UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Rafael de Moura Speroni

OBTENÇÃO DE DADOS PARA MINERAÇÃO DA UTILIZAÇÃO DA WEB

Dissertação de Mestrado submetida à Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação.

Prof. Dr. Fernando Álvaro Ostuni Gauthier

Florianópolis, Setembro de 2003.

OBTENÇÃO DE DADOS PARA MINERAÇÃO DA UTILIZAÇÃO DA WEB

Rafael de Moura Speroni

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação, Área de Concentração Sistemas de Conhecimento e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Prof. Dr. Fernando Álvaro Ostuni Gauthier
Coordenador do Curso

Banca Examinadora

_____

Prof. Dr. Fernando Álvaro Ostuni Gauthier (Orientador)

_____

Prof. Dr. Dalton Francisco de Andrade

_____

Prof. Dr. Rogério Cid Bastos

Para Valdemar e Theresinha Speroni,
meus pais, mestres e grandes amigos.

AGRADECIMENTO

Agradeço a Deus.

A meus pais, Valdemar e Theresinha e minha irmã Mônica, pelo apoio incondicional. Vocês são meu exemplo de vida e meu porto seguro.

À Lilia, pelo amor, compreensão e paciência, estando sempre disposta a me ouvir e acompanhar.

Ao professor Gauthier, pela orientação e incentivo no desenvolvimento deste trabalho.

Ao colega e amigo Renato Noal, pela colaboração direta no trabalho.

Aos professores Dalton Andrade e Paulo Ogliari pelo interesse e auxílio sempre acessível.

À Verinha, por ser sempre tão atenciosa e sorridente, tornando o trabalho junto à secretaria sempre mais agradável.

Aos meus familiares que, mesmo à distância, estão sempre muito presentes em minha vida, em especial ao tio Ênio, meu padrinho, sempre me esperando com um churrasco.

Aos grandes amigos com quem tive o prazer de conviver em Florianópolis. Piveta, Ricardo, Adamô e Neves, com quem dividi apartamento. Noal, Sílvia, Cássia, Zago, Adriano, Régis, Cassiano, Marcelle, Beto, Ana, Barreto, Guto, Balzan, Thirzá, Cristiano, Dione, André (Guidão), parcerias de respeito para uma festa.

Ao Bar do Guidão, pela Skol gelada e pelo espaço pro meu violão. Que a fonte não seque.

Ao Rancho do Jacuí e seus sócios, pelos momentos de descontração na beira d'água. Mesmo estando longe, sempre que posso, faço questão de comparecer e rever estes bons amigos.

A todos aqueles amigos que direta, ou indiretamente, colaboraram para a realização deste trabalho.

SUMARIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CARD	Classification And Regression Trees
CERN	European Organization for Nuclear Research
CLF	Common Log Format
ECLF	Extended Log Format
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IP	Internet Protocol
ISP	Internet Service Provider
NCSA	National Center for Supercomputer Applications
OLAP	On-Line Analytical Processing
XML	Extensible Markup Language

RESUMO

O crescimento da complexidade, tamanho e tráfego nos sites da Web, faz com que tarefas, como o projeto de sites, necessitem de cada vez mais atenção. Um importante auxílio nestas tarefas é a análise de como o site está sendo utilizado. A mineração da utilização da Web aplica técnicas de mineração de dados aos registros de visitas a sites, normalmente contidos em arquivos de log de servidores Web. Este trabalho apresenta uma metodologia para coleta e transformação dos dados, baseando-se em uma ontologia para a representação do conhecimento envolvido, adicionando conteúdo semântico aos documentos, obtendo-se arquivos de log enriquecidos. A metodologia proposta é aplicada em um site de controle acadêmico, no qual os alunos têm acesso a serviços de, entre outros, procedimentos de matrícula e documentos. Foi implementada, ainda, uma ferramenta para auxiliar no processo de coleta e transformação dos dados, referentes a 3 semanas de utilização do site, correspondentes aos períodos de matrícula e cancelamento de matrícula em disciplinas. Os dados são, posteriormente, submetidos a análises estatísticas, sendo aplicados métodos de agrupamento aos mesmos.

Palavras-chave: Mineração da Web; Seqüência de Requisições; Obtenção de Dados.

ABSTRACT

The complexity, size and traffic increase in Websites makes tasks (like sites development) need each time more attention. An important aid for these tasks is the site utilization analysis. Web Usage Mining applies data mining techniques to site visit logs, usually contained in Web server log files. This work presents a methodology for data collection and transformation, it is based in an ontology for the knowledge representation, adding semantic content to the documents, and obtaining enriched log files. The proposed methodology is applied in an academic control site, in which students access services as school registration and documents procedures. It was also implemented a tool to assist in the process of collection and transformation of data regarded to 3 weeks of site utilization, corresponding to the registration and registration canceling period in disciplines. The data is then submited to statistical analysis, and clustering methods are applicated.

Keywords: Web Mining; Clickstream; Data Obtention.

1 INTRODUÇÃO

1.1 Considerações Iniciais

A grande difusão em que se encontra a Internet faz com que, cada vez mais, haja uma grande quantidade de conteúdos sendo disponibilizados e acessados por usuários do mundo todo. Também as organizações aproveitam-se do alcance e facilidade de acesso para disponibilizar serviços e aplicações desenvolvidas para a Web. Cada acesso, requisição de documentos, por parte do usuário, faz parte da ininterrupta e quase infinita seqüência denominada "*clickstream*" (KIMBALL & MERZ, 2000), o fluxo ou corrente incessante de acessos às páginas do universo Web.

O crescimento da complexidade, tamanho e tráfego nos sites da Web, faz com que tarefas como o projeto de sites, necessitem de cada vez mais atenção. Um importante auxílio nestas tarefas, segundo COOLEY et al. (1999), é a análise de como o site está sendo utilizado.

Uma das bases para estas tarefas está na análise e interpretação dos milhões de registros armazenados diariamente nos logs dos servidores Web, que proporcionam, de uma maneira bastante desestruturada e desorganizada, um retrato da cadeia de acessos às páginas.

A mineração da utilização da Web é a aplicação dos métodos de mineração de dados à análise dos registros da utilização da Web, normalmente na forma de logs de servidores Web (STUMME et al., 2002).

A mineração da utilização da Web mostra-se, entretanto, uma ferramenta fundamental não só à descoberta de padrões de navegação, mas sua utilidade pode ser, também, de inestimável valor para todos os indivíduos e organizações envolvidas no projeto e implementação de sites Web.

Neste trabalho são descritos o desenvolvimento de um modelo para obtenção de dados para a mineração da utilização da Web e a implantação de uma ferramenta que auxilia neste processo de obtenção, a partir dos arquivos de log do servidor Web. O modelo é testado em um portal acadêmico, levando-se em conta apenas o módulo de acesso restrito aos alunos, onde os mesmos estão devidamente identificados.

Os dados obtidos são, posteriormente, submetidos a técnicas de mineração de dados, sendo aplicados métodos de agrupamento, para a identificação de grupos que revelem informações a respeito da maneira como se dá a utilização do site.

1.2 Justificativa

As técnicas de mineração da utilização da Web (*Web Usage Mining*) vêm sendo alvo de muitas pesquisas, uma vez que proporciona informações a respeito da maneira como os usuários estão utilizando os serviços disponíveis em determinado site. Desta forma, sua aplicação mostra-se de grande utilidade nas tarefas de projeto e manutenção de sites, uma vez que os projetistas podem ficar a par de problemas de utilização que necessitem ser resolvidos, bem como possíveis melhorias. Este trabalho disponibiliza informações a respeito da aplicação de métodos de mineração da utilização da Web através do estudo em um site de controle acadêmico, e análise dos dados obtidos.

1.3 Objetivos

1.3.1 Objetivo Geral

Apresentar um modelo para a obtenção de dados de sessões de usuários a ser aplicado na mineração da utilização da Web.

1.3.2 Objetivos Específicos

- Identificar as estratégias para obtenção de dados e as técnicas para mineração da Web;
- Modelar a obtenção de dados sobre a utilização de um site, visando conhecer o tipo de visitas efetuadas;
- Realizar um estudo de caso, demonstrando a aplicabilidade do modelo proposto;
- Verificar a aplicabilidade das técnicas no auxílio às tarefas de projeto e manutenção, através dos resultados obtidos.

1.4 Organização do Trabalho

O presente trabalho está dividido em 9 capítulos.

- O primeiro capítulo apresenta uma introdução ao trabalho desenvolvido, bem como os objetivos do mesmo;
- No Capítulo 2 são apresentados os conceitos relativos à mineração da Web, bem como suas classificações;
- O Capítulo 3 discorre sobre a análise da seqüência de requisições feitas pelos usuários, seus conceitos e técnicas;
- No Capítulo 4 são apresentadas considerações a respeito de ontologias e suas aplicações;
- O Capítulo 5 trata dos conceitos relativos à mineração de dados e uma breve descrição dos métodos de agrupamento (*clustering*);
- No Capítulo 6 é apresentada a modelagem proposta para obtenção de dados para mineração da utilização da Web;
- O Capítulo 7 relata e apresenta a aplicação prática do modelo proposto, através de um estudo de caso;
- O Capítulo 8 apresenta a análise dos dados obtidos na aplicação do modelo, com a aplicação de métodos de agrupamento aos mesmos;
- No Capítulo 9 são apresentadas as conclusões relativas ao trabalho, bem como algumas sugestões para trabalhos futuros.

2 MINERAÇÃO DA WEB (WEB MINING)

2.1 Considerações Iniciais

No desenvolvimento e manutenção de sites da Web, existem, dentre outros, dois tipos de problemas que os projetistas enfrentam: a compreensão de quais são as tarefas que as pessoas estão tentando realizar no site, e a identificação das dificuldades encontradas na execução das mesmas (HONG & LANDAY, 2001).

Para OBERLE et al. (2003), mineração da Web é a aplicação de técnicas de Mineração de Dados ao conteúdo, estrutura, e utilização dos recursos da Web, auxiliando na descoberta de estruturas tanto locais quanto globais (modelos ou padrões) entre as paginas Web. A mineração da Web corresponde ao uso das técnicas de mineração de dados para descobrir e extrair, automaticamente, informações de documentos e serviços da Web (KOSALA & BLOCKEEL, 2000).

Uma distinção é, geralmente, feita entre a mineração da Web que opera nos próprios recursos da Web (comumente diferenciada ainda entre mineração de estrutura e conteúdo), e a que opera na utilização que os visitantes fazem destes recursos.

As pesquisas em mineração da Web envolvem, segundo BORGES & LEVENE (1999), OBERLE et al. (2003), três áreas: Mineração de Conteúdo, Mineração de Estrutura e Mineração da Utilização.

2.2 Mineração do Conteúdo/Texto da Web (Web Content/Text Mining)

A mineração do conteúdo da Web é uma forma de mineração de texto, conforme afirma CHAKRABARTI (2000). O principal recurso da Web que está sendo minerado é uma página Web. Procura-se extrair, portanto, informações relevantes do próprio conteúdo dos documentos e páginas da Web, contidas no código que as gera (HTML, scripts, etc.).

Este tipo de mineração pode tirar proveito da natureza semi-estruturada do texto das páginas Web. As marcações HTML das atuais páginas, e ainda, a marcação das páginas em XML, proporcionam informação relacionada não apenas à aparência das páginas, mas também à estrutura lógica.

Aplica-se este tipo de mineração quando se deseja, por exemplo, detectar co-ocorrências de termos em textos.

2.3 Mineração da Estrutura da Web (Web Structure Mining)

A mineração da estrutura da Web opera, normalmente, na estrutura de *hyperlinks* das páginas Web. O principal recurso minerado, segundo BERENDT (2002), é um conjunto de páginas, abrangendo desde um simples site à Web como um todo, procurando, desta maneira, inferir o conhecimento com base na própria organização dos documentos Web e nas ligações entre eles.

As técnicas explicitam informações adicionais (normalmente implícitas) contidas na estrutura do hiper-texto. Além disso, uma importante área de aplicação é a identificação da relevância relativa de diferentes páginas que aparecem semelhantes quando analisadas isoladamente por seu conteúdo.

É comum, ainda, a aplicação das duas técnicas em conjunto, a chamada mineração de c**onteúdo/estrutura**. STUMME (2002) cita o exemplo do algoritmo de classificação de páginas utilizado pelo Google, um site de busca, que determina a relevância das páginas pelo número de vezes que outras páginas as citam. Neste caso, os *hiperlinks* entre as páginas representam uma estrutura sobre as páginas. Por outro lado, os *hiperlinks* são parte do conteúdo textual das páginas, e as duas abordagens são levadas em conta para a classificação das páginas retornadas.

2.4 Mineração da Utilização da Web (Web Usage Mining)

Mineração da utilização da Web é a aplicação dos métodos de mineração de dados à análise dos registros da utilização da Web, normalmente na forma de logs de servidores Web (STUMME, 2002).

O comportamento do usuário, em um determinado site, é estudado através da análise do registro das requisições a páginas e documentos, feitas pelos visitantes a um site da Web. Conforme cita COOLEY (1999), as requisições são coletadas em um arquivo tipo log nos servidores Web, constituindo a principal fonte de dados a serem minerados, na busca de informações sobre a utilização.

O conteúdo e estrutura das páginas e de um site, em particular, revelam as intenções do seu projeto. O comportamento daqueles que o utilizam, por outro lado, pode revelar estruturas adicionais e relações que não haviam sido projetadas. Para BORGES (1999), entender as preferências de navegação dos usuários é um passo essencial para a melhoria dos serviços oferecidos pelo site.

É usual combinar a mineração de utilização da Web com análises de conteúdo e estrutura, de maneira a encontrar uma lógica entre os caminhos observados com freqüência e as páginas destes caminhos. Para OBERLE et al. (2003), isto pode ser feito utilizando-se uma variedade de métodos. Dentre as diferentes abordagens alguns autores, (FREITAS et al., 2000 e OBERLE et al., 2003), classificam as páginas em termos de uma ontologia pré-definida, enquanto outros, (TINGSHAO et al., 2003), se baseiam na extração de agrupamentos de palavras-chave representados por caminhos mais freqüentes. A própria ontologia pode ser feita manualmente ou aprendida semi-automaticamente, e a classificação das páginas de acordo com a mesma pode ser semi-automatizada de várias formas.

As aplicações de mineração da Web poderiam ser classificadas em duas categorias principais (KOSALA & BLOCKEEL, 2000): as de **aprendizado do perfil do usuário** ou modelagem do usuário em sistemas adaptativos (personalizadas) e **aprendizado de padrões de navegação** (não-personalizadas). Os usuários podem estar interessados em técnicas capazes de detectar suas necessidades e preferências, com possibilidade de personalização das páginas, por exemplo. Os projetistas e mantenedores de sites, entretanto, podem estar interessados em técnicas que os permitam promover melhorias no conteúdo e ergonomia, baseadas, por exemplo, no comportamento de seus usuários.

Para o melhor entendimento dos padrões de utilização, a análise pode levar em conta a semântica das URLs visitadas. OBERLE (2003) apresenta um *framework* para melhorar os registros de utilização da Web com semânticas formais baseados em uma ontologia que fundamenta o site.

2.5 Trabalhos Relacionados

As pesquisas em mineração da utilização da Web têm sido desenvolvidas em aplicações voltadas para comércio eletrônico e personalização de conteúdos, propagandas e manutenção e fidelização de clientes (KOHAVI, 2001). Boas experiências de navegação podem promover um aumento nas vendas e retorno de clientes. Más experiências de navegação podem provocar a perda de clientes insatisfeitos com os serviços oferecidos. Em seu trabalho o autor revisa os ingredientes necessários para o sucesso da Mineração de Dados, aplicando-os ao contexto dos dados referentes a comércio eletrônico.

ZAÏANE et al. (1998) aplicou técnicas de OLAP sobre os arquivos de log, onde, através de classificações e sumarizações, foram encontradas algumas correlações entre eventos e analisadas séries temporais. Foi desenvolvida uma ferramenta, chamada de WebLogMiner, com o objetivo de extrair conhecimentos implícitos em grandes arquivos de log. Verifica-se necessidade de uma melhoria nos arquivos de log, no sentido de que mais dados sejam neles armazenados, para que se possa efetuar um melhor gerenciamento da utilização.

NASRAUOI et al. (1999) descreve um algoritmo para aglomeração dos acessos em seções típicas. São definidos um conceito de "sessão de usuário" e uma medida de distância entre duas sessões. Foi utilizado o algoritmo de aglomeração competitiva CARD, utilizando lógica *Fuzzy*.

SPILIOPOULOU & FAULSTICH (1998) construiram árvores agregadas que representam seqüências freqüentes de acesso, sobre as quais foram aplicadas técnicas de mineração de dados. A ferramenta WUM (*Web Utilization Miner*) minera dados de log nos sites e descobre padrões de navegação na forma de grafos. Os logs são utilizados em sessões e, posteriormente transformados em uma estrutura de árvore. A mineração é feita em cima desta estrutura reduzida e são aplicadas heurísticas para melhorar a performance.

OBERLE (2003) propõe um *framework* para a melhoria dos registros de utilização da Web, com a utilização de semânticas formais baseadas em uma ontologia que sustenta o site. O trabalho descreve as ações do usuário utilizando a taxionomia da ontologia, capturando, assim, os interesses expressos em uma visita de determinado usuário a uma página. As idéias são aplicadas a um portal de que utiliza tecnologias de Web semântica.

A mineração da utilização da web vem sendo alvo de muitas pesquisas, uma vez que a verificação da forma como estão sendo utilizados os serviços em um site, pode demonstrar a existência de falhas, ou mesmo o grau de satisfação dos usuários. A mineração destes dados, na busca de informações conclusivas, depende diretamente da capacidade de obtenção e da qualidade dos dados.

A forma de captura dos dados, durante a utilização das páginas, será descrita na próxima seção, onde se pretende melhor conceituar o processo, e suas formas de aplicação.

