado no resultado final. Perguntas como se informações relevantes ou desejáveis foram encontradas ou explicitadas ou se os grupos constituem comportamentos suficientemente únicos dentro do proposto pelas metas ou ainda se os resultados são capazes de prever com precisão os resultados de testes ou mesmo os comportamentos dos alunos devem ser respondidas nesta etapa.

Na eventualidade de resultados insatisfatórios, é necessário determinar qual dos passos anteriores produziu uma saída inadequada e retornar a ele, refazendo todos os passos seguintes. Por exemplo, a idade dos alunos ou mesmo a faixa etária deles pode parecer ser um bom indicativo comportamental dos mesmos, mas, durante a qualificação, resultar em grupos não consonantes com os comportamentos ou resultados de testes dos alunos. Neste caso, um retorno a um passo anterior deve ser feito e um novo dado deve ser escolhido para realizar o estudo comportamental. Pode-se aplicar o K-Means sobre um conjunto de dados e os grupos resultantes serem insatisfatórios, levando à necessidade de escolher um algoritmo diferente para processar este conjunto de dados. Isto levaria provavelmente a um retorno ao passo de preparação, pois os dados necessitariam de uma nova formatação para uso com outro algoritmo. Inúmeros problemas podem ocorrer ao longo de todo o processo e pode ser necessário escolher novas fontes de dados, novos dados, em alguns casos até novas metas ou novos algoritmos para processamento dos dados para poder obter respostas melhores e mais significativas ao final de todo o procedimento.

3.4 Considerações Finais

Apesar de ser elaborada para seguir uma lógica linear, a metodologia permite uma certa liberdade na ordem de realização de seus dois primeiros passos, embora a continuidade dela ainda dependa da combinação e consonância dos resultados de ambos passos. A cada novo passo, é possível detectar potenciais problemas e retornar aos anteriores, reavaliando e alterando decisões e resultados.

Três figuras importantes participam do processo, ainda que uma mesma pessoa possa cumprir todas estes papéis: o educador, que é quem define o curso e também quem tem maior ciência do necessário para melhorá-lo; o especialista em software, responável pelo sistema e aquele que melhor conhece suas características e, finalmente, o especialista em mineração de dados, que é o responsável por unir os conhecimentos dos outros sobre o curso *online* ministrado e extrair as informações desejadas. Nem todos estes atores são necessários em todos os passos, mas cada qual exerce uma participação importante naqueles em que participa.

Para que se pudesse examinar melhor estes passos e como eles se relacionam, optou-se por realizar um estudo de caso. Ele é descrito no capítulo seguinte desta dissertação, detalhando extensivamente como se desenvolve cada passo da metodologia.

Neste capítulo, a partir de um modelo conceitual de *logs*, foi detalhada toda uma metodologia para a aplicação de técnicas de mineração de dados a ambientes educacionais *online*, examinando cada um dos passos componentes separadamente, seus atores responsáveis e apresentando exemplos de desenvolvimento destes passos, onde necessário.

Capítulo 4

Um Estudo de Caso com o Ambiente TIDIA-Ae

É apresentado neste capítulo um estudo de caso da metodologia discutida no capítulo anterior aplicado sobre o ambiente educacional TIDIA-Ae. São vistas as razões para sua escolha, assim como as possibilidades oferecidas pelo mesmo. É feita uma apresentação detalhada da aplicação de cada passo da metodologia, explicitando entradas e saídas para cada passo, assim como os processos envolvidos. Uma análise é realizada sobre os resultados obtidos destacando suas características principais.

4.1 Escolha de um Sistema Apropriado como Estudo de Caso

Foi selecionado para este estudo o ambiente TIDIA-Ae, desenvolvido no projeto **Tecnologias da Informação para Desenvolvimento da Internet Avançada**, da FAPESP[12]. A versão atual desse ambiente foi construída sobre a plataforma Sakai[15]. A versão do ambiente analisada foi criada no Laboratório e-labora, da UNICAMP, e foi adotada para utilização na Universidade Virtual do Estado de São Paulo (Univesp). A UNICAMP está realizando testes para avaliar sua adoção como plataforma de apoio aos cursos de graduação e de pós-graduação presenciais da universidade. Esse ambiente é uma opção vantajosa por ser um sistema em amplo desenvolvimento no país, assim como um derivado de um sistema desenvolvido e utilizado em escala mundial sem se distanciar muito do mesmo. Estas características dão a este trabalho o prospecto de poder ser expandido e integrado futuramente a esses sistemas, dadas as devidas adaptações.

4.1.1 TIDIA-Ae

Surgido da união de esforços entre a FAPESP e pesquisadores das principais universidades do estado de São Paulo, o projeto Tecnologia da Informação para o Desenvolvimento da Internet Avançada — Aprendizado Eletrônico tem o intuito de desenvolver um ambiente de colaboração e ferramentas de suporte e apoio ao ensino e aprendizagem com interações presenciais e a distância, síncronas e assíncronas[12].

Um de seus principais objetivos é o desenvolvimento de um amplo conjunto de ferramentas dedicadas ao conhecimento colaborativo e o ensino a distância. Essas ferramentas contemplam os três grandes grupos de ferramentas gerais de EaD — administração, coordenação e comunicação — e também ferramentas e conteúdos diversos. O *software* aberto sobre o qual este conjunto é desenvolvido permite extensas alterações conforme as necessidades de seus usuários. O ambiente de gerenciamento resultante deste esforço de pesquisa é chamado Ae.

O projeto é uma parte do projeto geral do TIDIA, financiado pela FAPESP, e tem associações com as instituições internacionais *IMS — Global Learning Consortium* e *Sakai Foudation*, que visam discutir as aplicações da tecnologia no aprendizado e seus resultados. O Ae visa beneficiar instituições de ensino, empresas e fundações em suas atividades educacionais, possibilitando a expansão do alcance do aprendizado eletrônico[12].

4.1.2 SAKAI

Definido como um Ambiente de Colaboração e Aprendizado (*Collaboration and Learning Environment*) por seus desenvolvedores, o Sakai é uma ferramenta de código aberto (*Open Source*) que engloba diversas funções voltadas para o ensino, o aprendizado e a pesquisa colaborativa.

Usado em mais de 200 universidades, faculdades e escolas ao redor do mundo [15], sua natureza de código aberto permite que as inúmeras alterações e adaptações feitas por cada entidade possam ser acrescentadas ao projeto original, ampliando suas capacidades. Seus componentes são modulares, permitindo aos usuários optar por aqueles que melhor se ajustam às suas necessidades. O Sakai tem ampla escalabilidade, podendo manter mais de 200.000 instalações simultâneas e até 20.000 usuários simultâneos, ainda que esse nível de capacidade seja bastante dependente de detalhes de implementação e padrões de uso[15].

Suas ferramentas base se dividem em Ferramentas de Colaboração Geral com suporte a anúncios, notícias, fóruns, *blogs* e outras funcionalidades similares; Ferramentas de Ensino e Aprendizado que focam na construção e disponibilização de cursos e testes; Ferramentas de Portifólio permitindo construir e partilhar portifólios entre os usuários e, finalmente, as Ferramentas Administrativas que permitem o controle dos aspectos gerais do sistema.

O SAKAI figura neste trabalho apenas como a base sobre a qual o TIDIA-Ae foi construído. Apesar de todo o trabalho ter sido desenvolvido com interesse único no TIDIA-Ae, ele pode ser adaptado e expandido para o SAKAI, dados os esforços de adequação necessários.

4.1.3 As Razões da Escolha

Em instância mais ampla, o TIDIA-Ae é uma opção vantajosa por ser um sistema em amplo desenvolvimento no país, assim como um derivado de um sistema desenvolvido e utilizado em escala mundial sem se distanciar muito do mesmo. Estas características dão a este trabalho o prospecto de poder ser expandido e integrado futuramente a estes sistemas, dadas as devidas adaptações.

Outro aspecto importante é que o TIDIA-Ae desenvolvido localmente oferece acesso rápido e direto às equipes de desenvolvimento e manutenção do mesmo, permitindo agilidade nas eventuais necessidades de interação com elas. Caracteriza-se também como um facilitador para uma eventual integração e adaptação do protótipo ao sistema.

O TIDIA-Ae também é utilizado na UNICAMP como um sistema auxiliar para cursos de aulas presenciais. Neste caso, é utilizado primariamente para dispor documentos para os alunos ou ligações para materiais externos. Embora esse tipo de uso para um sistema de educação online não seja raro, é suficiente para permitir uma análise da metodologia.

4.2 Compreensão do Sistema

A abordagem adotada, de mínima interferência com o sistema original para permitir a mineração de dados, limita as possibilidades de obtenção de dados, mas tem a vantagem de oferecer acesso imediato a dados já existentes. Apesar disso, corre-se o risco de esses dados serem insuficientes, além dos consequentes problemas de formatação e limpeza que demandam muitas vezes um extenso trabalho. Existe também a possibilidade de não existirem registros prévios, isto é, o sistema pode estar em início de uso ou simplesmente não possuir forma alguma de guardar registros de uso. Neste último caso, uma abordagem de mínima interferência pode não ser possível.

O passo inicial, compreensão do sistema, avaliou as possíveis fontes de dados no ambiente. Os registros de acesso dos usuários, mantidos pelos servidores de aplicação, geralmente são fontes significativas de informações. Diversos pesquisadores recorrem aos mesmos para obtenção de dados [11, 27, 33], algumas vezes usando ferramentas específicas para processá-los tais como WebSIFT [6] e o WebLogMiner [34]. Ambas as ferramentas são capazes de aplicar rotinas de mineração de dados e descobrir uma variedade de padrões a partir de *web logs*. No entanto, esses registros proveem apenas um traçado rudimentar das navegações e atividades do aprendiz no ambiente; uma etapa significa-

tiva de limpeza e transformação deve ser realizada para que a informação possa ser utilizada pelos algoritmos de mineração de dados [32].

Para o estudo inicial foram obtidos os registros dos acessos ao servidor de aplicação (Tomcat) para cursos do segundo semestre de 2008. Essa fonte de dados limita a qualidade da informação obtida, pois desconsidera a estrutura interna das ferramentas do ambiente. Adicionalmente, os acessos dos usuários são omitidos dos registros devido a configurações específicas do sistema realizadas pelos administradores, pois são dados de pouca valia para eles e sua presença dificulta a identificação e solução de problemas na execução do sistema.

Assim, outras fontes de dados foram procuradas dentre as estruturas internas da plataforma Sakai, que possui mais de duzentas tabelas em sua base de dados interna. O TIDIA-Ae acrescenta outras tabelas a essas, caracterizadas pela presença do prefixo tidia na composição do nome. As tabelas (Fig. 4.1) visam descrever e complementar aspectos e ferramentas diferentes do ambiente. Há o registro de eventos relacionados a acessos, mantido na tabela SAKAI_EVENT. Outras tabelas foram utilizadas para complementar os dados obtidos, como SAKAI_SESSION, que mantém detalhes referentes a cada sessão de uso do sistema, e SAKAI_USER_ID_MAP, que relaciona os identificadores internos dos usuários usados nas mais diversas tabelas com a base de dados de usuários SAKAI_USER.

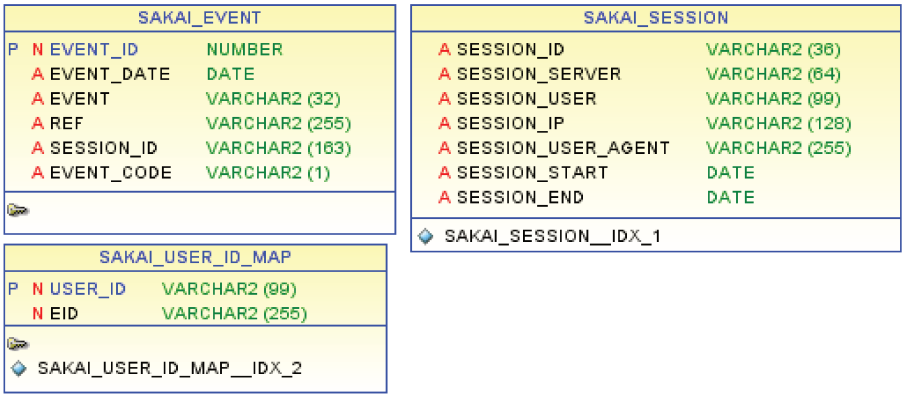

Fig. 4.1: As tabelas replicadas do Tidia-Ae/Sakai

Tal qual a base original, foram replicadas as estruturas das tabelas com seus respectivos dados em uma base de dados usando o gerenciador de banco de dados *MySQL*. A versão usada aqui é a 5.1.31.

As tabelas selecionadas correspondem a apenas uma parte do modelo conceitual apresentado na seção 3.1 do capítulo 3. A tabela SAKAI_EVENT, a única dentre as exploradas que se relaciona diretamente com o modelo, representa a classe **Evento**, agregando informações a respeito de cada atividade realizada dentro do ambiente do SGA. Embora SAKAI_USER seja equivalente à classe **Usuario** do modelo, ela não foi utilizada neste estudo. SAKAI_SESSION teve aplicação significativa no estudo realizado, mas não possui uma relação direta com o modelo apresentado. Isso porque

a tabela pode ser definida como um registro de tempo de utilização contínua do Sistema, agregando basicamente a uma ligação direta entre um grupo seqüencial de eventos descritos pela tabela SAKAI_EVENT.

Dada a grande quantidade de tabelas na base de dados do TIDIA-Ae e também as restrições de tempo, um estudo mais detalhado dos conteúdos e funções de todas elas não foi possível, assim não foram identificadas e trabalhadas tabelas que representassem as demais partes do modelo conceitual. A inclusão destas tabelas agregaria uma maior capacidade de descrição dos dados analisados ao fim do estudo, mas tornariam todo o processo de interpretação dos resultados mais custoso.

Selecionadas as tabelas da base de dados do sistema a serem utilizadas como fonte de dados, conclui-se o primeiro passo da metodologia. A seção seguinte mostrará a realização do segundo passo, **Determinação dos Objetivos da Mineração de Dados**, descrevendo a forma como foram tratados os dados e definidas as metas para este estudo e também o terceiro passo, **Preparação**.

4.3 Determinação dos Objetivos e Preparo dos Dados

O preparo dos dados pra a mineração é diretamente influenciado pelos objetivos determinados pelos pesquisadores. Visando demonstrar este aspecto, decidiu-se por apresentar ambos os passos, Determinação dos Objetivos e Preparação, em conjunto nesta seção.

Para o estudo de caso, decidiu-se explorar os dados disponíveis estabelecendo metas mais comuns a trabalhos de mineração de dados de análise do uso do sistema. Buscou-se levar em consideração as características de apoio à aula presencial do TIDIA-Ae na UNICAMP, de modo que a análise final tenha alguma valia para ele e outros SGAs que sejam empregados de forma similar.

4.3.1 Sumário de Dados de Uso

As métricas mais comumente aplicadas para analisar a utilização desse tipo de sistema envolvem o quanto o sistema é usado e acessado pelos usuários, em particular pelos alunos. Trabalhos como os de Castro et al.[5] e Hsu et al.[17], por exemplo, usam algumas métricas similares. Elas visam identificar padrões de uso dos sistemas pelos alunos, alunos com comportamento irregular, ou sobre indivíduos que ignoram o sistema quase completamente, alertando o tutor sobre a eventual necessidade de tomar ações preventivas. Portanto, para este estudo foi estabelecido o objetivo de identificar grupos de usuários conforme o tempo que permanecem logados no sistema e a quantidade de sessões de uso do sistema.

No ambiente estudado, essa informação é obtida da tabela SAKAI_SESSION, com dados relativos à quantidade de sessões e tempo total de acesso do sistema para cada usuário. Cada usuário é

definido por um registro na tabela SAKAI_USER_ID_MAP. Para conter os sumários de dados, a base de dados analítica DATA_SUMMARY foi criada; a tabela USER_SUMMARY dessa base mantém os dados processados sobre as sessões. A Fig. 4.2 representa esta tabela.

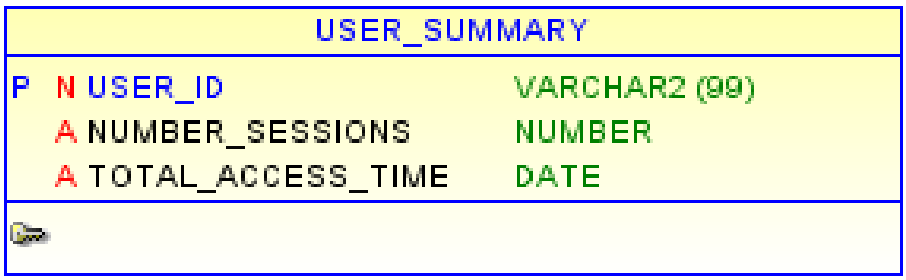

Fig. 4.2: A tabela USER_SUMMARY

Estabelecido o objetivo de identificação de grupos de usuários, os dados foram organizados de forma a destacar as informações relevantes para a análise desejada. Assim, esses dados foram organizados como vetores de duas dimensões: a primeira consiste no número total de sessões de um usuário e a segunda no tempo total de uso do sistema por ele, calculado a partir da soma aritmética dos tempos de cada sessão.