3 ANÁLISE DE SEQÜÊNCIA DE REQUISIÇÕES (CLICKSTREAM ANALYSIS)

A navegação na Web gera uma imensa fonte de dados relativos ao comportamento de indivíduos durante suas interações em sites da Web. Segundo LI (2000), à seqüência de requisições (cliques) gerada pelos usuários enquanto movem-se através das páginas de um site, dá-se o nome de "*clickstream*".

"Essa seqüência de cliques é literalmente um registro de cada gesto efetuado por cada visitante a cada site da Web". (KIMBALL & MERZ, 2000).

A seqüência de requisições contém um grande montante de dados quantitativos que podem ser usados para a melhoria do funcionamento do site, conforme afirma LI (2000). Quando propriamente minerada, a análise da seqüência de requisições pode prover respostas a questões específicas, como quais as páginas mais acessadas, o produto mais procurado, a maneira como as pessoas chegam ao site, onde as sessões são interrompidas. Tais respostas podem vir a auxiliar, posteriormente, no desenvolvimento e melhoria dos serviços oferecidos.

Os dados obtidos apresentam, ainda, caráter qualitativo, uma vez que a análise da seqüência de requisições pode revelar padrões de uso no site em estudo e fornecer um maior entendimento a respeito do comportamento dos usuários (ANDERSEN et al., 2000). A seqüência não é somente mais uma fonte de dados tradicionais no ambiente Web, e sim uma coleção desenvolvida de fontes de dados, pois diversas são as formas de registrar o comportamento dos usuários, como visto em KIMBALL & MERZ (2000).

Embora os dados estejam desorganizados, têm grande potencial para tornarem-se informações importantes e anteriormente desconhecidas, sobre a utilização dos sites (KIMBALL & MERZ, 2000). O simples registro de todas as ações do usuário, entretanto, não é suficiente para analisar seu comportamento, uma vez que este tipo de dados, sem tratamento, não mostra, de forma geral, os caminhos percorridos e opções feitas pelo usuário.

3.1 Sessões

O conceito de sessões é muito importante no processo de mineração da Web, pois é base para a identificação dos passos do usuário em um site. NASRAOUI et al. (1999) definem uma sessão como sendo uma seqüência de acessos temporalmente compactos feitos por um usuário.

Para ANDERSEN et al. (2001), a vantagem, quando são levadas em conta as sessões, é a possibilidade de seguir os passos do usuário através do site.

Uma vez que os Servidores Web, tipicamente, não guardam os nomes de usuários, costuma-se definir uma sessão de usuário como sendo os acessos derivados de um mesmo endereço IP, tais que o espaço de tempo decorrido entre eles seja menor que um limite pré-especificado.

A identificação de um usuário, através de seu endereço IP, entretanto, pode não ser tão confiável, uma vez que tal endereço pode ser dinamicamente determinado pelo provedor (ISP) através do qual o usuário obtém acesso, podendo, até mesmo, ser modificado dentro de uma mesma sessão. (SILVERSTON, 2002).

3.2 Tipos de Usuários

Para KIMBALL (2000), o "visitante" requer uma atenção em especial, visto que, em sistemas Web, existem diferentes tipos de usuários:

Usuários completamente anônimos, identificados apenas pelo endereço IP, que se revela de valor questionável, uma vez que ele apenas identifica uma porta de saída do provedor de internet do usuário. Nestes casos, tais portas são, normalmente, configuradas para serem designadas dinamicamente, o que impossibilita reconhecer um usuário em sessões diferentes.

Usuários que aceitaram *cookies*, arquivos armazenados na máquina do visitante. Tal método fornece uma identificação confiável para a máquina, pois ele é verificado em cada pagina acessada pelo visitante. Pode-se ter certeza, portanto, de que pelo menos a máquina responsável pela sessão é a mesma utilizada anteriormente.

Usuários identificados, que não apenas podem ter aceitado um *cookie*, mas que em algum momento revelaram seu nome ou outra informação que os identifique. É um usuário que está devidamente identificado no sistema.

Dependendo, então, do tipo de resultado esperado, deve-se levar em conta os tipos de usuários que estarão envolvidos no processo de coleta dos dados.

3.3 Coleta de Dados

Em muitos casos, os arquivos de log do Servidor Web são uma ótima fonte de dados, pois todos os tipos de servidores populares emitem tais arquivos automaticamente, conforme

SWEIGER (2002). Há um registro no arquivo de log para cada transação HTTP, e tais arquivos seguem formatos padronizados.

Uma entrada de log de utilização contém, normalmente, o registro do percurso de uma página de origem até uma página de destino, incluindo o IP da máquina cliente que originou a chamada, o tipo de método de acesso realizado (POST ou GET), além de outros dados, a depender do padrão utilizado pelo servidor Web.

Em sites de maior porte, é comum a distribuição da hospedagem das páginas em diferentes servidores Web. Neste caso, cada servidor é responsável pela resposta para as requisições às páginas nele hospedadas e, conseqüentemente, armazenam seus registros de acessos em arquivos de log próprios.

A Figura 1 exemplifica parte de uma visita a um site, cujas páginas de conteúdo estático estão hospedadas no “Servidor 1”, e as páginas de conteúdo dinâmico no “Servidor 2”. Durante a visita, o usuário requisita, em primeiro momento, uma página de conteúdo estático. O “Servidor 1”, ao disponibilizá-la, registra em seu arquivo de log a solicitação do usuário. Em seguida, o usuário faz requisição a uma página de conteúdo dinâmico. Esta requisição é enviada ao “Servidor 2”, que disponibiliza o conteúdo solicitado, armazenando em um arquivo de log esta requisição.

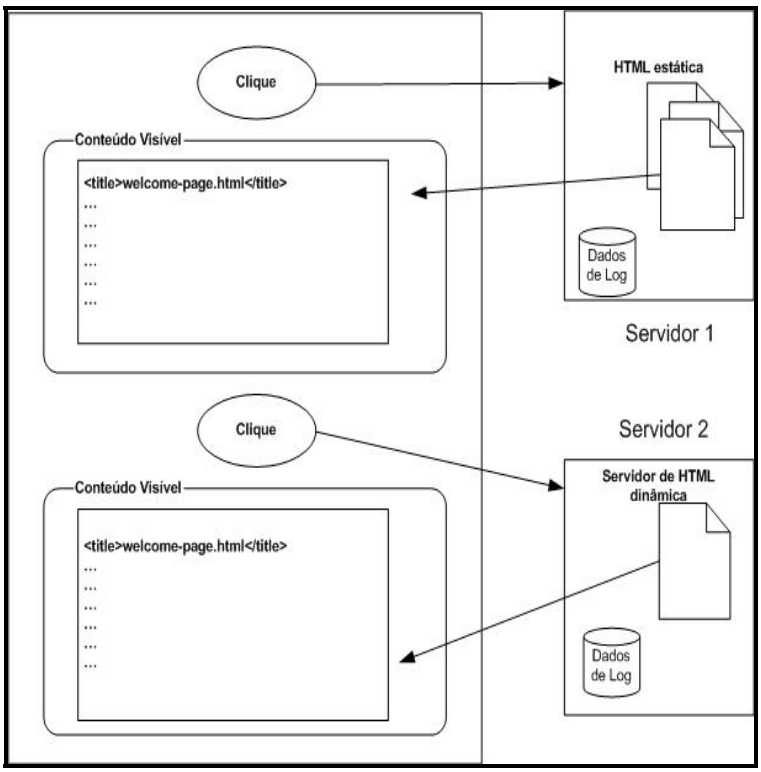

Figura 1 – Páginas distribuídas em diferentes Servidores Web (Fonte: Adaptado de KIMBALL & MERZ)

O formato padrão para arquivos de log, *Common Log Format* – CLF, foi especificado pelo CERN e NCSA como parte do protocolo HTTP. Este padrão, entretanto, por ser bastante limitado em termos das informações que armazena, foi ampliado posteriormente pelo W3C para o *Extended Log Format* – ECLF (HALLAM-BAKER & BEHLENDORF, 1996), quando foram adicionadas mais informações aos registros. Ao se utilizar o *Extended Log Format*, é possível que sejam especificados quais os campos que se deseja gravar no log.

Para LI (2000), é necessário, entretanto, tomar cuidado no processo de coleta de dados, uma vez que a seqüência de cliques provê uma grande quantidade de dados. A cada requisição ao site, tem-se um ponto de dado. Logo, é importante levar em conta o custo para a posterior análise dos mesmos.

3.4 Filtragem dos Dados

Os registros provenientes dos logs do Servidor Web, por possuírem alimentação direta do mesmo, são, normalmente, muito carregados, com informações que podem não ser relevantes para o caso de análise de comportamento de usuários. Desta forma, o acesso a uma simples

página Web provoca a gravação de várias entradas de log no servidor, para cada uma das imagens, scripts ou outros arquivos carregados juntamente com a página.

Para a mineração de utilização, normalmente, apenas as entradas de log referentes aos acessos às páginas HTML serão de interesse, uma vez que os demais arquivos, especialmente imagens, são carregados automaticamente, e, por isso, nem sempre serão úteis para um sistema que procure minerar os padrões de navegação do usuário, pois não foram explicitamente solicitados por este (BOULLOSA, 2002).

Para remover imagens, uma abordagem simples é retirar do log todas as entradas associadas às extensões conhecidas para elas, tais como GIF, JPG, JPEG, etc. O mesmo pode ser feito em relação a arquivos de sons ou outras fontes multimídia.

KIMBALL & MERZ (2000) propõem um método alternativo de captura, fornecendo uma fonte de dados independente da atividade do site da Web, chamado de **Servidor de Registros Nulos**.

O servidor de registros nulos é um servidor da Web, cuja missão primária não é entregar conteúdo, mas sim aceitar dados de logs. É um servidor Web na medida em que aceita requisições, mas não entrega dados significativos em resposta a essas solicitações, e sim uma resposta nula mínima. O servidor hospeda apenas um arquivo de conteúdo mínimo, como uma imagem transparente de um único pixel.

Esta alternativa é particularmente aplicável à captura de dados de um site da Web corporativo altamente distribuído com uma diversidade de servidores departamentais individuais.

Nas páginas em que se deseja capturar os dados relativos à visita dos usuários, é colocada uma requisição a este arquivo contido no servidor de registros nulo. Os arquivos de log deste servidor irão conter os registros de requisições destas páginas, estejam elas hospedadas em servidores distribuídos ou não.

A Figura 2 representa o funcionamento do Servidor de Conteúdo Nulo.

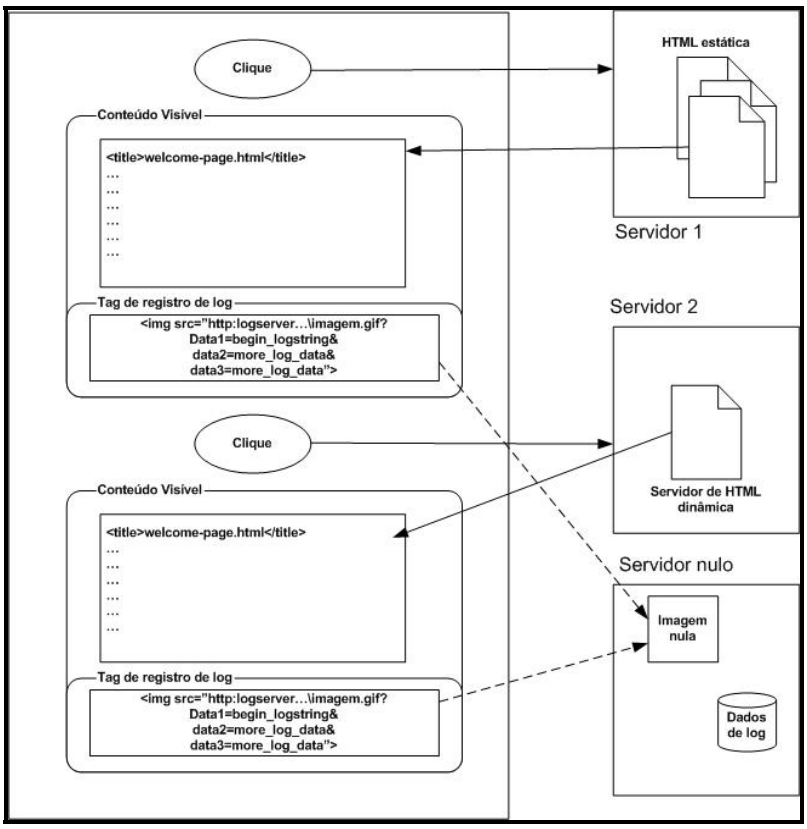

Figura 2 – Funcionamento do servidor de registros nulo (Fonte: KIMBALL & MERZ, 2000)

O processamento dessas solicitações significará um aumento muito pequeno no tempo de descarga e, essencialmente nenhum aumento se for colocado na parte inferior de uma página. Desta forma, mesmo se houver um retardo na recuperação desta pequena imagem, a página será apresentada normalmente (KIMBALL & MERZ, 2000).

Para CASERTA (2001), a imagem do *clickstream* está finalmente sendo transformada de um contador de ações para uma fonte crucial de informações sobre os usuários. Assim, a análise da seqüência de requisições mostra-se de grande valia no processo de obtenção de dados a respeito da utilização do site.

A forma de implementação deve visar a obtenção dos dados, de maneira que se tenha condição de identificar quais as requisições que correspondem a cada usuário. O significado dos dados obtidos será, posteriormente, analisado, a fim de que se tornem informações potencialmente úteis.

A próxima seção descreve conceitos a respeito de ontologias, e sua utilização em sistemas computacionais, no auxílio da representação do conhecimento.

4 ONTOLOGIAS

4.1 Considerações Iniciais

O processo de conceituação, segundo GRUBER (1993), implica em definir um corpo de conhecimento, representado formalmente, sendo necessário, para isto, axiomas lógicos que restrinjam as possíveis interpretações dos termos definidos.

A tarefa de extração de informação pode ser vista como um problema de compatibilidade semântica entre um padrão definido por usuários, e uma porção de informação escrita em linguagem natural. Para GUARINO (1997), a compreensão ontológica do padrão precisa ser adequadamente especificada, e comparada com as implicações ontológicas dos dados. Assim, ontologias consistem em teorias de vários tipos expressando o significado de vocabulários compartilhados, que serão utilizados nestas tarefas.

No contexto de compartilhamento de conhecimento, o termo Ontologia é utilizado com o significado de uma **especificação de uma conceitualização**. Ou seja, uma ontologia é uma descrição (como uma especificação formal) dos conceitos e relacionamentos que possam existir para uma pessoa ou uma comunidade de agentes.

Os sistemas de representação de conhecimento tradicionais são, tipicamente, centralizados, sendo necessário que todos compartilhem exatamente as mesmas definições de conceitos em comum. Este controle central, entretanto, tendo seu tamanho e escopo crescente, torna o sistema de difícil gerenciamento.

"Web Semântica é uma extensão da Web atual, na qual é dado um significado bem definido à informação, melhorando a forma como as pessoas e os computadores trabalham em conjunto. (BERNERS-LEE, 2001)".

4.2 Razões para a Utilização de Ontologias

A utilização de ontologias traz vantagens como a discriminação dos tipos de conhecimento e sua representação explícita aumentando a flexibilidade no gerenciamento do conhecimento. Algumas das razões que levam à utilização de ontologias são descritas a seguir (NOY & MCGUINNESS, 2001).

O **compartilhamento da compreensão da estrutura da informação** através das pessoas e agentes de software é um dos objetivos mais comuns no desenvolvimento de ontologias.

Proporcionar **reutilização de conhecimento do domínio**. Uma determinada ontologia, desenvolvida por um grupo de pesquisadores, pode, simplesmente ser reutilizada, por outros, para seus domínios. Ainda, a construção de uma grande ontologia pode ser feita através da integração de várias outras existentes, que descreverão partes do grande domínio.

Separar o conhecimento a respeito do domínio do conhecimento operacional. Pode-se, por exemplo, descrever a tarefa de configuração de um produto e seus componentes, de acordo com uma especificação, e implementar um programa que faça esta configuração, independente dos produtos e componentes.

Analisar o conhecimento do domínio é possível, uma vez que esteja disponível uma especificação declarativa dos termos. A análise formal dos termos é extremamente valiosa para reutilizar e estender ontologias existentes.

Normalmente a ontologia de um domínio não é um objetivo final. O desenvolvimento de uma ontologia está, portanto, relacionado à definição de um conjunto de dados e sua estrutura para utilização por outras aplicações.

Não há uma maneira ou metodologia “correta” para o desenvolvimento de ontologias. NOY & MCGUINNESS (2001) propõem um processo possível para o desenvolvimento, dividindo-o em alguns passos:

Determinar o domínio e escopo da ontologia, desejando-se responder a perguntas como:

- Qual é o domínio que a ontologia vai cobrir?
- Para que a ontologia será utilizada?
- Para que tipos de questões as informações nela contida irão prover respostas?
- Quem irá utilizar e manter a ontologia?

As respostas para estas perguntas podem modificar-se durante o processo de modelagem da ontologia, mas auxiliarão a delimitar seu escopo. Uma forma de delimitação do escopo é a utilização de **questões de competência**, perguntas que possam ser respondidas baseadas na ontologia.

Levar em consideração ontologias já existentes, para, se possível reutilizá-las, refinando e especializando-as.

Enumerar termos importantes na ontologia. É usual a utilização de uma lista de termos sobre os quais deseja-se tratar e explicar.

Definir as classes e sua hierarquia, tarefa que pode ser realizada levando-se em conta abordagens "*top-down*", onde são definidos conceitos gerais, com uma subseqüente especialização. Na abordagem "*bottom-up*" são definidas as classes mais específicas, com um posterior agrupamento em conceitos gerais. Pode-se utilizar, ainda, uma combinação dos dois processos, onde são definidos os conceitos mais salientes para, depois, especializar e generalizá-los apropriadamente.

Criar instâncias das classes anteriormente definidas.