Tendo em vista algumas das métricas usadas por Castro et al. [5], que incluem a duração média das sessões de um usuário e a média de *hits* por sessão do usuário, aqui na forma da média de eventos, foi estabelecida uma segunda organização dos dados agregando informações relativas a cada sessão, com sua duração e a quantidade de eventos registrada assim como o usuário responsável. Ao longo do estudo, no entanto, esta abordagem foi descartada devido a restrições de tempo e também limitações na capacidade descritiva dos dados disponíveis.

4.3.2 Sumário de Acesso a Documentos e Mídia

A natureza do uso do ambiente neste estudo denota um forte uso do mesmo como repositório de documentos e mídias. Para o estudo foram analisados os dados relacionados à disponibilização e acesso a documentos dos tipos mais comuns no sistema, imagens (JPG, GIF, BMP), textos (PDF, DOC), áudio (MP3) e apresentações (PPT).

Foi estabelecido o objetivo de agrupar estes documentos de forma a descobrir relações de relevância entre eles e os conteúdos e formatos dos mesmos. Isto é, determinar quais documentos são mais acessados, quais os tipos desses documentos e quais conteúdos que eles apresentam e se estas características se mostram similares com os demais documentos pertencentes aos mesmos grupos.

Os dados foram preparados em duas etapas. Primeiro, duas tabelas foram construídas com dados extraídos da tabela de eventos do ambiente (Fig. 4.3). A primeira, DOCUMENT_UPLOAD_SUMMARY,

Fig. 4.3: As tabelas que sumarizam as interações com documentos

contém informações a respeito de eventos de *upload* de novos documentos no sistema, associados aos eventos *content.new*. A segunda, DOCUMENT_READ_SUMMARY, é similar, mas descreve eventos de leitura, revisão e remoção desses documentos. Estas tabelas foram estabelecidas para concentrar somente informações relevantes relacionadas aos documentos do sistema e as possíveis interações com os mesmos. Elas não são tão úteis ao processo de mineração em si, mas servem como um ponto de partida mais prático para o preparo dos dados para a mineração, uma vez que limitam os dados àqueles relacionados a documentos.

A segunda etapa do preparo dos dados, foi iniciada definindo uma organização vetorial de dimensão variável. Ela contém as diferenças de tempo entre o momento de inclusão do documento no sistema e cada acesso subseqüente. O tamanho do vetor varia por depender da máxima quantidade de acessos registrada no sistema. Devido à necessidade de equalizar estas dimensões, os valores ausentes foram preenchidos com nulos. Esta abordagem foi estabelecida buscando considerar o máximo possível de detalhes disponíveis a respeito dos acessos aos documentos na mineração de dados.

Uma segunda abordagem, mais contida, registra, para cada documento, três dados: a quantidade total de acessos, a diferença entre o momento de inclusão e o primeiro acesso, e a diferença de tempo entre o primeiro e o último acesso. Esta, foi elaborada devido ao temor de que as amplas diferenças dimensionais da primeira e a subseqüente equalização através de nulos gerasse resultados distorcidos e insatisfatórios.

Quando um usuário tem falhas na conexão ao tentar acessar uma página ou documento, é comum que ele tente atualizar a página do navegador consecutivas vezes até que a página seja carregada. Essa atitude resulta em diversas requisições ao sistema e uma parte significativa dos acessos de leitura dos documentos é originada dessa forma, como observado pelo curto e não raro nulo intervalo de tempo registrado entre os vários eventos de leitura de um documento em uma mesma sessão. Uma limpeza realizada foi considerar um único acesso de leitura a um mesmo documento por sessão para eliminar a influência deste problema.

4.4 Mineração de Dados

Para alcançar os objetivos propostos na sessão anterior, é necessária a aplicação de alguma técnica de mineração de dados. Tendo estabelecido objetivos focados no agrupamento dos dados disponíveis, é selecionado nesta etapa a forma como estes objetivos serão abordados e que meios para alcançá-los serão utilizados. Assim, apresentam-se os algoritmos selecionados para processar os dados e as razões de sua escolha, além das particularidades inerentes as implementações utilizadas dos mesmos. Também nesta seção é feita a apresentação de uma abordagem alternativa, consolidando dados relativos a documentos em grupos através de critérios simples ao invés do uso de sofisticados algoritmos de mineração de dados.

4.4.1 Algoritmos de agrupamento

O primeiro algoritmo de agrupamento selecionado para aplicação neste estudo foi o K-means. Ele é um dos mais conhecidos métodos de agrupamento particional e também é um dos conceitualmente mais simples. Estes fatores influenciaram bastante em sua escolha, uma vez que a baixa complexidade dos vetores de dados preparados para a mineração de dados não exige o uso de algoritmos complexos para serem processados. Sua popularidade também o torna um bom ponto de partida para o estudo, visto que oferece uma base de comparação mais comum para outros trabalhos. O algoritmo utilizado foi adaptado conforme as necessidades do trabalho realizado nesta dissertação a partir de uma das versões apresentadas por Scarberry [26] que utliza o K-means para apresentar uma comparação entre aplicações de processamento simples e aplicações de processamento concorrente.

O segundo algoritmo utilizado foi o Mapa Auto-Organizável, que possui propriedades de redução de dimensionalidade minimizando a perda da topologia do espaço original. Como apresentado anteriormente na seção 4.3.2, um dos preparos realizados dos dados relativos a documentos e mídia previu uma grande dimensionalidade de vetores, o que tornou o Mapa Auto-Organizável uma escolha interessante para o processamento. A facilidade de acesso a ferramentas para aplicação deste algoritmo também foram um fator considerado na escolha do algoritmo: foi utilizado o *SOM Toolbox* [2], um conjunto de ferramentas desenvolvido para o ambiente MATLAB e recomendadas por Kohonen [20].

K-Means aplica um cálculo de distância euclidiana simples entre os vetores e os centros para identificação de grupos e tem como critério de parada a não ocorrência de deslocamentos dos centros entre uma interação e outra ou o alcance de um número limite de iterações. Como parâmetros de entrada são fornecidos ao algoritmo os vetores de dados normalizados, a quantidade máxima de grupos desejada, o número máximo de iterações e um valor base para seleção aleatória de coordenadas como centros iniciais para os grupos.

O *SOM Toolbox* não requer muitos ajustes para a execução do Mapa Auto-Organizável. Experimentos foram executados contemplando diferentes tamanhos de mapas. O treinamento em cada execução é feito com uma parcela aleatória diferente dos dados de entrada, definindo um conjunto de células em um plano representando os dados de entrada. A seguir, o *SOM Toolbox* estabelece grupos de células através de um mecanismo interno onde ele realiza a aplicação de uma sequência de execuções de um algoritmo K-Means próprio e apresentando apenas o melhor resultado ao final. A ferramenta define a quantidade máxima de grupos que poderão ser identificados através da expressão: $\sqrt{dlen}$, onde *dlen* representa quantidade de vetores de entrada, isto é, a quantidade de linhas da matriz de entrada.

A seção a seguir apresenta uma aproximação alternativa para o tratamento dos dados referentes a documentos, cujo objetivo é explorar algumas possibilidades de organização e visualização dos dados sem requerer o uso de algoritmos sofisticados de mineração de dados.

4.4.2 Consolidação Alternativa de Dados

A organização e visualização de informações significativas que possam encontrar-se espalhadas em meio aos dados disponíveis também é um objetivo comum da Descoberta de Conhecimento em Bases de Dados. Desta maneira, optou-se por apresentar uma alternativa à aplicação dos algoritmos de mineração de dados. Aproveitando-se dos dados de acessos a documentos e mídia, criou-se um sistema de relatórios de acessos a esses arquivos pelos usuários para informar tanto professor quanto aluno da forma como estes documentos estão sendo acessados.

Inicialmente o trabalho foi desenvolvido apoiando-se em um sumário de dados particular construído de forma similar aos sumários listados anteriormente. Buscando uma maior integração com o sistema e transparência, foram utilizadas as suas bases de dados originais, como descrito a seguir.

Para o sumário para o professor, a partir da identificação do usuário, uma série de consultas automáticas é realizada à base de dados buscando todos os documentos que o usuário colocou no sistema e informações de quando a inclusão foi realizada, o total de acessos realizado a ele, quantos usuários diferentes realizaram esses acessos e quando foi acessado pela última vez.

O sumário para o aluno, de maneira semelhante, busca dados referentes às leituras de documentos por parte do usuário indicado, construindo uma lista com os documentos acessados, quantidade de acessos e os momentos do primeiro e último acessos de cada um.

Neste estudo foram obtidas duas saídas para cada caso: um arquivo **XML**, visando a integração com outras aplicações, e uma visualização desses dados em HTML direcionada ao educador e ao aluno.

4.5 Interpretação dos Dados encontrados no Ambiente

No decorrer do desenvolvimento deste trabalho observou-se falhas na capacidade de descrição dos dados, devido à característica do sistema como uma ferramenta de apoio às aulas presenciais. Este aspecto cria uma não obrigatoriedade do uso do sistema pelo aluno, pois ele pode apoiar-se de maneira compensatória em colegas ou até mesmo no próprio professor.

Assim, um aluno que não acessa o sistema pode não estar necessariamente defasado em relação aos que o usam, dado que ele tenha interesse suficiente para procurar outras formas ou fontes de acesso às informações dispostas pelo professor através da ferramenta. Documentos pouco acessados de um professor, possivelmente são copiados e distribuídos por uns poucos alunos a toda a turma. De fato, é provável que estes comportamentos se tornem ainda mais notáveis em cursos pouco ligados à tecnologia de informação, onde o hábito de se utilizar de recursos *online* ou mesmo de computadores possa não existir em uma grande parcela dos alunos.

Estas condições do estudo limitam a qualidade das conclusões obtidas sob o ponto de vista pedagógico, pois os dados obtidos referem-se apenas ao uso do ambiente como complemento a aulas presenciais. As seções seguintes apresentam os resultados obtidos através dos experimentos realizados com os dados disponíveis, apresentando interpretações dos resultados obtidos com a mineração de dados.

4.6 Qualificação dos Resultados dos Agrupadores

Inicialmente são analisados os resultados obtidos com as execuções dos algoritmos de agrupamento. São examinados os resultados para o conjunto de dados de uso e em seguida para os dados de acesso a documentos. Em ambos os casos, diferentes configurações foram usadas para K-Means, com os resultados gravados em arquivos XML para averiguação. Para importação dos dados no ambiente MATLAB, para treinamento e processamento com mapas auto-organizáveis, a entrada foi realizada com arquivos com valores delimitados por vírgulas (CSV) e os resultados gravados como imagens e como arquivo texto com a relação de grupos, células e membros, para facilitar a comparação com resultados obtidos pelo K-Means.

4.6.1 Resultados para os Dados de Uso

Ao traçar os dados de uso disponíveis sobre um plano cartesiano (Fig. 4.4), torna-se possível determinar grupos de forma visual. De fato, com algum esforço para determinar, de forma manual, limiares que definam grupos específicos, pode-se agrupar os usuários sem a necessidade de um algoritmo sofisticado de mineração de dados. Apesar disto, decidiu-se por insistir na aplicação dos

algoritmos escolhidos sobre os dados disponíveis, no intuito de observar seu funcionamento e quais resultados eles podem gerar com esses dados.

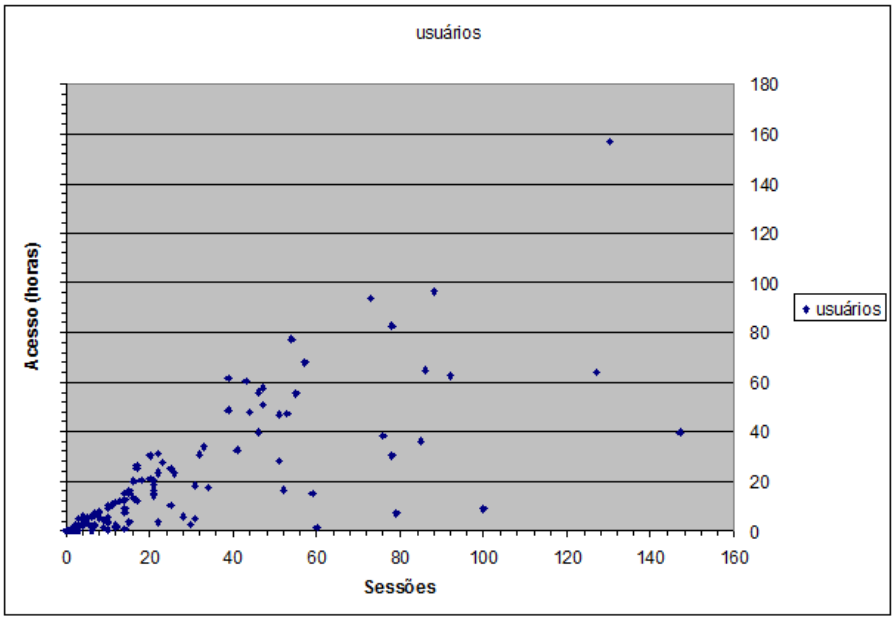

Fig. 4.4: Os dados de uso dispostos no plano cartesiano.

Uma dificuldade de configuração do K-means é estimar a quantidade adequada de grupos. Assim, o algoritmo foi executado diversas vezes observando o número máximo de clusters diferentes de 6, 8, 10 ou 12, tentando encontrar um resultado que se adequasse melhor à quantidade de vetores de dados disponíveis.

Como mencionado anteriormente na seção 4.4.1, a implementação utilizada do K-means tem como parâmetros de entrada os vetores de dados normalizados, a quantidade máxima de grupos desejada, o número máximo de iterações e um valor base para seleção aleatória de coordenadas como centros iniciais para os grupos. Esse último parâmetro, baseado na geração pseudo-aleatória de valores oferecida pela linguagem de programação JAVA, não se revelou como grande influenciador dos resultados, pois variações de seus valores afetam o resultado significativamente apenas quando muito grandes, com diferenças de várias ordens de grandeza. Valores muito baixos geralmente levam à identificação de um número de grupos menor que o máximo proposto ao algoritmo, pois os centros acabam sendo inicializados muito próximos uns dos outros, motivo pelo qual foi adotado um valor suficientemente alto para minimizar a influência dessa variável nos resultados. Assim, seu valor foi fixado, após alguma experimentação, em 10000.

A característica determinante definida pelos dados disponíveis é o interesse do usuário no sistema e no conteúdo disponbilizado pelo professor através dele. Um usuário que acessa o sistema poucas vezes no semestre, certamente tem um interesse insuficiente no mesmo e pode ter seu desempenho

no curso comparado ao interesse demonstrado. Se ele tiver um bons resultados, pode significar que os recursos do sistema estão sendo mal utilizados pelo educador e portanto a parcela *online* do curso pode precisar de uma revisão para agregar mais conteúdo e relevância a ela. Se ele tiver maus resultados, o curso *online* está oferecendo material essencial adequadamente, mas está falhando em atrair o aluno a utilizá-lo em seu aprendizado. Outros padrões de comportamento observáveis são, por exemplo, alunos que acessam frequentemente o sistema por breves períodos de tempo podem ter interesse elevado no sistema, mas não estão encontrando novos conteúdos para explorar. Alunos que, por outro lado, acessam poucas vezes o sistema, mas passam muito tempo utilizando-o a cada acesso provavelmente o acessam apenas quando notificados diretamente pelo professor da disponibilidade de novos conteúdos.

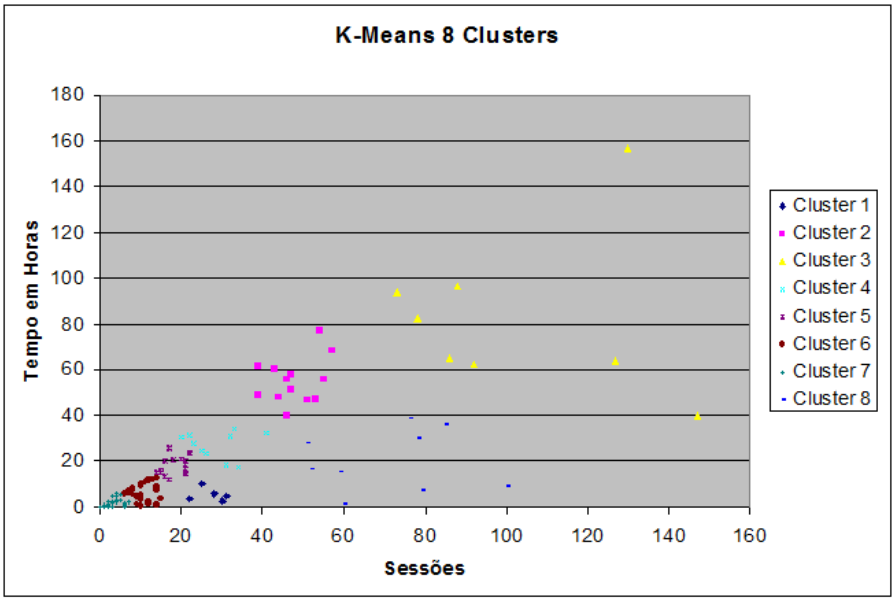

Fig. 4.5: Os resultados obtidos pelo K-Means ajustado para um máximo de 8 grupos.