A representação da utilização de ontologias pode ser exemplificada através de um suposto portal onde são gerenciados os dados referentes a projetos e pesquisadores. A Figura 3 representa a estrutura e relacionamentos entre as páginas do suposto exemplo. As páginas possuem ligações entre elas e através de uma página principal, onde existe um Menu para escolha das páginas.

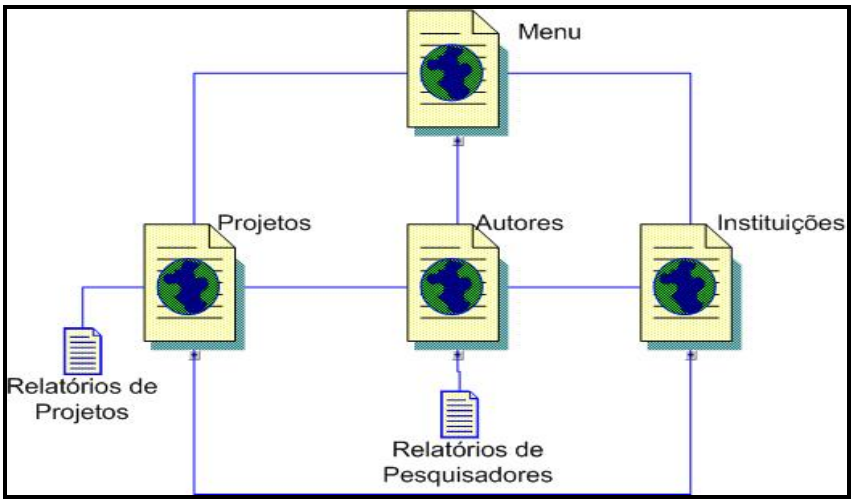

Figura 3 – Estrutura do portal de exemplo

O conhecimento contido no portal refere-se a projetos, seus autores e instituições. Identificados os conceitos envolvidos, pode-se representar a ontologia do portal através da Figura 4.

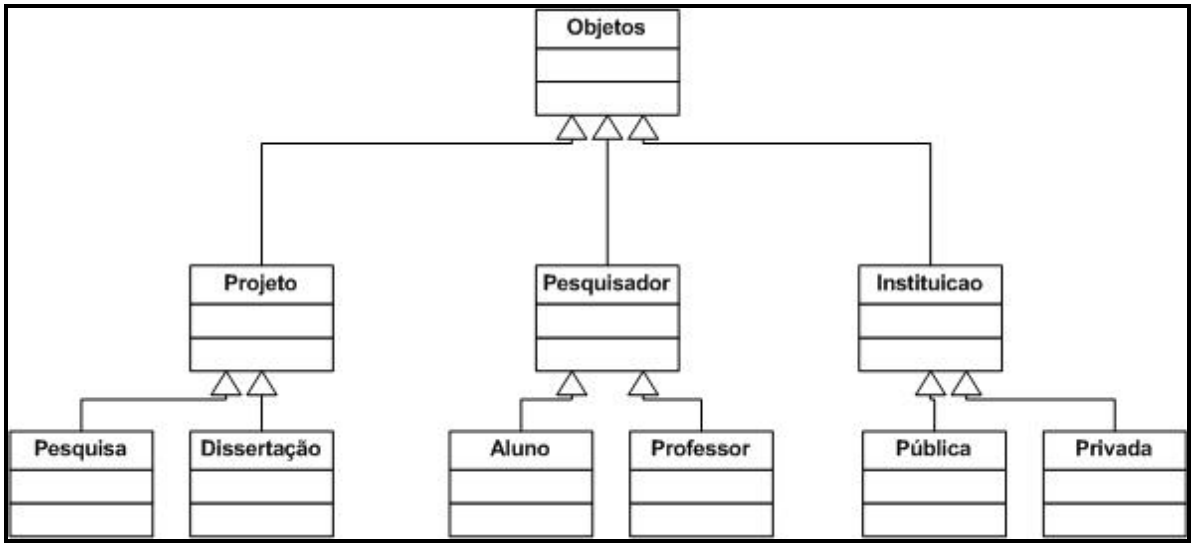

Figura 4 – Exemplo de representação de ontologia

A ontologia do exemplo representa objetos de três classes: a classe **projeto**, especializada em **pesquisa** e **dissertação**; **pesquisador**, especializada em **aluno** e **professor**; **instituição**, que se especializa em **pública** e **privada**.

A estrutura do site, ou a maneira como suas páginas estão relacionadas não representa, de forma explícita, a forma como este conhecimento está relacionado.

Na próxima seção, pretende-se apresentar os conceitos e terminologias a respeito de Mineração de Dados. São apresentadas, ainda, técnicas utilizadas para análise de conjuntos de dados.

5 MINERAÇÃO DE DADOS E DESCOBERTA DE CONHECIMENTO

Este tópico examina os conceitos fundamentais a respeito da Mineração de Dados e da Descoberta de Conhecimento. São discutidas a metodologia e terminologias bem como sua importância e formas de aplicação.

5.1 Conceituação

A agilidade de resposta na recuperação de informação é um dos principais motivos da utilização de sistemas de bancos de dados. A forma como os dados estão armazenados, entretanto, pode não revelar, de maneira clara, as informações desejadas. Devido à grande quantidade de dados com que trabalham os sistemas de bancos de dados atuais, os relacionamentos e padrões, existentes nos mesmos, não estão explícitos, sendo necessária a aplicação de técnicas específicas para este fim.

A mineração de dados, ou "*data mining*", corresponde à atividade automática ou semi-automática de exploração e análise de grandes quantidades de dados com o propósito de neles descobrir regras e padrões antes desconhecidos (BERRY & LINOFF, 1997). Por possuir natureza interdisciplinar, faz uso de áreas como a estatística, inteligência artificial, teoria dos grafos, além da teoria de banco de dados.

O problema da mineração de dados parte do pressuposto de que os grandes bancos de dados do mundo real são verdadeiras "minas" de conhecimento (DEOGUN *et al.*, 1997), onde repousam informações de grande valor que podem ser encontradas através de técnicas e algoritmos adequados.

O desenvolvimento das pesquisas em mineração de dados é justificado pelas variadas aplicações práticas que podem ser associadas à descoberta do conhecimento, previamente ignorado, armazenado em grandes bancos de dados.

A mineração de dados é também conhecida, por alguns autores, como **mineração de bancos de dados** ("*database mining*") ou ainda, segundo PIATESTKY-SHAPIRO (2000), **descoberta de conhecimento em bancos de dados** ("*knowledge discovery in databases*") .

BERRY & LINOFF (1997) referem-se à descoberta de conhecimento como um dos "estilos" de mineração de dados. Seria uma abordagem "*bottom-up*", em que, partindo-se dos

dados, tenta-se chegar a um conhecimento previamente ignorado. Outra abordagem, de testes de hipóteses, seria uma tentativa "*top-down*" de provar (ou negar) idéias previamente concebidas.

Tais abordagens correspondem, respectivamente, aos dois tipos clássicos de inferência conhecidos como indução e dedução. Do ponto de vista do aprendizado, a indução parte de casos particulares (dados de treino) para os gerais, tentando desenvolver um modelo explicativo. Já a dedução, parte do geral – um modelo prévio – para o particular – os dados (CHERKASSKY & MULIER, 1998).

Para muitos pesquisadores, entretanto, a descoberta de conhecimento é uma área mais ampla, da qual a mineração de dados é apenas uma etapa, voltada, principalmente, para os métodos de produção de conhecimento. FAYYAD *et al.* (1996), que definem a descoberta de conhecimento como a extração não-trivial de informações potencialmente úteis, previamente desconhecidas e implícitas em dados brutos, dividem seu processo nas seguintes etapas, conforme representado na Figura 5:

a) **definição dos domínios** onde serão realizadas as análises e quais os objetivos do processo de descoberta de conhecimento;

b) **criação de um conjunto de dados**, através da seleção entre as diferentes fontes de dados disponíveis;

c) **pré-processamento dos dados**, incluindo a retirada dos dados desnecessários e o tratamento daqueles que estão indisponíveis ou que possam conter alguma incerteza;

d) **transformação dos dados**, adequando suas dimensões e variáveis de maneira apropriada, para que estes estejam coerentes com as necessidades dos métodos que serão utilizados na próxima etapa;

e) **mineração de dados**, etapa que envolve efetivamente as técnicas e algoritmos que produzirão o conhecimento procurado;

f) **análise e interpretação dos resultados** encontrados na etapa anterior.

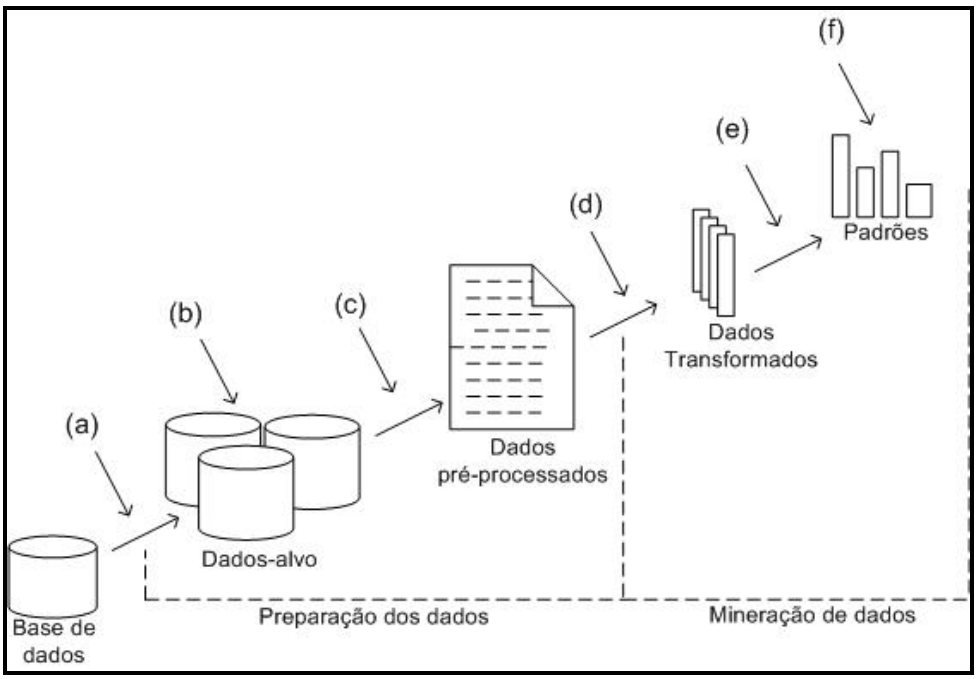

Figura 5 – Etapas do processo de Descoberta de Conhecimento (Fonte: Adaptação de AMARAL & AMARAL, 2002)

Estas etapas podem ser repetidas quantas vezes sejam necessárias para a obtenção de resultados satisfatórios.

Após este processo, a consolidação do conhecimento é feita através de sua incorporação no desempenho do sistema. A verificação e resolução de conflitos potenciais podem, portanto, ser feita com base no conhecimento prévio.

5.2 Modelagem

O processo de modelagem consiste na criação de um modelo que represente um conjunto de dados (GROTH, 1998). Costuma-se utilizar a modelagem quando se tem um nível maior de conhecimento da área e da relação que se deseja ajustar.

Normalmente, um modelo não é uma representação fiel do conjunto de dados, caso contrário, pode resultar em excessos de treinamento e, dependendo da sua utilização, tornar-se um modelo específico (GROTH, 1998).

Modelos podem, ainda, ser utilizados na previsão de eventos futuros, uma vez que, ainda que os dados históricos não proporcionem uma visão do futuro, os padrões tendem a se repetir. Assim, segundo GROTH (1998), a construção de um modelo que represente um conjunto de dados pode servir para que sejam feitas previsões a partir deles.

5.3 Preparação dos Dados

Para que sejam aplicadas as técnicas de mineração de dados, é necessário que os dados sejam preparados, a fim de que estes estejam limpos e consistentes quando analisados.

5.3.1 Limpeza dos Dados

A limpeza é uma parte muito importante no processo de preparação dos dados, uma vez que a simples análise dos dados brutos pode ser prejudicada devido a problemas de notação, escala e, até mesmo, erros tipográficos. Neste passo deve ser feita a limpeza dos dados de maneira que os incorretos ou incompletos sejam ignorados. Com isto é feita uma purificação dos dados utilizando operações básicas, como a eliminação de dados fora de padrão (AMARAL & AMARAL, 2002).

Exemplificando, pode-se supor um determinado conjunto de dados de clientes, provenientes de diferentes bases de dados. A variável que representa o sexo do cliente é tratada, em uma das bases, como sendo “M” e “F”. Outra base de dados, entretanto, pode utilizar, para esta variável, os valores “0” e “1”. No processo de limpeza deve-se padronizar as variáveis, para que utilizem a mesma notação.

5.3.2 Dados Ausentes

No tratamento de grandes volumes de dados, é comum que ocorra a falta de dados. Observações cujos dados não estejam completos, quer seja por indisponibilidade do dado na ocasião da inclusão, ou por falhas na revisão.

O tratamento destes casos é necessário para que os resultados do processo de mineração sejam confiáveis. Uma alternativa como solução desse problema é o uso de técnicas de imputação, que consistem em fazer a previsão individual dos dados ausentes, completando os dados para posterior análise.

5.3.3 Derivação de Dados

Muitos dos dados contidos nas bases de dados apresentam relacionamentos entre si. Quando existe, portanto, a necessidade de dados que não estejam disponíveis, mas que podem ser obtidos através da combinação ou transformação de outros disponíveis, recebem a denominação

de derivados. São dados que podem, normalmente, ser calculados a partir dos dados existentes, como, por exemplo, a idade de um cliente a partir de sua data de nascimento.

5.4 Alguns Métodos para Mineração de Dados

5.4.1 Árvores de Decisão

Para BERRY & LINOFF (1997), uma árvore de decisão é um modelo de previsão que pode ser visto como uma árvore. Cada ramo da árvore é uma questão classificatória, e suas folhas são partições do conjunto de dados com suas classificações.

Uma vantagem na utilização deste método é o fato de ser facilmente compreensível. As regras obtidas caracterizam-se por ser coletivamente exaustivas e mutuamente exclusivas.

Pode-se citar, como exemplo de algoritmos de árvore de decisão: CART ("*classification and regression trees*"), CHAID ("*chi-squared automatic induction*"), ID3 e C4.5 (QUINLAN, 1993), baseados em árvores intuitivas, SLIQ (**"***supervised learning in quest***"**).

5.4.2 Geração de regras de associação

A descoberta de regras de associação, segundo SAVASERE (1995) é aplicada a bancos de dados de transações, onde cada transação é composta por um conjunto de itens, e no qual procura-se descobrir quando a presença de um conjunto de itens implica na presença de um outro item na mesma transação.

A geração de regras de associação é muito utilizada em sistemas comerciais, especialmente naqueles direcionados à área de vendas, e é conhecida popularmente como "*market basket analysis*", já que, para atingir o seu objetivo de encontrar grupos de itens que ocorram juntos, analisa uma situação semelhante à que ocorre quando se utiliza uma cesta de supermercado (a transação).

5.4.3 Classificação

A classificação, também chamada de **aprendizado supervisionado**, funciona com base na utilização de um mapeamento prévio para grupamentos especificados (GROTH, 1998). A denominação de aprendizado supervisionado é justificada pelo fato de que é um supervisor externo quem fornece a entrada e a saída desejadas.

Os modelos de classificação tentam rotular e colocar registros em classificações previamente existentes, além de lhes adicionar outras informações tais como a probabilidade de ocorrência em determinado contexto.

As técnicas de classificação permitem o desenvolvimento de perfis de itens com atributos em comum. Estes perfis podem ser usados para classificar novos itens, quando adicionados ao banco de dados.

5.4.4 Agrupamento (“clustering”)

“Análise de agrupamentos engloba uma variedade de técnicas e algoritmos cujo objetivo é separar objetos em grupos similares.” (BUSSAB, et. al).

O agrupamento ou segmentação, também chamado de *Clustering* ou **aprendizado não-supervisionado**, é um método cujo funcionamento consiste em agrupar os dados mais parecidos, com base na distância entre eles. Criam grupos menores a partir de grandes conjuntos de registros, levando em consideração as características comuns entre eles.

Na análise de agrupamentos (“*clustering analysis*”), são reunidos os itens cujas características são semelhantes. A diferença para a classificação é que, neste caso, não se tem conhecimento de quais serão os grupos resultantes até que o processo seja concluído, enquanto que na classificação os grupos são pré-definidos.

Deste modo, busca-se construir, no agrupamento, modelos que encontrem itens similares entre si, a partir de alguma métrica de “distância” entre estes, que serão assim colocados juntos em novos agrupamentos (“*clusters*”) ou em agrupamentos já existentes. Para tanto, podem ser utilizados métodos geométricos, estatísticos ou mesmo redes neurais.

5.4.5 Memory–based reasoning

Estes métodos procuram fazer predições de novos itens de dados, a partir de itens já conhecidos, procurando pelos vizinhos mais próximos a estes últimos e combinando seus valores para encontrar valores de predição e classificação. Uma de suas vantagens é a possibilidade de aprendizado sempre que novos itens sejam acrescentados ao banco de dados (BERRY & LINOFF, 1997).

5.4.6 Redes neurais e algoritmos genéticos

As redes neurais são uma das técnicas mais comuns de mineração de dados, pela sua ampla aplicabilidade (MITCHELL, 1997). Utilizam modelos que procuram reproduzir de maneira mais simplificada as conexões neuronais do cérebro. Por esse método, procura-se, a partir de um conjunto de treino, aprender padrões gerais que possam ser aplicados à predição e classificação.

Dois dos principais problemas das redes neurais são as dificuldades associadas ao entendimento dos modelos por ela produzidos e sua grande sensibilidade ao formato dos dados de entrada.