A Fig. 4.5 apresenta os grupos identificados pelo K-Means, ajustado para oito grupos, onde é possível identificar diferentes grupos de interesse por parte dos usuários. Uma quantidade significativa de usuários concentra-se abaixo do quadrado definido entre 40 horas e 40 sessões e o algoritmo concentrou, portanto, um número razoável de grupos nesta região. Note que alguns desses usuários, particularmente os do grupo 6, excedem em muito os limiares propostos, o que pode requerer a atenção especial do educador para entender as razões desse uso demasiado.

Sendo o TIDIA-Ae utilizado como apoio a aulas presenciais e uma plataforma de distribuição e compartilhamento de documentos e mídia, considerou-se para este estudo que a região delimitada entre 20 e 40 horas de uso e entre 20 e 40 sessões de acesso como o comportamento padrão desejável dos alunos. Neste escopo, o K-means permitiu flexibilizar esses limiares criando grupos que por não

se encaixarem totalmente dentro deles. Esses grupos podem ajudar a qualificar mais adequadamente os comportamentos individuais dos alunos conforme a similaridade de seus acessos com os demais membros do grupo a qual pertencem e não tanto através dos limiares estabelecidos.

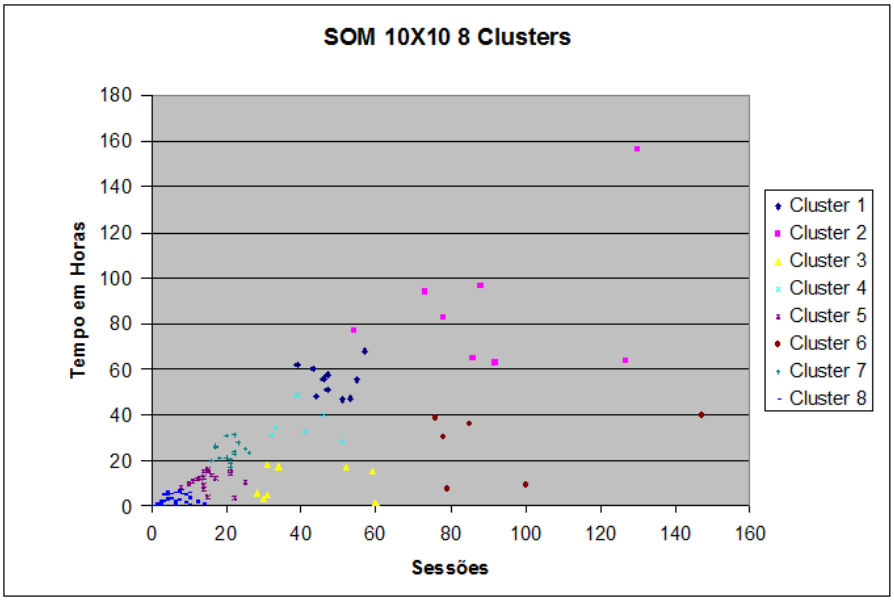

Fig. 4.6: Os resultados obtidos pelo Mapa Auto-Organizável.

A Fig. 4.6 apresenta um resultado do processamento dos dados através dos mapas auto-organizáveis. Os resultados se mostraram mais variáveis ao longo de diversas execuções que o K-Means. Após o treinamento do mapa e distribuição dos dados entre as várias células do mesmo, o algoritmo aplica diversas vezes uma implementação interna à ferramenta do K-Means sobre o mapa. São usados diversos valores para o número de grupos, retornando o melhor resultado para cada caso com base na soma dos erros quadráticos, e o índice Davies-Bouldin é calculado para cada um desses resultados. O melhor resultado é apresentado pela ferramenta como resultado final [3].

A cada execução do algoritmo através do SOM Toolbox, o treinamento foi realizado novamente antes do mapeamento. Este fator pode representar a razão pela grande variedade de resultados diferentes obtidos, indo de resultados com três grupos até resultados com doze grupos. Como especificado anteriormente na seção 4.4.1, a quantidade máxima de grupos que poderão ser identificados é definida através da expressão: $\sqrt{dlen}$, onde $dlen$ representa quantidade de vetores de entrada. No caso deste estudo, $dlen = 140$ e em grande parte das execuções os resultados apresentavam quantidades de grupos próximas do valor definido pela raiz quadrada de $dlen$, isto é, aproximadamente 12 grupos. Para comparação, é apresentado o resultado com oito grupos. Este resultado, embora possua uma linha geral de grupos similar ao do K-Means, é diferente o suficiente para possuir grupos exclusivos.

Como dito anteriormente, a característica determinante definida pelos dados disponíveis é o in-

teresse do usuário no sistema e no conteúdo disponbilizado pelo professor através dele. O estudo permitiu formalizar uma maneira automatizada de definir os grupos de interesse sem se prender a critérios rígidos de agrupamento, isto é, a imposição empírica de limiares para o agrupamento. Para esta função, ambos algoritmos propostos se portaram de forma similar e eficaz, embora o K-means tenha tido uma gama de resultados mais uniforme.

Por serem os dados referentes a todos os usuários do sistema dentro de um período de um semestre sem distinção de curso ou função (isto é, aluno ou professor), os resultados aqui obtidos são bastante limitados. Tendo os cursos como critério para separar os usuários, seria possível determinar grupos de comportamentos específicos para cada curso e permitir a comparação entre eles ou permitir aos educadores qualificarem o interesse de seus alunos como indivíduos ou grupos no sistema e nas facilidades educacionais oferecidas por ele.

4.6.2 Resultados para o Acesso a Documentos e Mídia

Para os trabalhos de processamento dos dados relativos ao acesso a documentos e arquivos de mídia, foram utilizadas as mesmas configurações iniciais dos agrupadores aplicadas aos dados de uso. Ao longo do desenvolvimento, poucos ajustes se revelaram necessários além dos relativos às adequações aos formatos dos dados aplicados.

A análise do resultado com os dois tipos de sumários criados na seção 4.3.2 levou à adoção do segundo tipo. A dimensionalidade elevada do primeiro tipo de sumário, estabelecida pelos poucos documentos significativamente acessados (mais precisamente 55 dimensões) somada à grande quantidade de documentos pouco ou mesmo nunca acessados, gera uma grande quantidade de nulos acrescentados artificialmente para ajustar a dimensão, o que tornou os resultados pouco informativos.

Verificou-se ao longo do estudo que o nome de cada documento na tabela de eventos original consiste no caminho completo de diretórios até o documento, o que permitiria filtrar os documentos por curso e também eliminar a grande quantidade de documentos particulares de usuários para realizar o processamento. Para aproveitar-se dessas informações seria necessária uma filtragem baseada em texto. O custo de tempo para a construção e ajuste adequado deste filtro fez com que tal possibilidade não fosse utilizada.

Do grupo de documentos inicial identificados, 52 documentos foram eliminados da análise por nunca terem sido acessados. A Fig. 4.7 e a Fig. 4.8 apresentam os resultados obtidos para as duas opções de agrupamento.

Como com os dados de uso, as principais diferenças entre os resultados aqui obtidos estão nas áreas de concentração maior de pontos (documentos). A mais notável particularidade aqui é que os diferentes documentos de cada curso têm acessos bastante distintos, enquadrando-se em grupos diferenciados conforme a importância dada a eles pelos alunos, isto é, o quanto e quão frequentemente

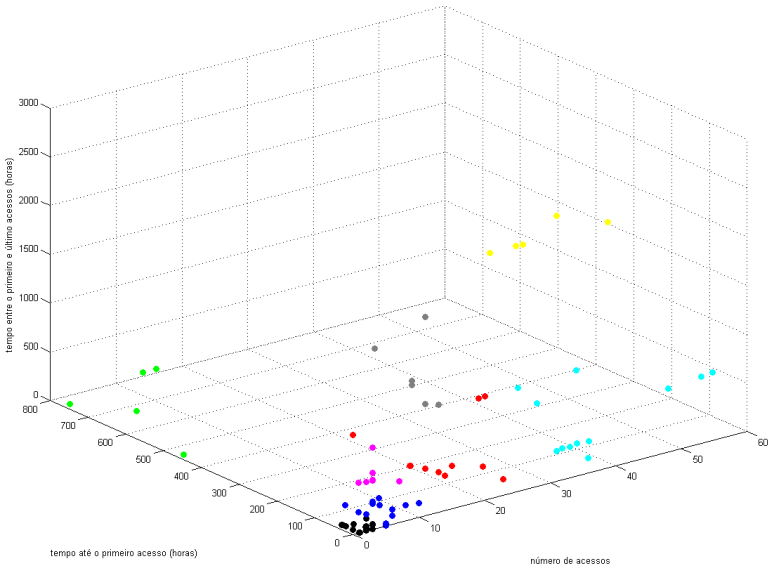

Fig. 4.7: Os resultados obtidos pelo K-Means com os dados referentes aos documentos.

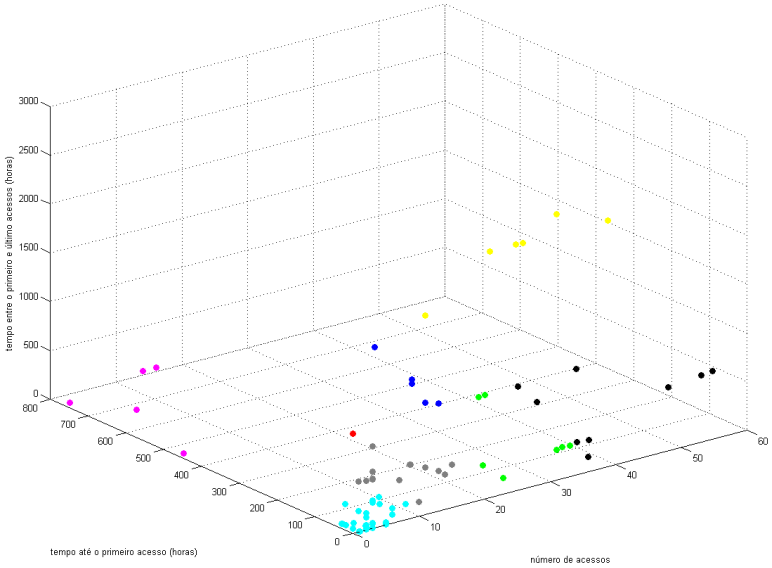

Fig. 4.8: Os resultados obtidos pelo Mapa Auto-Organizável com os dados referentes aos documentos.

foram acessados. Uma particularidade observada para este caso é que os grupos definidos refletem, em diversos casos, os tipos de documento e suas funções: arquivos de áudio de um mesmo curso têm quantidade de acessos similares; arquivos de texto com apostilas têm média de acessos distinta de

gabaritos e arquivos relacionados a avaliações. Desta forma, o resultado pode ajudar o professor a identificar documentos que não têm o padrão de acesso esperado, permitindo identificar recursos que não recebem a atenção devida pelos alunos e investigar potenciais causas para esse descaso.

4.7 Resultados da Consolidação Alternativa de Dados

Para a consolidação alternativa dos dados, o professor e o aluno podem visualizar o quadro de acessos a documentos do ambiente. O mecanismo resultante é eficiente, derivado de poucas tabelas como fontes de dados para construir essa representação sucinta do quadro de acessos aos documentos tanto para o professor quanto para o aluno. Para o caso do professor, pode ajudar a identificar alunos que não tenham obtido documentos relevantes ou documentos que não tenham, por alguma razão, atraído a atenção dos alunos da forma esperada. O relatório individual para o aluno, embora cumpra bem o objetivo de apresentar-lhe os documentos que ele acessou, servindo-lhe como histórico de fácil acesso pode, em futuras implementações que se apóiem em uma parte maior da estrutura de dados do sistema, ajudá-lo também a identificar os documentos relativos aos cursos de que ele participa que ele possa não ter acessado, evitando que os mesmos lhe escapem à atenção. Esta ferramenta em particular tem bom potencial para integração ao ambiente, tendo utilidade ainda maior em cursos não presenciais.

O processo desta consolidação origina dois arquivos, como mencionado na seção 4.4.2: um arquivo em XML contendo os dados em um formato que facilita um eventual uso por outras aplicações e uma página em HTML, formatada em grande parte por meio de um arquivo CSS, contendo as informações extraídas organizadas de forma a facilitar o entendimento para o educador ou mesmo para o aluno.

A Fig. 4.9 apresenta o relatório de acessos a documentos realizados por um determinado aluno no formato em que é apresentado a ele.

De maneira similar, a Fig. 4.10 apresenta o relatório geral de acessos a documentos apresentado ao professor.

As imagens foram obtidas utilizando o *Internet Explorer 8*, mas independem de um navegador específico para funcionar.

4.8 Considerações Finais

Neste capítulo foi realizado um estudo de caso da metodologia proposta no capítulo anterior sobre o ambiente educacional TIDIA-Ae. Todos os passos descritos para ela foram realizados em etapas separadas e os problemas observados foram discutidos e as soluções possíveis implementadas. Foram

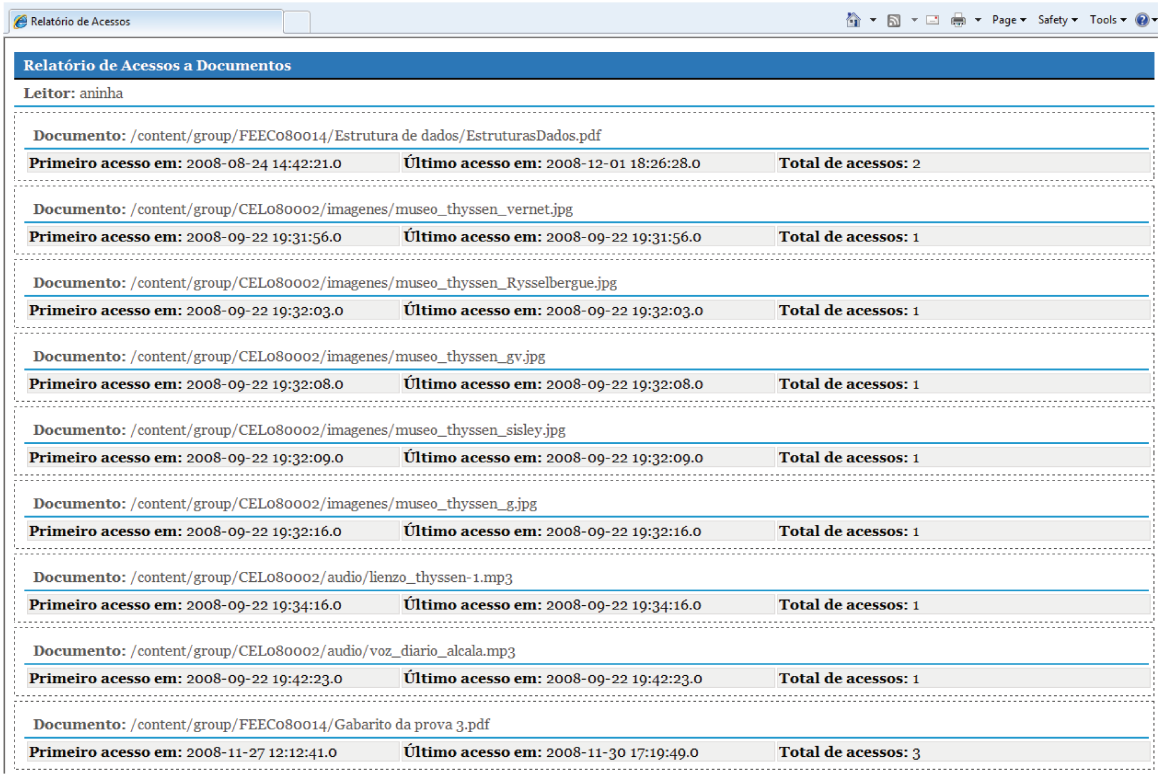

Fig. 4.9: Exemplo do resultado produzido por meio dos dados de acesso de um aluno.

comentadas as limitações encontradas nos dados extraídos do ambiente e suas causas. Os resultados do estudo de caso foram apresentados e comentados, suas potenciais aplicações em ambientes reais de ensino destacadas.