Os algoritmos genéticos (GA – “genetic algorithms”) utilizam as idéias e mecanismos da genética e seleção natural (operações de seleção, “cross-over”, mutação) para encontrar os parâmetros ótimos que descrevem funções preditivas

(MITCHELL, 1997). São desenvolvidas sucessivas gerações de soluções, até que apenas algumas “sobrevivam” e as funções encontradas convirjam para uma solução ótima.

5.5 Técnicas de Análise de Agrupamentos Utilizadas em Mineração de Dados

As técnicas de **Análise de Agrupamentos** têm por objetivo encontrar diferentes grupos em uma determinada base de dados. Havendo também a necessidade de ser determinado o número e as características desses grupos (EVERITT, 1993).

Um Agrupamento (*cluster*) é um subconjunto de todos os possíveis subconjuntos distintos da população, com características em comum (EVERITT, 1993). Um Agrupamento é um conjunto de objetos com características semelhantes.

As técnicas de análise de agrupamentos dividem-se em dois tipos; as de **agrupamento hierárquico** e as de **agrupamento não-hierárquico**, também chamadas de técnicas de agrupamento **por partição**.

No a**grupamento hierárquico**, o processo de identificação de agrupamentos é geralmente realimentado recursivamente, utilizando tanto novos dados quanto grupos já identificados previamente como entrada para o processamento. As técnicas hierárquicas podem, ainda, ser classificadas em **aglomerativas** e **divisivas** (BUSSAB et al., 1990).

Aglomerativas são aquelas onde, inicialmente, cada observação corresponde a um agrupamento e, a cada passo do agrupamento, vão sendo unidas a outros grupos. No final do processo, o resultado é apenas um grupo contendo todas as observações, conforme a Figura 6. Já

as técnicas **divisivas** são aquelas cujo processamento se inicia com apenas um grupo, que contém todas as observações. As etapas seguintes vão dividindo os dados em agrupamentos até que, no final do processo, cada observação corresponda a um grupo.

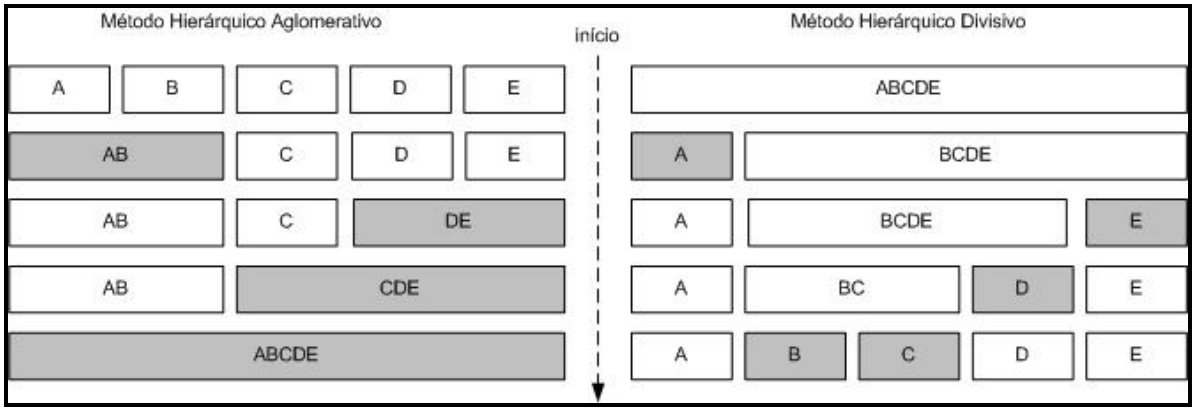

Figura 6 – Funcionamento dos Métodos Hierárquicos

Dentre os métodos de agrupamento hierárquico, pode-se citar o **Método da Centróide**, o **Método das Médias das Distâncias**, o **Método da Ligação Simples** (Vizinho Mais Próximo), o **Método da Ligação Completa (**Vizinho Mais Longe**)**, e o Método Ward's.

No **agrupamento não hierárquico (**também conhecido como Técnica de Partição**)**, o processo procura por um número pré-determinado de agrupamentos distintos, sem levar em consideração o resultado do passo anterior. Os *n* elementos são agrupados em um número pré-determinado de *k* agrupamentos. O método mais comum é o **Metodo das k-Médias (***k-means*). Um problema sério é não se conhecer o número ideal de agrupamentos (*k*).

No método das k-médias, uma vez escolhido o número de agrupamentos (*k*), são escolhidos os elementos que servirão como "sementes" para os grupos. Esta escolha pode ser feita aleatoriamente, definida pelo usuário, ou seguindo alguma outra ordem. Cada grupo terá um valor médio, chamado de centróide. O processamento consiste em calcular, para cada elemento a ser agrupado, sua distância para com os *k* grupos, adicionando-se o elemento ao grupo e recalculando-se o centróide do grupo. Após os elementos serem agrupados é calculada o grau de homogeneidade interna de cada um dos grupos, através da *Soma dos Quadrados dos Erros,* inspirado em Análise de Variância. Quanto menor for este valor, mais homogêneos são os elementos dentro de cada grupo e "melhor" será a partição (BUSSAB et al., 1990).

Na próxima seção, pretende-se apresentar uma metodologia para a representação, na forma de ontologia, do conhecimento envolvido entre as páginas e documentos de um site, bem como para o processo de coleta dos dados da sua utilização.

6 MODELO PARA OBTENÇÃO DE DADOS PARA WEB MINING

6.1 Considerações Iniciais

Nesta seção, pretende-se apresentar uma metodologia para a representação da estrutura do conhecimento envolvido, entre as páginas e documentos de um site, na forma de uma ontologia. É apresentado, ainda, o procedimento para a realização da coleta dos dados relativos à sua utilização, além da transformação e pré-processamento dos mesmos, para posterior análise.

A aplicação de técnicas de Mineração de Dados aos dados provenientes deste processo visa a obtenção de padrões de utilização que levem a conclusões relevantes para a manutenção e projetos de ampliação e reestruturação do site.

6.2 Levantamento da ontologia do site

A análise do comportamento dos usuários consiste na verificação e avaliação da maneira como estes estão utilizando o site, em relação à estrutura existente. Segundo Cooley (1999) a visão dos projetistas sobre como o site deveria ser utilizado, bem como o conteúdo das páginas, está inerente na sua estrutura. Cada ligação entre as páginas existe porque os projetistas acreditam que, de alguma maneira, estas páginas estejam relacionadas.

O levantamento da estrutura, portanto, consiste em identificar e representar a maneira como as páginas estão dispostas no site, bem como as ligações entre elas e possíveis caminhos a serem seguidos dentro do site.

Com base no estudo a respeito do site, seu projeto, a estrutura entre suas páginas e serviços oferecidos, é possível que seja delimitado o conhecimento envolvido. A simples representação de sua estrutura, com base na maneira em que se dá o relacionamento entre as páginas, através de seus *hiperlinks*, pode não ser suficiente para demonstrar a forma em que o conhecimento está disposto no sistema.

O levantamento de sua ontologia mostra-se, portanto, uma maneira de ser representado o conhecimento envolvido no sistema, a fim de tornar mais claro o entendimento, possibilitando, ainda, futuras extensões, além de uma possível reutilização para outros projetos que envolvam conhecimento de domínio semelhante. Deve-se identificar os objetos envolvidos, organizando-os em classes, de acordo com suas características semelhantes.

6.3 Inclusão de Conteúdo Semântico nas Páginas do Site

Uma vez que tenha sido feito o levantamento da estrutura e ontologia do site, pode-se atribuir rótulos de identificação, para cada página visitada ou ação efetuada pelo usuário, com a finalidade de prover às mesmas uma identificação compreensível, ou seja, com significado possível de ser mapeado para a ontologia.

A utilização destes rótulos permite que as páginas requisitadas, ou ações realizadas, pelo usuário sejam reconhecidas posteriormente nos arquivos de log. Para isto, a cada requisição contida nas páginas e ações que tenham sido determinados como pontos de estudo, deve-se enviar os rótulos, juntamente com identificadores de usuário e sessão, na forma de parâmetros, para que, assim, estes dados fiquem registrados nos arquivos de log do Servidor Web.

A coleta dos dados utiliza-se dos arquivos de log, gerados pelo servidor Web, como fonte de dados. A utilização dos rótulos e identificadores vem a auxiliar o processo de coleta, uma vez que estes dados irão enriquecer os arquivos de log.

6.4 Captura das Requisições do Usuário

O processo de coleta de dados sobre o comportamento dos usuários consiste em capturar suas atividades, ou seja, os passos que seguiu durante sua visita ao site.

Sugere-se, aqui, o conceito dos servidores de registros nulos, proposto por Kimball (2001). Conforme visto no Capítulo 3, é proposta a criação de um novo Servidor Web, cujo objetivo principal são seus arquivos de log, que ficarão centralizados e contendo apenas as requisições à imagem de conteúdo nulo, feitas pelas páginas especificadas.

Dependendo, da maneira como está distribuída a hospedagem das páginas do site, pode-se optar pela não criação de um novo servidor Web, para o armazenamento dos registros, uma vez que se a hospedagem estiver centralizada em apenas um servidor Web, os registros das atividades dos usuários estarão, conseqüentemente, centralizados nos arquivos de log deste Servidor.

A opção pela utilização do mesmo Servidor Web para hospedagem das páginas e da imagem de conteúdo nulo, apresenta o inconveniente de que os registros referentes às requisições da imagem nula estarão misturados aos demais registros referentes às páginas do site, com suas imagens e arquivos solicitados. Para resolução deste problema, é feita uma simples filtragem nos arquivos de log. A filtragem consiste, basicamente, na pesquisa de cada linha do arquivo,

buscando pelo nome da imagem de conteúdo nulo. Aquelas que contiverem a referencia ao arquivo são relevantes ao estudo, as demais são ignoradas.

O produto da filtragem será um arquivo de log contendo apenas os registros das chamadas à imagem, que correspondem às ações dos usuários, obtendo-se um arquivo semelhante ao que seria obtido, caso tivesse sido feita a opção pela criação de um Servidor de Conteúdo Nulo separado. Em termos formais, estes dados são geralmente a descrição das transações realizadas pelos usuários na forma de uma seqüência T de URLs u: $T=[u_i]$, onde T corresponde a uma seqüência de requisições, e u_i a URL $i_{ésima}$ URL visitada pelo usuário dentro desta seqüência.

6.5 Mapeamento das URLs para os objetos da ontologia

A utilização de uma ontologia, visando uma melhor compreensão do conhecimento, permite que as requisições dos usuários sejam analisadas através de uma abordagem voltada aos objetos, componentes da ontologia, e não às páginas, especificamente. Para isto, é necessário que se faça um mapeamento dos rótulos e demais parâmetros identificadores das requisições, armazenados nos arquivos de log, para os objetos da ontologia.

O mapeamento consiste em transformar a seqüência *T* de URLs em uma seqüência de objetos contidos na ontologia: *m(u)=O.* Desta maneira, pode-se determinar que, a requisição de uma URL *u*, indica um interesse no objeto *O.*

Dentre as formas de fazer esse mapeamento, uma delas, utilizada em Oberle (2003), consiste em considerar cada objeto ou conceito como sendo um valor em uma dimensão de informação. Por exemplo, a URL que representa o conceito de “confirmação de matricula em 3 disciplinas”, pode ser *verdadeiro* na dimensão “realizou matricula” e pode ter o valor *3* na dimensão “número de disciplinas matriculadas”.

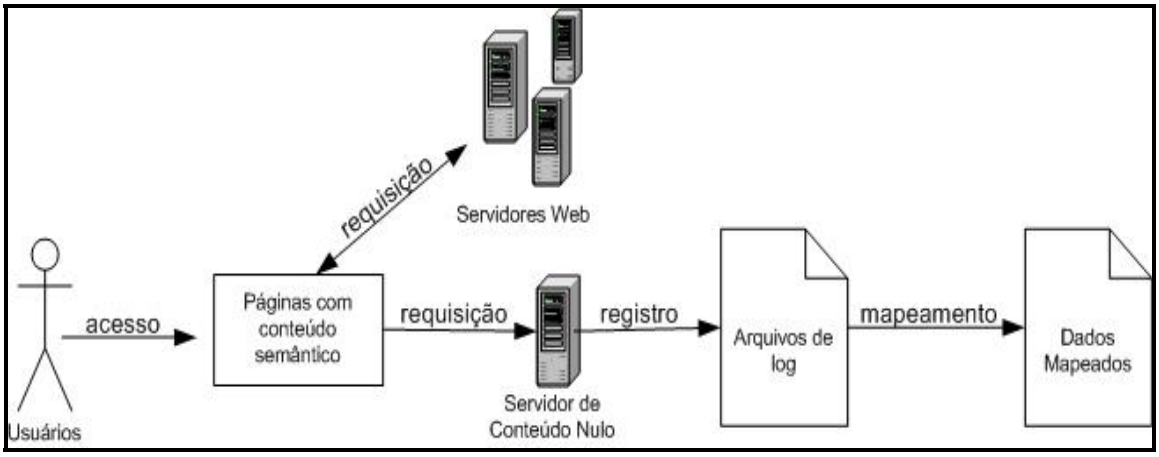

Figura 7 – Modelo para obtenção dos dados

A Figura 7 mostra a maneira como se dá o processo proposto para coleta dos dados sobre a utilização do site. O modelo proposto compreende, ainda, uma etapa anterior, de levantamento da ontologia e preparação das páginas e documentos. A ilustração aborda, entretanto, apenas as etapas compreendidas entre o acesso dos usuários e seu mapeamento aos objetos da ontologia.

Posterior à coleta, o processo de pré-processamento dos dados irá depender dos métodos de mineração de dados a serem aplicados.

Na próxima seção são apresentadas a aplicação do modelo proposto e a ferramenta desenvolvida para auxílio na coleta dos dados.

7 APLICAÇÃO DO MODELO PARA O CASO EM ESTUDO

7.1 Considerações Iniciais

Nesta seção são apresentados os procedimentos para implantação do modelo proposto para Mineração da Utilização da Web (Web Usage Mining). A aplicação do modelo foi feita no sistema acadêmico do Curso de Pós Graduação em Ciência da Computação da Universidade Federal de Santa Catarina.

O site é composto por, basicamente, quatro seções. Uma seção de acesso irrestrito, através da qual os usuários em geral, sem identificação, podem obter informações e avisos a respeito do curso, suas linhas de pesquisa, professores e disciplinas.

As outras partes do sistema são de acesso restrito, ou seja, para conseguir acesso, os usuários devem estar identificados. Uma, para acessos por parte dos professores, onde estes podem acompanhar e gerenciar informações a respeito dos alunos sob sua orientação ou matriculados nas disciplinas que ministra, bem como os candidatos a ingresso no curso, em época de seleção de alunos. A secretaria do curso conta com uma seção, também de acesso restrito, onde pode ser feito o acompanhamento da situação dos alunos sob orientação de cada professor do curso, bem como a dos alunos em processo de seleção para ingresso no curso.

Na realização deste trabalho levou-se em conta, apenas, a parte do sistema cujos usuários são os alunos do curso, e apenas conseguem acesso mediante informação de seu número de matrícula e senha. Nesta seção, os usuários efetuam suas matrículas e cancelamento, desde que dentro dos devidos prazos. Existem, ainda, outros serviços aos quais os usuários podem ter acesso por meio das páginas, que são melhor descritos no levantamento da ontologia.

7.2 Levantamento da Ontologia do Site

Para o levantamento da ontologia do site, levou-se em conta a maneira como as páginas estão estruturadas, suas ligações e possibilidades de navegação.

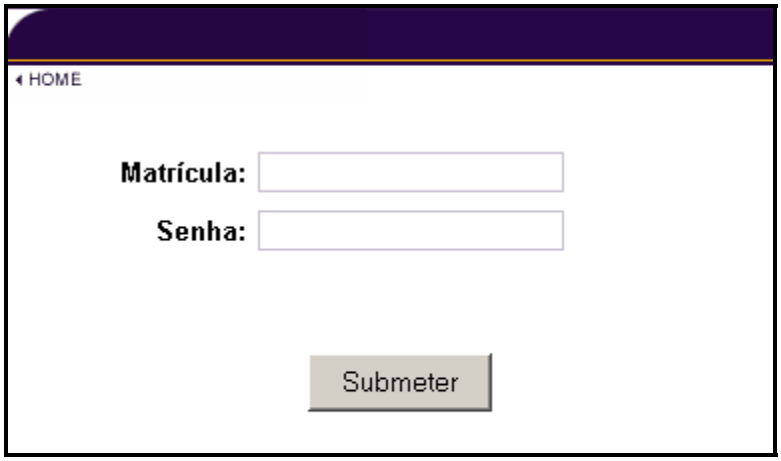

Figura 8 – Index – Página para entrada no sistema

A Figura 8 apresenta a pagina inicial, onde o usuário deve informar seu número de matrícula e senha para ingressar no sistema.

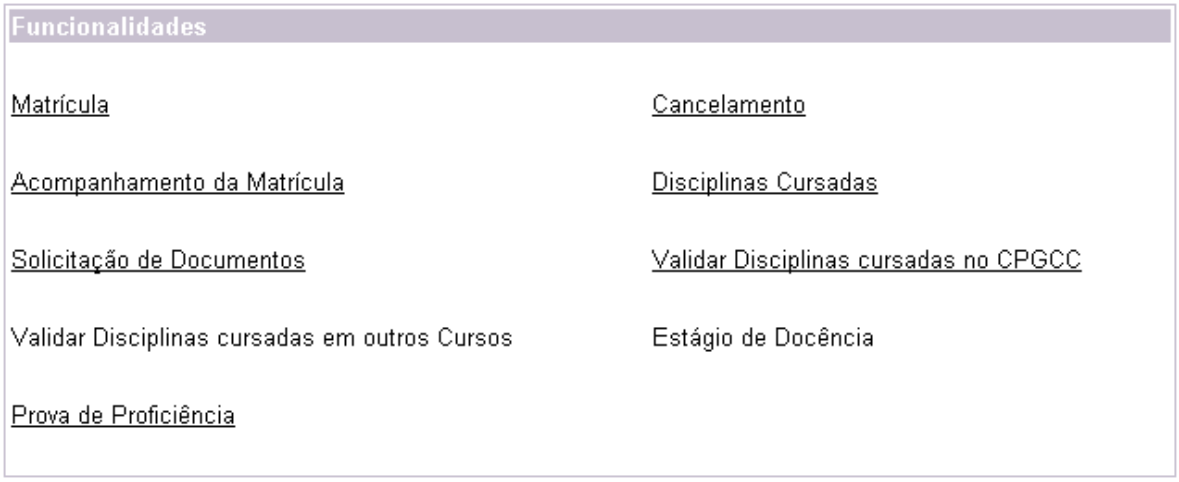

Figura 9 – Funcoes - Página principal para escolha dos serviços

Na Figura 9 é apresentada a página principal, onde o usuário faz a opção pelos serviços que deseja efetuar.

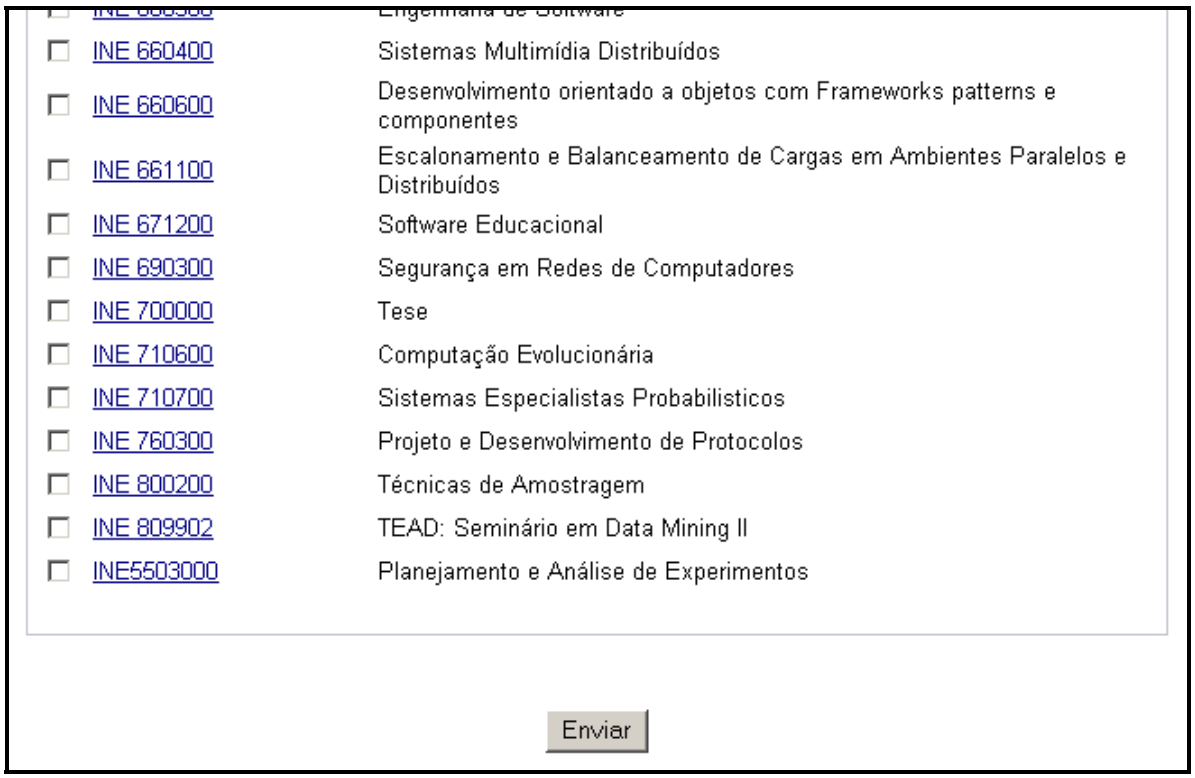

Figura 10 – Form_Matr – Página do formulário de matrícula

A Figura 10 apresenta um trecho da página do formulário para o procedimento de matrícula, onde o aluno faz a opção pelas disciplinas em que deseja se matricular.

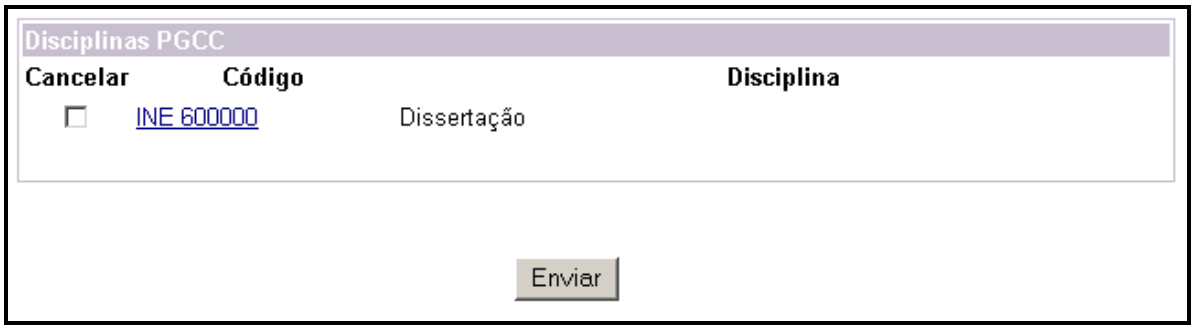

Figura 11 – Form_Canc – Página do formulário de Cancelamento de Matrícula

A Figura 11 é apresentado um trecho da página onde está disponibilizado o formulário para cancelamento das disciplinas em que o aluno está matriculado.

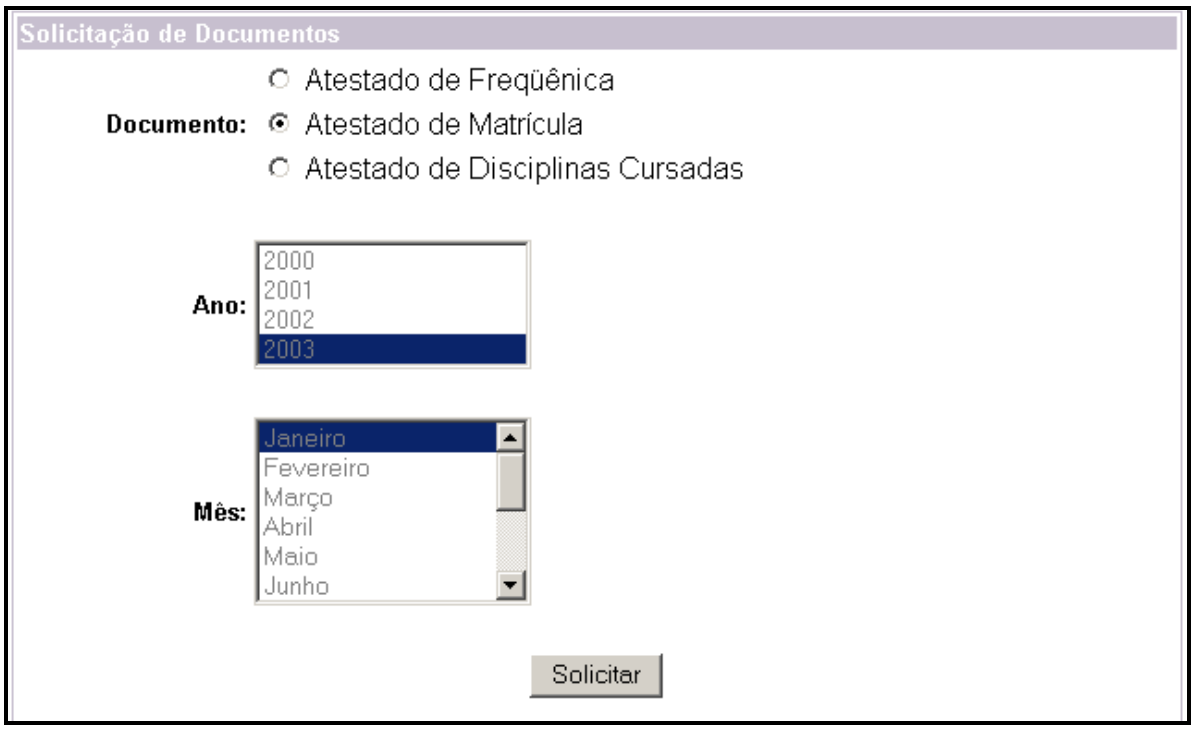

Figura 12 – Solic_Disc – Formulário para Solicitação de Documentos

Na Figura 12 é mostrado um trecho da página onde é disponibilizado o formulário para solicitação de documentos junto à secretaria. O usuário faz a opção pelo tipo de documento desejado e, no caso de Atestado de Freqüência, o mês e ano.

Disciplinas Cursadas

Disciplina	Ano	Período	Conceito
INE 610200 - Inteligência Artificial	2001	1º	A
INE 650400 - Computação Distribuída	2001	1º	A
INE 650500 - Banco de Dados	2001	1º	B
INE 660300 - Engenharia de Software	2001	1º	A
INE 610300 - Redes Neurais I	2001	2º	A
INE 610500 - Introdução à Robótica	2001	2º	A
INE 690100 - Sistemas Operacionais Seguros	2001	3º	A
INE 769902 - TAESD: Programação Orientada a Objeto	2002	1º	A
INE 600000 - Dissertação	2002	2º	*
INE 600000 - Dissertação	2002	3º	*
INE 600000 - Dissertação	2003	1º	*
INE 660900 - Seminário de Sistemas de Informação Baseados na Web I	2003	1º	*
INE 600000 - Dissertação	2003	2º	*

Disciplinas Validadas - Outros Cursos

Disciplina	Ano	Período	Conceito
Nenhum registro encontrado			

Disciplinas Validadas no CPGCC

Disciplina	Ano	Período	Conceito
Nenhum registro encontrado			

Figura 13 – Disciplinas – Página para verificação das disciplinas cursadas

A Figura 13 apresenta trecho da página onde são mostradas as disciplinas já cursadas e as que o aluno está matriculado.

A maneira como as páginas estão estruturadas está representada na Figura 14, e sua descrição consta na Tabela1.

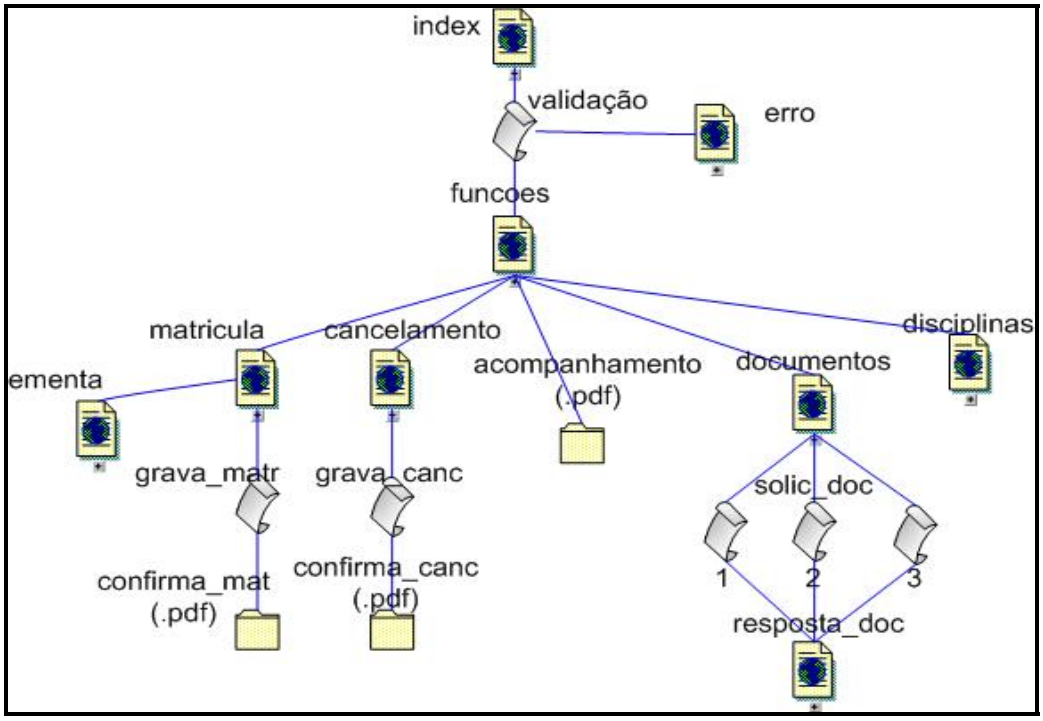

Figura 14 – Levantamento da Estrutura do site

Embora a realização deste levantamento possibilite uma visão da maneira como as páginas estão estruturadas e relacionadas, não explicita quais os serviços à disposição dos usuários. Assim, procede-se com a realização de um levantamento das classes em instâncias de serviços oferecidos.

Em termos da ontologia, o site trata de uma série de serviços disponibilizados aos alunos. Basicamente, os serviços oferecidos podem ser classificados em 3 classes: os relacionados a **procedimentos**, os relacionados a **documentos**, e aqueles relativos às **disciplinas**. A ontologia apresenta, ainda, classes virtuais, que podem auxiliar na representação do conhecimento, embora não cheguem a corresponder a ações específicas.

Os serviços relacionados a **procedimentos** especializam-se em:

- procedimento de **matrícula**, especializado em **confirmação da matrícula** e abandono do procedimento;
- procedimento de **cancelamento de matrícula**, especializado em **confirmação do cancelamento** e abandono do procedimento.

Os serviços relacionados a **documentos** são especializados em 3 classes: **solicitação de documentos** à secretaria do curso, para posterior retirada, documentos de **resposta a procedimentos**, e documentos para **visualização em página**.

Serviços de **solicitação de documentos** especializam-se em:

- Solicitação de **atestado de matrícula**;
- Solicitação de **atestado de freqüência**;
- Solicitação de **atestado de disciplinas cursadas**;

Serviços de **resposta a procedimentos** são especializados em:

- **Resposta de Matrícula**, comprovante da realização do procedimento de matrícula;
- **Resposta de Cancelamento**, comprovante da realização do procedimento de cancelamento de matrícula;
- **Resposta de Acompanhamento**, documento para acompanhamento da situação da matrícula do aluno.

A especialização referente à **visualização em página** refere-se ao documento através do qual o aluno pode ter acesso aos seus conceitos nas **disciplinas** que já cursou, ou está matriculado.

Os serviços relativos às disciplinas são aqueles onde os alunos podem obter informações a respeito das disciplinas oferecidas pelo curso. Suas **ementas** podem ser acessadas através destes serviços.

A representação destes objetos, na forma de uma ontologia, bem como sua relação e estrutura, está na Figura 15.

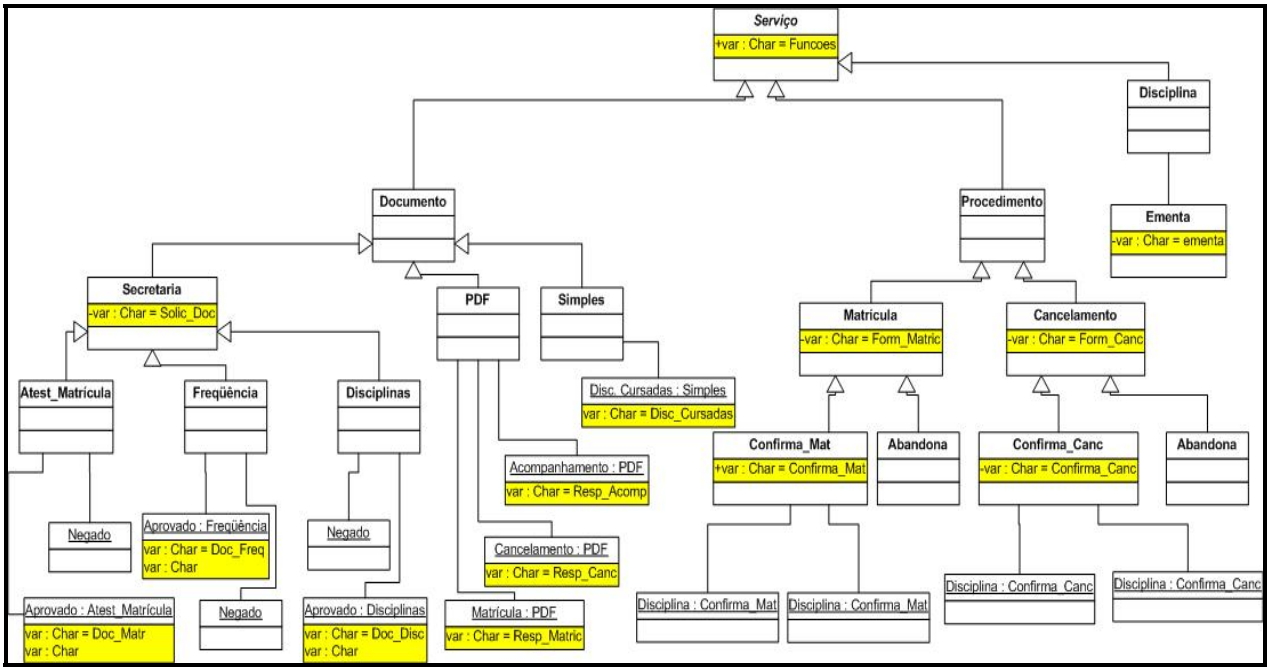

Figura 15 -Levantamento da Ontologia do site

Objeto	*Descrição*
Função	Representa o momento de escolha dos serviços a serem realizados
Solic_Doc	Relacionado ao serviço de solicitação de documentos
Doc_Matr	Serviço de solicitação de atestado de matrícula confirmado
Doc_Freq	Serviço de solicitação de atestado de freqüência confirmado
Doc_Disc	Serviço de solicitação de atestado de disciplinas cursadas confirmado
Resp_Acom	Serviço de documento de resposta do acompanhamento emitido
Disciplinas	Serviço de documento de disciplinas cursadas emitido
Ementa	Serviço de verificação de ementa de disciplinas
Form_Matr	Início do serviço para procedimento de matrícula
Confirma_Matr	Solicitação de matrícula nas disciplinas escolhidas
Resp_Matr	Serviço de documento de resposta do proc. de matrícula emitido
Form_Canc	Início do serviço para procedimento de cancelamento de matrícula
Confirma_Canc	Solicitação de cancelamento nas disciplinas escolhidas
Resp_Canc	Serviço de documento de resposta do cancelamento emitido

Tabela 1 – Descrição dos objetos sob estudo

Esta representação, na forma de uma ontologia, permite um melhor entendimento do conhecimento envolvido no site. A Tabela 2 descreve os objetos da ontologia que estão sendo considerados para o mapeamento.