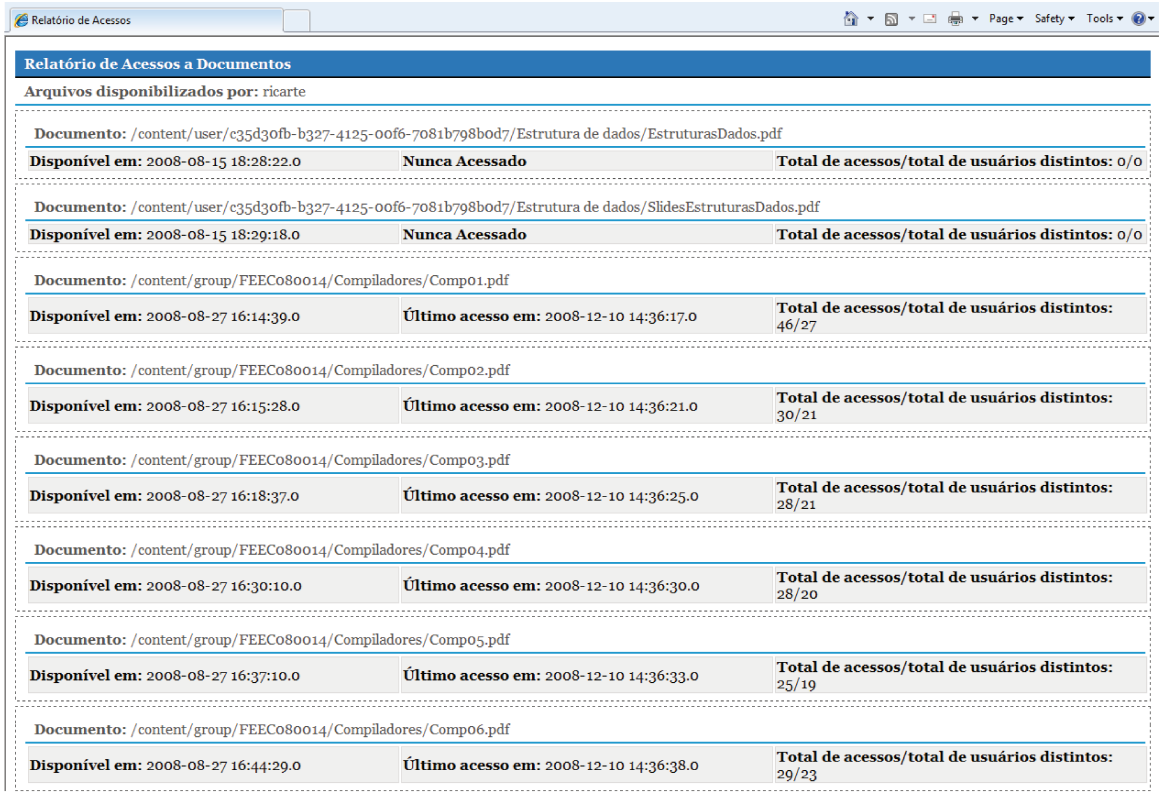

Fig. 4.10: Exemplo do resultado produzido por meio dos dados de acesso aos documentos disponibilizados por um professor.

Capítulo 5

Conclusão

A crescente disponibilidade de computadores e acesso à Internet tornou o ensino a distância através de sistemas educacionais *online* uma atraente e cada vez mais explorada opção. A ausência de contato direto entre professor e aluno torna a percepção, qualificação e adequação dos materiais dos cursos difícil. Isso ocorre porque o processo de obtenção das informações necessárias para estas tarefas torna-se complexo e custoso.

No entanto, dados capazes de oferecer um retorno apropriado ao educador podem existir dentro destes sistemas educacionais. Mas o tempo e complexidade envolvidos na extração e análise destes dados torna esta opção impraticável para o educador, que freqüentemente é responsável por uma grande quantidade de cursos e alunos. No mais, poucos sistemas educacionais oferecem ferramentas adequadas para a realização deste tipo de trabalho.

Faz-se necessário, portanto, um meio de analisar esses dados e tentar obter deles as informações necessárias ao educador. Técnicas de mineração de dados são uma opção viável e frequentemente aplicada para esta tarefa. O trabalho realizado nesta dissertação buscou formalizar uma metodologia capaz de orientar um pesquisador pelos passos necessários para aplicação destas técnicas. Passos estes que envolvem identificação, seleção, preparo, processamento e interpretação dos dados.

Um estudo de caso foi realizado apoiando-se em um sistema de ensino a distância, o TIDIA-Ae, sendo a instância analisada utilizada como apoio a aulas presenciais. Este fato somado às restrições quanto à quantidade de dados extraídos do ambiente educacional resultou em particularidades que dificultaram a demonstração da metodologia, embora em momento algum a tenham invalidado.

Com o uso da metodologia foi possível examinar e delimitar etapas distintas do processo necessário para a realização da mineração de dados sobre os dados de um ambiente educacional *online*. Ela permitiu organizar o fluxo do trabalho realizado e identificar os problemas que ocorreram em cada etapa permitindo resolvê-los no ponto em que oferecem maior interferência no desenvolvimento do trabalho. A metodologia não resolve os problemas em si, tampouco evita aqueles oriundos particu-

larmente das informações disponíveis. Isto ficou evidente nos pontos em que os dados se revelaram insuficientes, incompletos ou inadequados para a realização dos processos propostos de forma que gerassem resultados significativos para obter-se um retorno para os educadores.

Espera-se de um ambiente de aprendizado *online* dados capazes de descrever em variados níveis de detalhamento os hábitos de uso do ambiente pelos alunos e educadores. Materiais acessados, sequências de páginas e conteúdos vistos; materiais depositados no sistema, tanto pelos professores quanto alunos; resultados de testes, frequência de acessos, tempo gasto em cada conteúdo, concentrações de acessos em períodos determinados de tempo, todos são informações plausíveis de serem encontradas nos registros armazenados de um bom sistema educacional. Embora o TIDIA-Ae utilizado pela UNICAMP esteja preparado para coletar e manter estes dados de maneira organizada, a natureza do uso do sistema como auxiliar às aulas presenciais retira do aluno a obrigatoriedade de utilizar-se do sistema em seu aprendizado.

A baixa dimensionalidade dos dados analisados exclui a necessidade de métodos sofisticados de classificação. Os dados podem ser facilmente separados de forma visual uma vez traçados em planos bi ou tridimensionais. No entanto, um estudo que seja realizado envolvendo uma quantidade maior de dados, mesclando no conjunto dados de outras tabelas do TIDIA-Ae e resultando em entradas de dados de maior dimensão, poderá se beneficiar do uso destas técnicas. A análise dos dados torna bastante aparente a não obrigatoriedade do uso do sistema pelos alunos: documentos considerados importantes para os cursos, tais quais apostilas, são acessados por metade da classe apenas e em muitos casos, menos que isso; muitos cursos têm inúmeros alunos cadastrados que sequer acessam o sistema ou o fazem apenas uma ou duas vezes ao longo de todo o semestre. Pelo fato de o sistema ser usado apenas como apoio para cursos presenciais, a interação pessoal direta entre os alunos permite que os documentos disponibilizados no sistema, uma vez obtidos por um aluno, possam ser copiados extensivamente entre os colegas sem maiores dificuldades. Cursos não dependentes de tecnologia certamente têm alunos que acessam menos estes tipos de recursos, principalmente sem um estímulo adequado por parte do professor, pois esses alunos não sentem necessidade de interagir com sistemas computacionais. Um problema potencial identificado no ambiente TIDIA-Ae em particular é a falta de um recurso de *awareness* que ajude o aluno identificar novas atualizações dos conteúdos dos cursos, uma vez que ele acesse o sistema. Existe uma gama de ferramentas grande disponível no sistema e sem este recurso para guiar o aluno, ele pode se sentir desestimulado de revirar inúmeras páginas em busca de novas informações. O próprio educador pode ter dificuldades em notar requerimentos dos alunos, tais como perguntas postadas num fórum do ambiente, por exemplo. Assim, um recurso que aponte já na página inicial do sistema novos conteúdos e alterações gerais de interesse do usuário, poderiam melhorar significativamente a usabilidade do sistema, o interesse do aluno em usá-lo e também a quantidade e relevância das informações de uso gravadas pelo sistema durante os cursos.