7.3 Inclusão de Conteúdo Semântico nas Páginas do Site

A inclusão de conteúdo nas páginas consiste na sua identificação, através de rótulos que serão adicionados às mesmas, para, posteriormente, serem mapeados para a ontologia. Além do rótulo que identifica as páginas, outros parâmetros podem, ainda, ser adicionados, conforme na Figura 16.

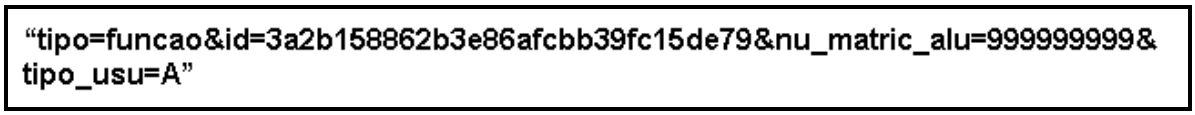

Figura 16 – Exemplo de rótulo de página

O exemplo representa a passagem de parâmetros identificadores da requisição, que são separados pelo caractere “&”. Neste exemplo, existe a presença dos parâmetros “**tipo**”, cujo valor é igual “**função**”, seguido do parâmetro “**id**”, cujo valor é “**3a2b158862b3e86afcbb39fc15de79**”, e do parâmetro “**tipo_usu**” com valor “**A**”.

Os parâmetros aqui citados são os utilizados neste trabalho, onde **tipo** refere-se ao rótulo da página requisitada, **id** refere-se ao identificador da sessão do usuário, **nu_matric_alu** é o identificador único do usuário e **tipo_usu** indica o tipo de usuário, se aluno ou professor. Além destes, são utilizados mais cinco parâmetros (**P1**, **P2**, **P3**, **P4**, **P5**) que são utilizados para o envio de informações adicionais, em caso de necessidade.

A maneira como os parâmetros são adicionados às paginas para que sejam enviados, na forma de requisições, para registro da utilização é detalhada no próximo item.

7.4 Captura das Requisições do Usuário

Para a captura dos cliques é criada uma imagem com conteúdo nulo, aqui chamada de “**null.gif**”, disponibilizada no servidor Web. Adicionando-se, então, ao final de cada página, uma chamada HTML a esta imagem, pode-se esperar um aumento insignificante no tempo de descarga das mesmas.

```
<img src="http://.../img/null.gif?tipo=funcao&id=>05fc8969cb19c0cc4b65bb39d924dc7c&
nu_matric_alu=999999999&tipo_usu=A" width= "0" heigth= "0">
```

Figura 17 – Código para chamada da imagem de conteúdo nulo

A Figura 17 apresenta um exemplo do código HTML necessário para que haja uma requisição à imagem de conteúdo nulo. Um código semelhante a este, colocado ao final de uma página da qual se deseja obter dados de utilização envia uma requisição ao servidor hospedeiro da imagem.

Estas chamadas irão resultar em um registro nos arquivos de log do servidor, conforme a Figura 18.

```
2003-03-20 01:16:42 200.180.16.8 – 150.162.60.53 443 GET /img/null.gif tipo=função
&id=05fc8969cb19c0cc4b65bb39d924dc7c&nu_matric_alu=999999999&tipo_usu=A 200
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+98)
```

Figura 18 – Registro armazenado no arquivo de log

O trecho aqui apresentado representa um registro, armazenado no arquivo de log do servidor, correspondente à requisição do exemplo, colocada ao final de uma página. Alem da data, hora, endereço IP, que são parte do padrão de armazenamento dos arquivos de log, os dados passados intencionalmente, na forma de parâmetros, indicam uma requisição à página de "função", feita pelo usuário, cuja identificação é "999999999", além do identificador da sua sessão, e tipo de usuário "A", indicando ser um aluno.

Conforme foi citado anteriormente, os arquivos de log do servidor contêm, além das requisições à imagem de conteúdo nulo, registros de todas as solicitações de páginas e arquivos solicitados, uma vez que está sendo utilizado o mesmo servidor Web. É necessário, portanto, que seja feita uma filtragem nos arquivos de log.

O resultado da filtragem, em busca dos registros significativos, será um arquivo de log contendo apenas os dados das chamadas à imagem, que correspondem às ações dos usuários. Em termos formais, estes dados são, geralmente, a descrição das transações realizadas pelos usuários na forma de uma seqüência T de URLs u: $T=[u_i]$.

Um exemplo destas seqüências é apresentado na Figura 19. Observe-se que informações de mais de uma sessão estão presentes no exemplo, uma vez que os identificadores de sessão ("**id**") apresentam valores diferentes.

Figura 19 – Trecho de arquivo de log após a filtragem

Analisando o trecho do arquivo, é possível perceber que todas as linhas mostradas no exemplo representam requisições à imagem de conteúdo nulo (GET /img/null.gif). Estas linhas são, portanto, referentes ao conteúdo desejado. Nota-se, ainda, que o conteúdo realmente importante está presente após a URL, que são justamente os parâmetros, e que o trecho representado na Figura 19 apresenta requisições de várias seções, de usuários diferentes.

7.5 Mapeamento das URLs para os objetos da ontologia

Uma vez que os dados provenientes dos arquivos de log fazem referência aos rótulos de cada ação ou acesso a página, é necessário que seja feito um mapeamento para os objetos da ontologia. Consiste, portanto, em transformar a seqüência T de URLs em uma seqüência de objetos contidos na ontologia: m(u)=O.

Nesta abordagem, a seqüência das páginas estáticas e dinâmicas é mapeada em um conjunto de objetos o_{ij} onde i representa o número da observação e j representa a dimensão de informação.

Para a maioria das dimensões de informação foram utilizadas variáveis binárias que representam a existência do conceito ou objeto.

7.6 A Ferramenta Desenvolvida

Para auxiliar no processo de coleta de dados, foi desenvolvida uma aplicação, cujo objetivo principal é realizar a filtragem dos dados nos arquivos de log, e retirar destes os dados

interessantes. A ferramenta conta com configurações a respeito da localização dos arquivos de log originais, e dos resultantes da filtragem, bem como o horário em que se deseja realizar a filtragem. Uma vez que os servidores Web, em sua configuração padrão, geram arquivos de log diários, pode-se fazer a filtragem com horário pré-estabelecido. A Figura 20 apresenta o formulário de configuração da ferramenta, onde são definidas as pastas com que a aplicação irá trabalhar para o gerenciamento dos dados.

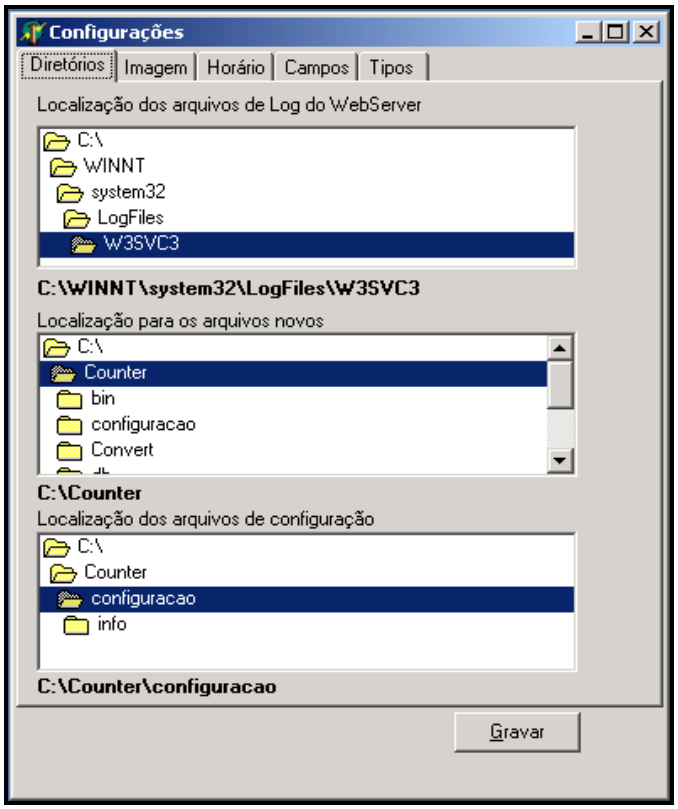

Figura 20 – Formulário de configurações da ferramenta

Algumas configurações são feitas, ainda, em arquivos do tipo “XML”. Nestes são definidos, por exemplo, os parâmetros para a realização do mapeamento, bem como a estrutura dos vetores de representação dos dados. A Figura 21 apresenta o trecho do arquivo de configuração de mapeamento.

```
<?xml version="1.0" standalone="yes"?>
<DATAPACKET Version="2.0">
<METADATA>
<FIELDS>
<FIELD attrname="ORIGEM" fieldtype="string" WIDTH="32"/>
<FIELD attrname="PARAMETRO" fieldtype="string" WIDTH="32"/>
<FIELD attrname="DESTINO" fieldtype="string" WIDTH="32"/>
</FIELDS>
</METADATA>
<ROWDATA>
<ROW ORIGEM="can_grava" PARAMETRO="1" DESTINO="CONFIRMA_CANC"/>
<ROW ORIGEM="cancelamento" PARAMETRO="" DESTINO="FORM_CANC"/>
<ROW ORIGEM="disciplinas" PARAMETRO="" DESTINO="DISCIPLINAS"/>
<ROW ORIGEM="documentos" PARAMETRO="" DESTINO="FORM_DOC"/>
<ROW ORIGEM="ementa" PARAMETRO="" DESTINO="EMENTA"/>
<ROW ORIGEM="erro" PARAMETRO="" DESTINO="ERRO_LOGIN"/>
<ROW ORIGEM="sempermissao" PARAMETRO="" DESTINO="ERRO_LOGIN"/>
<row ORIGEM="funcao" PARAMETRO="" DESTINO="FUNCAO"/>
<ROW ORIGEM="matricula" PARAMETRO="" DESTINO="FORM_MATR"/>
```

Figura 21 – Trecho do arquivo para configuração do mapeamento

O mapeamento para os objetos da ontologia pode basear-se, além dos rótulos dados para cada ação, nos demais parâmetros que são juntamente passados a cada requisição, representados, na Figura 18, por “PARAMETRO”. Um exemplo para isto é o caso de quando um aluno solicita um atestado de matrícula à secretaria. O formulário para solicitação de documentos à secretaria é o mesmo, seja para solicitação de atestados de matrícula, freqüência ou de disciplinas cursadas. A solicitação funciona em uma página de conteúdo dinâmico, onde conforme a opção selecionada pelo usuário, será enviada uma requisição para “solic_doc”, com um valor diferente (“2” para freqüência, “3” para matricula e 4 para disciplinas cursadas) para o parâmetro “P1”.

Desta forma um registro, no arquivo de log contendo uma requisição semelhante a “...tipo=solic_doc P1=3 ...” pode ser mapeada para, por exemplo “...tipo=DOC_MATRIC”, se esta for sua configuração de mapeamento. Assim, o funcionamento do arquivo de mapeamento é exemplificado com auxílio da Tabela 3.

Origem	Parâmetro	Destino
Solic_doc	2	DOC_FREQ
Solic_doc	3	DOC_MATR
Solic_doc	4	DOC_DISC

Tabela 2 – Exemplo de configuração de mapeamento

Neste caso, o item “Destino” corresponde ao rótulo atribuído à página, ou ação, o item “Parâmetro” indica uma informação adicional ao rótulo, e o item “Destino” refere-se ao objeto da ontologia a ser mapeado.

Com este mapeamento é possível, portanto, que cada linha do arquivo de log seja verificada, afim de que se identifique o objeto a que se refere. Feito isso, a transformação para a forma de uma matriz torna-se mais simples, uma vez que cada objeto da ontologia será representado por uma coluna, e as linhas serão equivalentes aos registros. Cada registro pode ser transformado para um vetor, onde o objeto a que ele se refere tem valor “1”, e os demais tem valor “0”. Para os casos de Procedimentos de Matrícula e de Cancelamento de Matrícula, entretanto, este valor pode ser diferente de “1” e “0”, pois é armazenado o número de disciplinas envolvidas no procedimento.

Id	Doc_Matr	Doc_Freq	Doc_Disc	Resp_Matric	Resp_Canc	Resp_Acomp	Funcoes	Form_Matric	Form_Canc
00a249ba8d760005ed9d148f39a364db	0	0	0	0	0	0	1	0	0
00a249ba8d760005ed9d148f39a364db	0	0	0	0	0	0	0	0	0
00a249ba8d760005ed9d148f39a364db	0	0	0	0	0	0	1	0	0
027a42e5d0cda95e27798c201110c84e	0	0	0	0	0	0	1	0	0
027a42e5d0cda95e27798c201110c84e	0	0	0	0	0	0	0	1	0
027a42e5d0cda95e27798c201110c84e	0	0	0	0	0	0	0	0	0
027a42e5d0cda95e27798c201110c84e	0	0	0	0	0	0	0	0	0
027a42e5d0cda95e27798c201110c84e	0	0	0	1	0	0	0	0	0
027a42e5d0cda95e27798c201110c84e	0	0	0	0	0	0	1	0	0
027a42e5d0cda95e27798c201110c84e	0	0	0	0	0	1	0	0	0
027a42e5d0cda95e27798c201110c84e	0	0	0	0	0	0	1	0	0
027a42e5d0cda95e27798c201110c84e	0	0	0	0	0	1	0	0	0
05d85e3fee09c6555fd90b6ad1f11349	0	0	0	0	0	0	1	0	0
05d85e3fee09c6555fd90b6ad1f11349	0	0	0	0	0	0	0	1	0
05d85e3fee09c6555fd90b6ad1f11349	0	0	0	0	0	0	0	0	0
05d85e3fee09c6555fd90b6ad1f11349	0	0	0	0	0	0	0	0	0
05d85e3fee09c6555fd90b6ad1f11349	0	0	0	1	0	0	0	0	0
05d85e3fee09c6555fd90b6ad1f11349	0	0	0	0	0	0	1	0	0
069f31fa2b0d27cd202cc1f3befd6b2d	0	0	0	0	0	0	1	0	0
069f31fa2b0d27cd202cc1f3befd6b2d	0	0	0	0	0	0	0	0	0
069f31fa2b0d27cd202cc1f3befd6b2d	0	0	0	0	0	1	0	0	0
069f31fa2b0d27cd202cc1f3befd6b2d	0	0	0	0	0	0	0	1	0
07f804aa41d6842efd8da9e9a8eeaeed	0	0	0	0	0	0	1	0	0

Figura 22 - Trecho de matriz resultante do mapeamento de arquivo de log

A Figura 22 mostra um trecho do arquivo de log filtrado, já mapeado para os objetos da ontologia. Os objetos com valor “1” indicam sua requisição, os com valor “0”, indicam que não foram requisitados.

O mapeamento dos arquivos de log filtrados é feito diariamente, assim que ocorre a filtragem, de acordo com a configuração de horário. Existe, entretanto, a opção de geração de um arquivo com os dados mapeados, contendo dados de vários dias, para facilitar a posterior análise.

Visando a facilidade de instalação, optou-se pelo desenvolvimento da aplicação sem a utilização de um Sistema Gerenciador de Banco de Dados. Optou-se, ao invés disso, pela

utilização de arquivos do tipo “XML”, que podem ser manipulados e visualizados de maneira simples.

A aplicação possui, ainda, um formulário onde é possível fazer a visualização dos arquivos “XML”.

A próxima seção apresenta um estudo de caso de mineração dos dados obtidos. O objetivo da análise realizada é verificar a possibilidade de utilização dos mesmos na melhoria dos serviços oferecidos pelo site.

8 MINERAÇÃO DOS DADOS

A mineração da Web pode ser uma importante ferramenta no auxílio às tarefas de manutenção e projeto de sites. Esta seção descreve a aplicação de métodos de mineração aos dados obtidos no presente trabalho. O objetivo desta mineração é mostrar a utilidade destes dados na verificação das sessões dos usuários. São aplicados métodos de agrupamento (*clustering*) aos dados, na intenção de classificá-los de acordo com os tipos de sessões existentes.

Foram considerados os dados relativos ao acesso dos usuários ao site, em um período de três semanas, sendo a primeira delas referente ao prazo para matrícula dos alunos, e as outras duas o período permitido aos mesmos para efetuarem cancelamentos de matrícula. Considerando-se, portanto, tal período como crítico para o funcionamento do sistema, uma vez que o número de acessos cresce consideravelmente, deseja-se descobrir, através da analise da utilização, quais as reais necessidades de melhoria no site, bem como os serviços mais importantes e aqueles que, caso necessário, podem ser desabilitados, no período, visando um melhor desempenho.

8.1 Pré-Processamento dos dados

A etapa de pré-processamento dos dados coletados visa prepará-los para que os métodos de Mineração de Dados sejam aplicados. De acordo com os tipos de questões que se deseja responder, os dados podem ser tratados de maneiras diferentes, voltando-se para a abordagem desejada.

Deseja-se, neste trabalho, identificar o tipo de utilização que está sendo feita no sistema através da análise das sessões feitas pelos usuários, ou seja, cada visita do usuário ao site é analisada como um todo. Desenvolveu-se, então, uma transformação, no sentido de que sejam analisadas as sessões inteiras e não somente seus registros.

O primeiro passo para a transformação foi o agrupamento das requisições correspondentes a cada sessão. Os dados são, portanto, ordenados pelo seu identificador de sessão e, posteriormente, pelo horário. Desta maneira, cada sessão tem suas requisições agrupadas na ordem em que aconteceram, podendo-se, portanto, verificar qual a seqüência de passos seguida pelo usuário.

Sendo uma sessão composta por *m* requisições, e cada requisição representada por um vetor com dimensão *n*, é possível que esta sessão seja representada, então, por uma matriz M de dimensões *m* x *n*.

$$M = \begin{bmatrix} x_{11} & x_{12} & \Lambda & x_{1n} \\ x_{21} & x_{22} & \Lambda & x_{2n} \\ \text{M} & \text{M} & \Lambda & \text{M} \\ x_{m1} & x_{m2} & \Lambda & x_{mn} \end{bmatrix}$$

Para cada sessão, representada pela matriz M, é possível calcular o vetor R, de dimensão *n*, da seguinte forma:

$R = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \Lambda & \bar{x}_n \end{bmatrix}$, sendo $\bar{x}_i = sign\left(\sum_{j=1}^{m} x_{ij} \right)$, tal que $x_{ij}, \bar{x}_i \in \{0,1\}$ e função sinal definida por $sign(x) = \begin{cases} 1, x > 0 \\ 0, x = 0 \end{cases}$.

O resultado final será, portanto, uma matriz cujo número de colunas é igual ao número de variáveis dos vetores, e o número de linhas será igual ao número de sessões em estudo.

Id	Doc_Freq	Doc_Disc	Resp_Matric	Resp_Canc	Resp_Acomp	Funcoes	Form_Matric	Form_Canc	Form_Do
00a249ba8d760005ed9d148f39a364db	0	0	0	0	0	1	0	0	0
027a42e5d0cda95e27798c20111 0c84e	0	0	1	0	1	1	1	0	0
05d85e3fee09c6555fd90b6ad1f11349	0	0	1	0	0	1	1	0	0
069f31fa2b0d27cd202cc1f3befd6b2d	0	0	0	0	1	1	1	0	0
07f804aa41d6842efd8da9e9a8eeaeed	0	0	1	0	0	1	1	0	1
0b68879ab1cd12d95cbeeb320a3dbeea	0	0	0	0	1	1	0	0	0
0cff89fde83783ba9b1ce291ef772b16	0	0	1	0	0	1	1	0	0
0e64db9f8903a81bc52b984958bf756b	0	0	1	0	1	1	1	1	1
1165f752c062a5842d1cc3e29520211f0	0	0	0	0	0	1	0	0	0
16313dd1549d803e9e326be30c053fca	0	0	0	0	0	1	1	0	0
189e615f78050d2ba03d9f1d5726f1d3	0	0	0	0	0	1	1	0	0
1ac3ace020e31cb7253bfd867a9ffedc	0	0	0	0	0	1	1	0	0
1bd7f5e47b500c00ba1023af5404aed3	0	0	1	0	0	1	1	0	0
1c208443bc94e7f87f1097b0d8e52832	0	0	1	0	0	1	1	0	1
1fbe8f6ded3f9fe9b3d1515c0afa333d	0	0	0	0	0	1	0	0	0
2fc484da162747ef189b94f6b0bda9f3	0	0	1	0	1	1	1	0	0
322dc0fc92646b628151949aa04e90b0	0	0	1	0	0	1	1	0	0
3a5ae602f25cc3b246c59b0a7407c535	0	0	1	0	0	1	1	0	0
3c93529c9ca273b04c2800755d672e1a	0	0	0	0	0	1	1	0	0
3e95bed96d9a11a7b4c6bc7a2d514b82	0	0	1	0	1	1	1	0	0
3eb7278f3c4ea16556ed1c1b45b46ec8	0	0	0	0	0	1	1	0	0
3fda541693062b0c8e129c15d138d6e5	0	0	1	0	0	1	1	0	0
42352989558fff2b55c58e8c6b97b035	0	0	1	0	1	1	1	0	0
45ba1266fe1603c34ed34cfcf63ea6c5	0	0	0	0	0	1	0	0	0
489f5369bfe336ecdde33b1ea9063bbe	0	0	1	0	0	1	1	0	0
4b6b7e50b39a8d810da2596a9f4caf70	0	0	1	0	1	1	1	1	0
4d779bb3aef3fef7a16e3e5e251e32d8	0	0	1	0	0	1	1	0	1
4eda2fef8a8cf6b0febec5704b8931f5	0	0	0	0	1	1	1	0	0
502cb53971dfc9f185846ceaa55f7823	0	0	0	0	0	1	0	0	0

Figura 23 – Exemplo dos vetores referentes às sessões

A Figura 23 representa um trecho da matriz resultante da transformação, onde se pode observar que o campo *id* é diferente para cada linha, ou seja, cada sessão. Além disso, pode-se

perceber que em algumas linhas, tem-se valor “1” em mais de uma coluna, o que indica a união dos vetores de requisições.

Tendo sido feita esta transformação, é possível que estes dados sejam submetidos a uma análise, através de métodos de mineração, para que seja possível identificar tipos de sessões com características comuns, classificando-as.

8.2 Mineração dos dados

A realização da análise dos dados, utilizando técnicas de Mineração de Dados, busca, neste trabalho, a identificação de padrões nas sessões de usuários do site. Pretende-se, então, detectar possíveis problemas encontrados pelos usuários durante as visitas, verificando a necessidade de manutenção e atualizações de projeto.

A análise tem, ainda, a função de validar modificações, anteriormente efetuadas pelos projetistas, com base em solicitações por parte da secretaria do Curso, disponibilizando novos serviços aos usuários do site.

De posse dos dados pré-processados, foram aplicadas, então, técnicas de agrupamento (*clustering*), desejando-se identificar os tipos de sessões, separando-as em grupos de acordo com suas características em comum.

O conjunto de dados submetido à análise corresponde aos registros das visitas dos usuários ao site em três semanas. Os registros compreendem um total de 4.745 acessos aos objetos mapeados. O pré-processamento dos dados, unindo-se as requisições de cada sessão, resultou em um total de 773 observações, correspondendo, portanto, ao número de sessões.

Em um primeiro momento, foi utilizado o algoritmo de Junção (Joining (Tree Clustering)), que é um método hierárquico aglomerativo, onde cada observação inicia sendo um grupo, e com o decorrer do processo os grupos vão sendo agrupados até que o resultado final seja um único grupo, contendo todas as observações. Com a aplicação deste método foi possível a identificação de 4 grupos.

Identificado o número de grupos esperado, utilizou-se o método das **k-médias**, que é um método de partição, cujo número de agrupamentos em que se deseja dividir os dados deve ser pré-estabelecido. A Figura 24 representa graficamente os agrupamentos (*clusters*) obtidos com a aplicação do método das k-médias. O eixo X representa os objetos em que os acessos foram mapeados, enquanto que o eixo Y representa os valores médios. Uma vez que os dados

submetidos à análise apresentavam apenas valores “0” e “1”, os valores médios ficaram dentro do intervalo [0..1]. Um objeto cujo valor médio, em determinado agrupamento, seja próximo a “1”, é visto, portanto, como um objeto de importância relevante, uma vez que a maioria dos valores para este objeto, nas sessões compreendidas no cluster, tem valor igual a “1”.

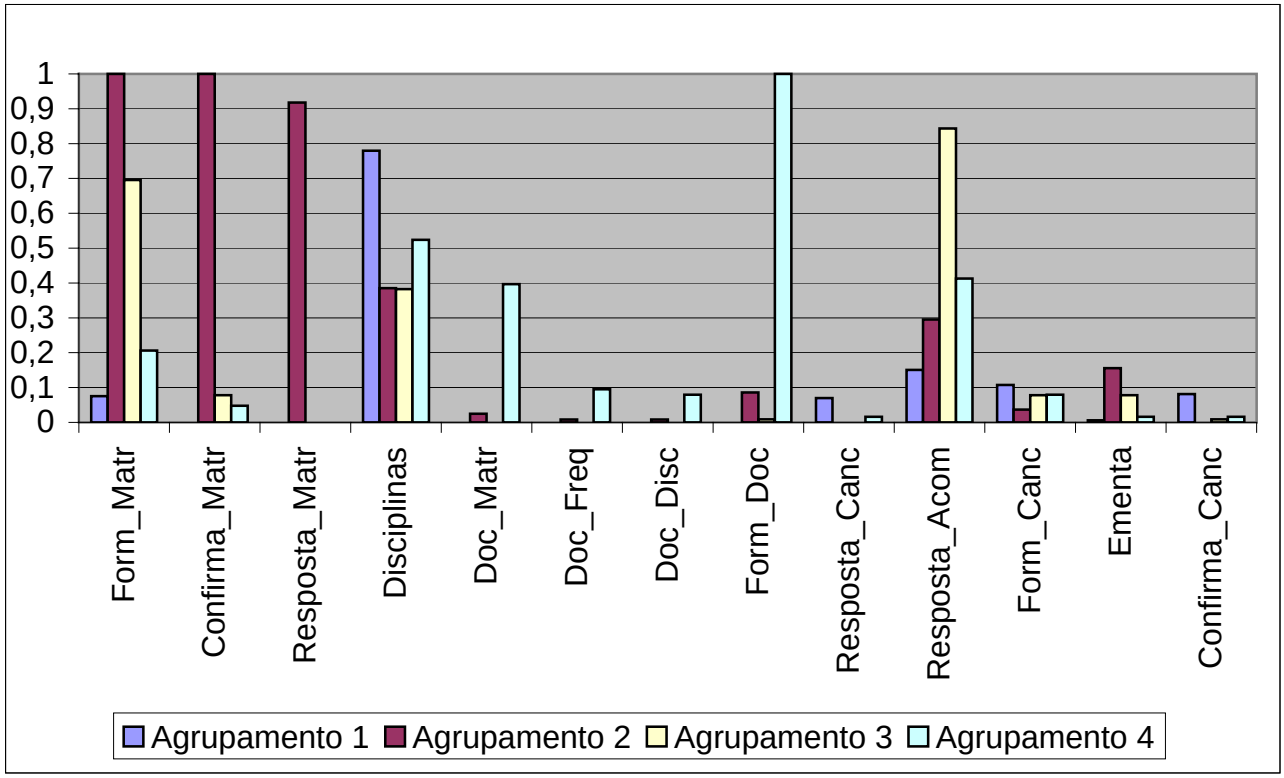

Figura 24 – Valores médios dos objetos nos agrupamentos

O Agrupamento 1 apresenta um valor médio alto apenas para o objeto “Disciplinas”, em torno de 0,7797. Isto indica uma característica das observações contidas neste grupo, onde o este objeto é acessado em cerca de 77,97% dos casos. O número de casos envolvidos neste agrupamento representa cerca de 45% do total.

No Agrupamento 2, pode-se perceber os maiores valores médios para os objetos “Form_Matr”, “Confirma_Matr” e “Resposta_Matr”, todos próximos a 1. A ocorrência de tais valores demonstra que as observações envolvidas neste agrupamento têm, como um padrão, a característica de acesso a estes 3 objetos. O número de casos representa cerca de 32% do total das observações.

Os maiores valores médios para o Agrupamento 3 correspondem aos objetos “Resposta_Acom” (acima de 0,8) e “Form_Matr” (cerca de 0,7). O número de observações, neste grupo, equivale a cerca de 15% dos casos.

Para o Agrupamento 4, o objeto com maior valor médio é o “Form_Doc”, igual a 1. Podem ser observados, ainda, valores próximos a 0,4 para os objetos “Doc_Matr”, “Resposta_Acom”, e 0,5 para o objeto “Disciplinas”. Cerca de 8% das observações estão compreendidas neste agrupamento.

8.3 Avaliação dos Resultados

De posse dos resultados, obtidos na aplicação do método das **k-médias**, deseja-se, então, fazer uma análise nos mesmos, a fim de identificar as características de cada um dos agrupamentos, bem como seu significado.

A Tabela 3 representa os valores médios, desvio-padrão e variância de cada objeto do Agrupamento 1.

Variáveis	**Média**	**Desvio-Padrão**
Ementa	0,005797	0,076028
Doc_Matr	0,000000	0,000000
Doc_Freq	0,000000	0,000000
Doc_Disc	0,000000	0,000000
Resposta_Matr	0,000000	0,000000
Resposta_Canc	0,069565	0,254782
Resposta_Acom	0,150725	0,358300
Form_Matr	0,075362	0,264359
Form_Canc	0,107246	0,309876
Form_Doc	0,000000	0,000000
Disciplinas	0,779710	0,415044
Confirma_Matr	0,000000	0,000000
Confirma_Canc	0,081159	0,273476

Tabela 3 – Valores médios no Agrupamento 1

Os valores médios para os objetos, no Agrupamento 1, demonstram que, em sua maioria, as observações correspondem a sessões onde os usuários desejam acessar o objeto “Disciplinas”. Este objeto refere-se ao serviço através do qual os alunos têm acesso a um documento, onde consta a lista de disciplinas que já cursou, ou está matriculado, com os conceitos disponíveis.

O serviço correspondente à lista de disciplinas e conceitos dos alunos é recente no site. A alternativa anterior era a solicitação de um Atestado de Disciplinas Cursadas à secretaria do curso, para posterior retirada. Como o documento não era emitido automaticamente, muitos eram os casos em que os alunos faziam solicitações informais para ficarem a par de seus conceitos, causando um transtorno maior à secretaria, principalmente em períodos próximos aos de matrícula.

Informados do problema, os responsáveis pelo projeto e manutenção do site sugeriram a disponibilização de um documento, sem valor legal, através do qual fosse possível aos alunos obter seus conceitos com facilidade. O serviço foi criado, portanto, a partir de uma verificação informal de um problema, sem que se tenha, até então, uma verificação real de sua importância.

A análise do Agrupamento 1 sugere um padrão de sessões onde o objetivo principal dos usuários é o acesso ao serviço correspondente à lista de disciplinas cursadas e seus conceitos. Pode-se afirmar, portanto, que a verificação da necessidade da criação do serviço foi válida, uma vez que o número de sessões com este objetivo é bastante expressivo, sendo responsável pelo agrupamento com maior número de observações.

Também é interessante notar que o valor médio para os objetos “Form_Doc”, “Doc_Matr”, “Doc_Freq”, “Doc_Disc” é “0” para este agrupamento. Estes objetos são referentes aos serviços de solicitação de documentos junto à secretaria. Percebe-se, então, que nenhuma das observações refere-se a este tipo de serviço.

A Tabela 4 mostra os valores médios, desvio-padrão e variância para os objetos das observações contidas no Agrupamento 2.

Variáveis	Média	Desvio-Padrão
Ementa	0,155738	0,363352
Doc_Matr	0,024590	0,155191
Doc_Freq	0,008197	0,090349
Doc_Disc	0,008197	0,090349
Resposta_Matr	0,918033	0,274879
Resposta_Canc	0,000000	0,000000
Resposta_Acom	0,295082	0,457017
Form_Matr	1,000000	0,000000
Form_Canc	0,036885	0,188867
Form_Doc	0,086066	0,281037
Disciplinas	0,385246	0,487654
Confirma_Matr	1,000000	0,000000
Confirma_Canc	0,000000	0,000000

Tabela 4 – Valores médios no Agrupamento 2

No Agrupamento 2, os objetos “Form_Matr”, “Confirma_Matr” e “Resposta_Matr” apresentam valores altos. Os dois primeiros têm valor médio igual a 1, indicando que todas as observações do grupo têm valor 1, e o último tem valor médio 0,918, representando 91,8% das observações com valor 1 para este objeto.

O serviço relativo ao procedimento de matrícula é composto por um formulário, onde o aluno seleciona as disciplinas em que deseja matricular-se, uma confirmação, no momento em que o aluno submete o pedido, e uma resposta, um documento comprovante de sua matrícula. Estes componentes são mapeados para os objetos “Form_Matr”, “Confirma_Matr” e “Resposta_Matr”, respectivamente.

Analisando-se os resultados para o Agrupamento 2, é possível verificar que, em todos os casos, o usuário acessou o formulário de matrícula e confirmou sua solicitação de matrícula. O comprovante de matrícula (objeto “Resposta_Mat”), entretanto, não foi acessado em 8,2% das observações do agrupamento, ao contrário do esperado.

O documento, ao qual o objeto “Resposta_Mat” faz referência, é gerado de maneira dinâmica, no formato PDF e, portanto, o tempo de descarga do documento para o usuário é maior

do que aquele necessário para as demais páginas. Assim, é possível que usuários que façam acesso ao sistema através de uma conexão mais lenta tenham dificuldades neste ponto, deixando de receber seu documento comprovante de matrícula.

O Agrupamento 2, então, pode ser identificado como sendo aquele em que as sessões são referentes aos procedimentos de matrícula efetuados. Isto indica que, neste agrupamento, não existem observações referentes às semanas destinadas ao cancelamento de matrícula, o que pode ser confirmado pelos objetos “Confirma_Canc” e “Resposta_Canc”, que têm valor médio igual a “0”.

O fato de que, em alguns casos, o documento de resposta não é acessado será encaminhado à equipe de projeto e manutenção do site.

A Tabela 5 contém o valor médio, desvio-padrão e variância de cada um dos objetos das observações contidas no Agrupamento 3.

Variáveis	**Média**	**Desvio-Padrão**
Ementa	0,078261	0,269757
Doc_Matr	0,000000	0,000000
Doc_Freq	0,000000	0,000000
Doc_Disc	0,000000	0,000000
Resposta_Matr	0,000000	0,000000
Resposta_Canc	0,000000	0,000000
Resposta_Acom	0,843478	0,364939
Form_Matr	0,695652	0,462144
Form_Canc	0,078261	0,269757
Form_Doc	0,008696	0,093250
Disciplinas	0,382609	0,488151
Confirma_Matr	0,078261	0,269757
Confirma_Canc	0,008696	0,093250

Tabela 5 – Valores médios no Agrupamento 3

No Agrupamento 3, os objetos cujo valor médio é maior são o “Resposta_Acom”, com valor 0,843 e “Form_Matr”, com valor 0,695.