Trabalhos futuros envolvendo a aplicação da metodologia proposta devem se apoiar em dados mais adequados, preferencialmente aqueles gerados por cursos digitais não presenciais como os ministrados pela Univesp ou eventuais cursos neste formato que a própria UNICAMP venha a disponibilizar; embora não precisem se limitar necessariamente ao escopo do TIDIA-Ae ou dos ambientes usados no Brasil. Esta abordagem seria adequada particularmente para reavaliar os esforços aqui realizados em condições mais específicas e favoráveis à geração de resultados palpáveis aos educadores.

Para o caso particular do TIDIA-Ae em sua instância utilizada dentro da UNICAMP seria adequado um estudo mais aprofundado de sua base de dados antes do início dos trabalhos. Um extenso exame do amplo conjunto de tabelas que compõe a base de dados do TIDIA-Ae deve ser realizado para determinar quais dentre estas tabelas contém informações de bom potencial para a realização de estudos investigativos. Um problema que dificulta significativamente este estudo é o sistema gerenciador de banco de dados *MySQL* usado para construir essa base de dados, pois ele não permite a existência física de chaves lógicas entre as tabelas, tornando trabalhosa a identificação das relações entre as inúmeras tabelas existentes. Possivelmente será necessária a construção de um conjunto menor de tabelas capaz de concentrar os dados relevantes para um processo de mineração de dados de maneira análoga ao que foi feito neste trabalho, mas talvez já integrando-as ao sistema e tornando a coleta organizada destas informações neste formato pelo sistema natural e automática.

Finalmente, a integração do protótipo construído para geração dos relatórios de acessos a documentos ao TIDIA-Ae é um projeto potencialmente benéfico ao ambiente, permitindo uma nova camada de obtenção de informações para o educador podendo até gerar resultados parciais durante o curso como discutido em 4.4.2 e 4.7, ainda que um trabalho de adaptação do protótipo à estrutura interna do sistema TIDIA-Ae seja necessário.

Referências Bibliográficas

[1] admin. Teleduc. ensino à distância. http://www.teleduc.org.br/pagina/principal/. Acessado em 24/08/2010.

[2] Esa Alhoniemi, Johan Himberg, Juha Parhankangas, and Juha Vesanto. Som toolbox. http://www.cis.hut.fi/projects/somtoolbox/. Acessado em 07/06/2009.

[3] Esa Alhoniemi, Johan Himberg, Juha Parhankangas, and Juha Vesanto. Som toolbox. http://www.cis.hut.fi/projects/somtoolbox/package/docs2/kmeans_clusters.html. Acessado em 30/06/2009.

[4] Maria Angélica C. de Andrade Cardieri. Mecanismo de monitoramento do uso de recursos web para apoio à avaliação de ambientes. Master's thesis, Faculdade de Engenharia Elétrica e de Computação — UNICAMP, 2004.

[5] Félix Castro, Alfredo Vellido, Àngela Nebot, and Julià Minguillón. Finding relevant features to characterize student behavior on an e-learning system. In Hamid R. Arabnia and Hamid R. Arabnia, editors, *FECS*. CSREA Press, 2005.

[6] Robert Cooley, Pang-Ning Tan, and Jaideep Srivastava. Websift: The web site information filter system. In *Proceedings of the Web Usage Analysis and User Profiling Workshop*, pages 1–6, San Diego, CA, USA, 1999.

[7] Allistair Crossley and Yoav Shapira. Apache tomcat 7 (7.0.2) - logging in tomcat. http://tomcat.apache.org/tomcat-7.0-doc/logging.html. Acessado em 20/08/2010.

[8] Daniel Davis. Journaling - fedora repository 3.4 documentation - duraspace wiki. https://wiki.duraspace.org/display/FCR30/Journaling. Acessado em 23/08/2010.

[9] Victor Hugo Menendez Domínguez. Material detail. http://www.merlot.org/merlot/viewMaterial.htm?id=356173. Acessado em 23/08/2010.

[10] Leticia Dos Santos Machado and Karin Becker. O uso da mineração de dados na web aplicado a um ambiente de ensino a distância. In *I Workshop de Teses e Dissertações em Banco de Dados*, pages 117–121, 2002.

[11] Leticia Dos Santos Machado and Karin Becker. Distance education: a web usage mining case study for the evaluation of learning sites. In *Advanced Learning Technologies, 2003. Proceedings. The 3rd IEEE International Conference on*, pages 360–361, 2003.

[12] FAPESP. Sobre o projeto — tidia-ae. http://tidia-ae.incubadora.fapesp.br/portal/o-projeto. Acessado em 06/06/2009.

[13] Usama Fayyad, Gregory Piatetsky-shapiro, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Knowledge discovery and data mining: Towards a unifying framework. pages 82–88. AAAI Press, 1996.

[14] Inc. Fedora Commons. Home - fedora repository. http://fedora-commons.org/. Acessado em 24/09/2009.

[15] Sakai Foundation. Sakai project | collaboration and learning - for educators by educators. http://sakaiproject.org/. Acessado em 06/06/2009.

[16] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2000.

[17] Hui-Huang Hsu, Chun-Jung Chen, and Wen-Pin Tai. Towards error-free and personalized web-based courses. In *AINA '03: Proceedings of the 17th International Conference on Advanced Information Networking and Applications*, pages 99+, Washington, DC, USA, 2003. IEEE Computer Society.

[18] Adriana Justin Cerveira Kampff, Eliseo Berni Reategui, and José Valdeni de Lima. Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. *Novas Tecnologias na Educação*, 6(2), December 2008.

[19] Teuvo Kohonen. *Self-Organizing Maps, Third Extended Edition*. Springer, 2001.

[20] Teuvo Kohonen and Timo Honkela. Kohonen network. *Scholarpedia*, 2(1):1568, 2007.

[21] Manuel Prieto, Víctor Menéndez, Alejandra Segura, and Christian Vidal. A recommender system architecture for instructional engineering. In Miltiadis Lytras, John Carroll, Ernesto Damiani, and Robert Tennyson, editors, *Emerging Technologies and Information Systems for the*

Knowledge Society, volume 5288 of *Lecture Notes in Computer Science*, pages 314–321. Springer Berlin / Heidelberg, 2008.

[22] LLC. Red Hat Middleware. hibernate.org - hibernate. https://www.hibernate.org/. Acessado em 01/02/2010.

[23] Cristóbal Romero and Sebastián Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.*, 33(1):135–146, 2007.

[24] Cristóbal Romero, Sebastián Ventura, and Paul De Bra. Knowledge discovery with genetic programming for providing feedback to courseware authors. *User Modeling and User-Adapted Interaction*, 14(5):425–464, January 2004.

[25] Cristóbal Romero, Sebastián Ventura, and Enrique García. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1):368–384, 2008.

[26] Randall Scarberry. Hyper-threaded java - java world. http://www.javaworld.com/javaworld/jw-11-2006/jw-1121-thread.html. Acessado em 07/06/2009.

[27] Judy Sheard, Jason Ceddia, John Hurst, and Juhani Tuovinen. Inferring student learning behaviour from website interactions: A usage analysis. *Education and Information Technologies*, 8(3):245–266, September 2003.

[28] Carissa Smith. Fedora commons registry - fedora commons registry - duraspace wiki. https://wiki.duraspace.org/display/FCCommReg/Fedora+Commons+Registry. Acessado em 23/08/2010.

[29] Mário de Souza Neto. Direto online: Percepção de presença em ambientes de educação a distância baseados na web. Master's thesis, Faculdade de Engenharia Elétrica e de Computação — UNICAMP, 2004.

[30] Commons Documentation Team. Commons logging - user guide. http://commons.apache.org/logging/guide.html. Acessado em 20/08/2010.

[31] Moodle Trust. Moodle.org: open-source community-based tools for learning. http://moodle.org/. Acessado em 22/05/2010.

[32] Osmar R. Zaïane. Building a recommender agent for e-learning systems. In *ICCE '02: Proceedings of the International Conference on Computers in Education*, pages 55+, Washington, DC, USA, 2002. IEEE Computer Society.

[33] Osmar R. Zaïane and Jun Luo. Towards evaluating learners' behaviour in a web-based distance learning environment. In *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on*, pages 357–360, 2001.

[34] Osmar R. Zaïane, Man Xin, and Jiawei Han. Discovering web access patterns and trends by applying olap and data mining technology on web logs. *Advances in Digital Libraries Conference, IEEE*, 0:19+, 1998.

[35] Márcio Henrique Zuchini. Aplicações de mapas auto-organizáveis em mineração de dados e recuperação de informação. Master's thesis, Faculdade de Engenharia Elétrica e de Computação — UNICAMP, 2003.