Em um primeiro momento não se identificou uma causa para esta relação, uma vez que o objeto “Form_Matr” corresponde à parte do procedimento de matrícula, onde os usuários

selecionam as disciplinas que desejam cursar. O objeto “Resposta_Acom”, entretanto, corresponde ao serviço de documentos referente ao acompanhamento de matrícula, que contém as informações da situação atual de sua matrícula, e as disciplinas que está apto a cursar.

A Figura 25 mostra a aparência da página inicial do site, onde o usuário faz a opção pelo serviço que deseja realizar.

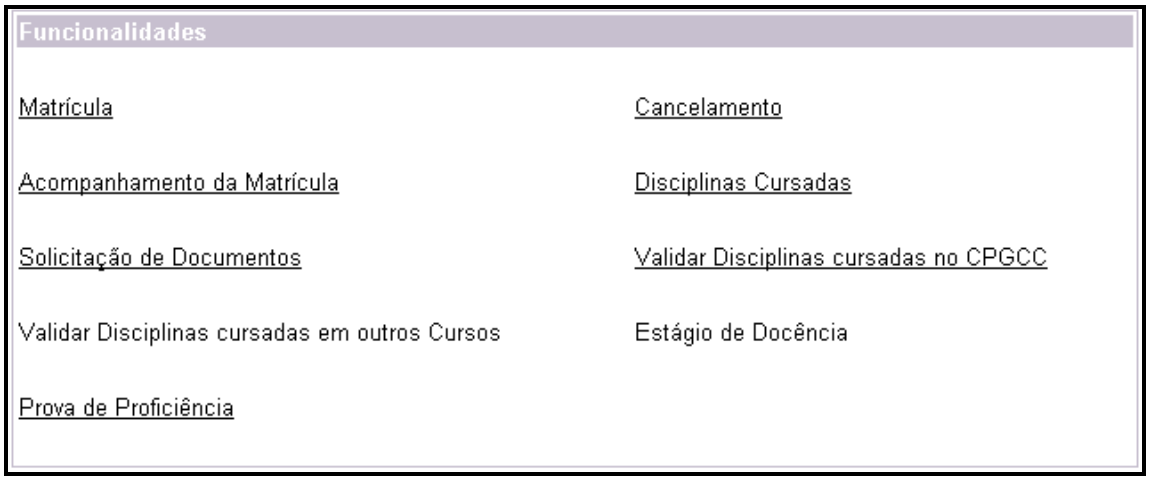

Figura 25 – Opções de serviços para realização no site

Uma vez que os objetos cujos valores são expressivos no agrupamento são relativos a serviços diferentes, é possível que esteja havendo, então, um problema de navegação, causado por falha ergonômica ou clareza na terminologia utilizada nas páginas.

Analisando-se a forma como estão dispostas as opções de serviços a serem realizados, pode-se concluir que, em muitos casos, os usuários estão sendo induzidos a entrar no link “Matrícula”, que na verdade leva ao formulário para início do procedimento de matrícula, quando desejam, na verdade, verificar sua atual situação perante as disciplinas escolhidas para cursar. Assim, é necessário que retornem à página de escolha dos serviços para, então, acessar o link “Acompanhamento de Matrícula”.

Verifica-se, ainda, que não existem solicitações de documentos à secretaria, uma vez que o valor médio para “Doc_Matr”, “Doc_Freq” e “Doc_Disc” é igual a 0.

Na Tabela 6 são apresentados o valor médio, desvio-padrão e a variância para os objetos das observações do Agrupamento 4.

Variáveis	Média	Desvio-Padrão
Ementa	0,015873	0,125988
Doc_Matr	0,396825	0,493169
Doc_Freq	0,095238	0,295901
Doc_Disc	0,079365	0,272479
Resposta_Matr	0,000000	0,000000
Resposta_Canc	0,015873	0,125988
Resposta_Acom	0,412698	0,496274
Form_Matr	0,206349	0,407935
Form_Canc	0,079365	0,272479
Form_Doc	1,000000	0,000000
Disciplinas	0,523810	0,503444
Confirma_Matr	0,047619	0,214669
Confirma_Canc	0,015873	0,125988

Tabela 6 – Valores médios no Agrupamento 4

No Agrupamento 4, o objeto que apresenta maior valor médio é o “Form_Doc”, com valor igual a 1. Além desse, os objetos com valores médios mais altos são “Disciplinas”, com valor 0,523, “Resposta_Acom”, com valor 0,412 e “Doc_Matr”, com valor 0,396.

O comportamento que pode ser identificado no grupo é que todos os casos têm relação com serviços de documentos. O objeto “Form_Doc” refere-se ao formulário de solicitação de documentos à secretaria do curso. Os demais valores médios para os objetos demonstram que, em sua maioria, os usuários acessam serviços de documentos, mas não somente os de solicitação à secretaria, uma vez que os objetos “Disciplinas” e “Resposta_Acom” apresentam valores em torno de 0,523 e 0,412, respectivamente.

A análise deste grupo pode sugerir que buscas por documentos como “acompanhamento de matrícula” e “disciplinas cursadas” estejam sendo feitas equivocadamente através do link para “solicitação de documentos”.

Em função dos resultados obtidos e de sua análise, é possível inferir que o comportamento dos usuários, no site estudado, revela possíveis necessidades de mudança na organização e apresentação do conteúdo das páginas. A mudança na ordem em que as opções de serviço são

apresentadas aos usuários pode ser interessante, no sentido de facilitar a navegação. Nomenclaturas inadequadas, utilizadas para os *hyperlinks*, podem induzir os usuários a possíveis enganos durante as visitas ao site. Nomes como “Acompanhamento de Matrícula”, por exemplo, merecem uma atenção especial, uma vez que os usuários costumam tentar solicitar o serviço através de outros caminhos.

A análise aqui apresentada serve como um exemplo da possibilidade de aplicação de técnicas de Mineração da Utilização da Web aos dados, obtidos através da aplicação do modelo proposto. Outros tipos de análise podem, ainda, ser feitos, levando-se em conta, por exemplo, a seqüência das requisições dos usuários nas sessões.

Na próxima seção são apresentadas as conclusões a respeito do trabalho desenvolvido, bem como algumas sugestões para trabalhos futuros.

9 CONCLUSÕES E TRABALHOS FUTUROS

9.1 Conclusões

A Mineração da Utilização da Web (Web Usage Mining) aplica técnicas de Mineração de Dados aos registros de utilização da Web. A Análise da Seqüência de Requisições (ClickStream Analysis), é uma forma de obtenção dos dados, uma vez que essa seqüência de cliques é literalmente um registro de cada gesto efetuado por cada visitante a cada site da Web. (KIMBALL, 2000).

Os arquivos de log, gerados pelos Servidores Web, são uma fonte de dados para o *clickstream*, uma vez que registram todas as requisições a documentos e páginas feitas pelos usuários dos sites neles hospedados. A utilização do conceito de Servidores Nulos auxilia no processo de obtenção dos dados, proporcionando uma forma de identificação das requisições em aqueles pontos pré-definidos como alvos de estudo, ainda que a hospedagem das páginas e documentos esteja distribuída em vários servidores.

A utilização de uma ontologia para representação do conhecimento envolvido no site, proporciona seu entendimento sob o ponto de vista do seu conteúdo, e não da estrutura e ligação entre as páginas. Desta maneira, torna-se possível a extensão e reaproveitamento deste conhecimento para melhoria do sistema, a utilização em outras aplicações ou a compreensão de qual o conhecimento que o usuário esta procurando no site.

O enriquecimento semântico dos arquivos de log proporciona uma melhoria na qualidade dos dados, uma vez torna possível a identificação precisa dos pontos que estão sendo submetidos à análise. Os arquivos passam a conter dados a mais do que o formato padrão, tornando-se portanto mais ricos em detalhes do seu conteúdo e, em conseqüência, mais úteis ao processo de análise.

Foi proposta uma modelagem da obtenção de dados para mineração da utilização da Web, baseada em uma ontologia, representando o conhecimento contido no site. Sugere-se a inclusão de conteúdo semântico aos documentos e páginas, na forma de rótulos que os identifiquem. É utilizado o conceito de servidor de conteúdos nulos para a captura das requisições dos usuários, através de arquivos de log enriquecidos pelo conteúdo semântico. Os dados obtidos na captura

são, posteriormente, mapeados para os objetos da ontologia, e dispostos na forma de um vetor que representa o objeto acessado em cada requisição.

O modelo foi aplicado em um site de controle acadêmico, cujos dados foram coletados em um período de três semanas. Foram aplicados, posteriormente, métodos de agrupamento (*clustering*), onde foi possível a identificação de 4 agrupamentos, pelo tipo se sessões dos usuários.

A aplicação das técnicas de Mineração da Utilização da Web, em nosso estudo, demonstrou sua validade no sentido de identificação de padrões de utilização. A análise dos resultados possibilitou a constatação da necessidade de possíveis modificações na apresentação e conteúdo das páginas, a fim de melhorar a qualidade dos serviços oferecidos. Identificou-se, por outro lado, a validade de modificações realizadas anteriormente, a partir de constatações dos projetistas e mantenedores, no sentido de disponibilizar novos serviços aos usuários e reduzir, assim, o volume de trabalho da secretaria do curso.

Desta maneira, pode-se verificar a validade da modelagem para obtenção dos dados, aqui descrita, uma vez que os resultados obtidos em sua aplicação vêm a auxiliar na manutenção e projeto do site estudado. A análise de agrupamentos aplicada foi, apenas, uma forma de demonstrar a possibilidade de descoberta de informações, a partir dos dados coletados.

9.2 Trabalhos futuros

Algumas sugestões podem ser feitas, ainda, para a realização de trabalhos futuros. Uma vez que a aplicação implementada visa, apenas, o auxílio no processo de coleta e transformação dos dados, não se buscou, em seu desenvolvimento, uma preocupação no sentido de análise. Sugere-se, portanto a implementação de interfaces que forneçam, por exemplo, uma representação gráfica, da seqüência de requisição nas sessões dos usuários.

Quanto à utilização dos dados obtidos, pode-se sugerir análises que levem em conta o tempo envolvido nas sessões, pois é uma variável que pode revelar outros padrões de utilização, como. Exemplificando, objetos que demandem de um período de tempo maior para serem compreendidos podem expressar a dificuldade de aprendizado que representam aos usuários.

A personalização de sites, de acordo com o perfil do usuário, pode ser implementada com base nos dados obtidos. A geração de regras de associação é uma alternativa cuja aplicação pode ser interessante. Através desta, é possível que sejam descobertas novas relações entre os

conteúdos. Um exemplo é a criação de sistemas de recomendação de produtos ou conteúdos em geral, baseados no histórico de utilização dos usuários com comportamentos semelhantes, ou mesmo na identificação do usuário. Nestes sistemas, os usuários recebem indicações de conteúdos que, de acordo com seu próprio histórico ou de usuários com perfil semelhante, são normalmente acessados.

10 REFERÊNCIAS BIBLIOGRÁFICAS

AMARAL, Fernanda C.; AMARAL, Eduardo. **Minería de Datos Aplicada al Mercado Corporativo**. VII Congresso Internacional de Ciencias de la Computación, CICC. Cochabamba, BO, 2002.

ANDERSEN, Jesper; LARSEN, Rune S.; GIVERSEN, Anders et al. **Analyzing Clickstreams Using Subsessions**. DOLAP'00, McLean, VA, USA, 2000.

BUSSAB, Wilton de O.; MIAZAKI, Édina S.; ANDRADE, Dalton F. **Introduçao à Análise de Agrupamentos**, Associação Brasileira de Estatística, 9º Simpósio Nacional de Probabilidade e Estatística. São Paulo, 1990.

BERENDT, Bettina; HOTHO, Andreas; STUMME, Gerd. **Towards Semantic Web Mining.** Horrocks and J. Hendler (Eds.), © Springer-Verlag Berlin Heidelberg, 2002.

BERRY, Michael J.; LINOFF, Gordon. **Data Mining Techniques: For Marketing, Sales, and Customer Support**. 1 ed., New York, USA, Wiley Computer Publishing, 1997.

BORGES, José; LEVENE, Mark. **Data mining of user navigation patterns**. In Web Usage Analysis and User Profiling, pp. 92-111. © Springer-Verlag as Lecture Notes in Computer Science, Vol. 1836, 1999.

BOULLOSA, José R. **Um ambiente para Mineração da Utilização da Web**. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, 2002.

CASERTA, Joe. (2001) **Clickstream Data Mart**. Intelligent Enterprise, Dezembro de 2001. Disponível em: www.intelligententerprise.com/011205/418warehouse1_1.shtml - acesso em Janeiro de 2003.

CHAKRABARTI, Soumen. **Data mining for hypertext: A tutorial survey**. SIGKDD Explorations,1:1–11, 2000.

CHERKASSKY, V., MULIER, F. **Learning From Data – Concepts, Theory and Methods**, 1 ed., New York, USA, John Wiley & Sons, Inc. 1998.

COOLEY, Robert; MOBASHER, Bamshad; SRIVASTA, Jaideep. **Data Preparation for Mining World Wide Web Browsing Patterns**. Knowledge and Information System © Springer-Verlag, 1999.

DEOGUN, J. S. , RAGHAVAN, V. V., SARKAR, A., et al., **Data mining: Research trends, challenges, and applications**. In: Roughs Sets and Data Mining: Analysis of Imprecise Data, pp. 9-45, Boston, MA, USA, Kluwer Academic Publishers. 1997.

EVERITT, B. S. **Cluster Analysis**. Third edition, Edward Arnold, London, 1993.

FAYYAD, Usama M; PIATETSKY-SHAPIRO, Gregory; PADHRAIC, Smyth et al. **Advances in knowledge discovery and data mining**. MIT-AIII Press, 1996.

FREITAS, Frederico L.; BITTENCOURT, Guilherme, CALMET, Jacques. **MASTER-Web: An Ontology-based Internet Data Mining Multi-Agent System**. International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet (SSGRR-2001), Scuola Superiore G. Reiss Romoli, L'Aquila, Itália, 2001.

GROTH, Robert. **Data Mining: a hands-on approach for business professionals**. Prentice Hall. New Jersey, 1998.

GRUBER, Tom. R. **Toward principles for the design of ontologies used for knowledge sharing**. International Workshop on Formal Ontology, Padova, Italy. 1993. Available as technical report KSL-93-04, Knowledge Systems Laboratory, Stanford University.

HONG, Jason; LANDAY, James. **WebQuilt: A Framework for Capturing and Visualizing the Web Experience**. Proceeding of WWW 10, Hong Kong, May 2001.

KIMBALL, Ralph; MERZ, Richard. **Data Webhouse: construindo o Data Warehouse para a Web**. Editora Campus, Rio de Janeiro, Brasil, 2000.

KIMBALL, Ralph. **The Special Dimensions Of The Clickstream**. Intelligent Enterprise, Janeiro de 2000, Disponível em: www.intelligententerprise.com em Novembro de 2002.

KOHAVI, Ron. **Mining E-Commerce Data: The Good, the Bad, and the Ugly**. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001

KOSALA, Raymond; BLOCKEEL, Hendrik. **Web Mining Research: A Survey**. SIGKDD Explorations, ACM SIGKDD, 2000.

LI, Richard. **Clickstream Data Warehousing**. ArsDigita Systems Journal, Agosto de 2000, Disponível em: www.redhat.com/asj/clickstream/ em Janeiro de 2003.

NASRAOUI, Olfa; FRIGUL, Hichem; JOSHI, Anupam et al. **Mining Web Access Logs Using Relational Competitive Fuzzy Clustering**. Eight International Fuzzy Systems Association World Congress - IFSA 99, 1999.

NOY, Natalya F.; MCGUINNESS, Deborah. **Ontology Development 101: A Guide to Creating Your First Ontology**. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

OBERLE, Daniel; BERENDT, Bettina; HOTHO, Andreas et al. **Conceptual User Tracking**, Web Intelligence, First International Atlantic Web Intelligence Conference, AWIC 2003, Madrid, Spain, May 5-6, Proceedings (pp. 155-164). Berlin: Springer, LNCS 2663, 2003.

PIATESTKY-SHAPIRO, G. **Knowledge Discovery in Databases: 10 years after**. SIGKDD Explorations, v. 1, n. 2 (Jan), 2000.

QUINLAN, J. R. **C4.5: Programs for machine learning**. San Mateo, CA, USA, Morgan Kaufmann, 1993.

SILVERSTON, Len. **Universal Data Models for Clickstream Analysis**. DM Review, Janeiro de 2002. Disponível em: www.dmreview.com, em Março de 2003.

SPILIOPOULOU, M.; FAULSTICH, L. C. **WUM: A Web Utilization Miner**. In: Proceedings of the EDBT Workshop, WebDB98, LNCS - Lecture Notes in Computer Science, v. 1590, Springer-Verlag, Valencia, Spain, 1998

STUMME, Gerd; HOTHO, Andreas; BERENDT, Bettina. **Usage Mining for and on the Semantic Web**. Proceedings of NSF Workshop, Baltimore, 77-86, Nov. 2002.

SWEIGER, Mark. **Web Server Log Files: The Ultimate Clickstream Consulting**. Clickstream Consulting, 2002. Disponível em: http://www.clickstreamconsulting.com em Janeiro de 2003.

TINGSHAO, Zhu; GREINER, Russ; HAEUBL, Gerald. **Predicting Where a Web User Wants to Go**. CHI2003, Workshop on Best Practices and Future Visions for Search User Interfaces, Florida, USA, 2003.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining – Pratical Machine Learning Tools and Techniques with Java Implementations**", Morgan Kaufmann Publishers, USA, 2000.

ZAÏANE, Osmar. R.; XIN, Man; HAN, Jiaewi. **Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs**. In: Proceedings of Advances in Digital Libraries Conference (ADL'98), pp. 19-29, Santa Barbara, CA, USA, 1998.