COPPE/UFRJ

METODOLOGIA DE MINERAÇÃO DE DADOS APLICADA A NAVEGAÇÃO DE DISPOSITIVOS MÓVEIS

Luis Carlos Couto de Souza

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Civil.

Orientador: Nelson Francisco Favilla Ebecken

Rio de Janeiro
Setembro de 2009

METODOLOGIA DE MINERAÇÃO DE DADOS APLICADAA NAVEGAÇÃO DE DISPOSITIVOS MÓVEIS

Luis Carlos Couto de Souza

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

Aprovada por:

_____

Prof. Nelson Francisco Favilla Ebecken, D. Sc.

_____

Prof[a]. Beatriz de Souza Leite Pires de Lima, D. Sc.

_____

Prof[a]. Myrian Christina de Aragão Costa, D. Sc.

RIO DE JANEIRO, RJ – BRASIL

SETEMBRO DE 2009

Souza, Luis Carlos Couto de

Metodologia de mineração de dados aplicada a navegação de dispositivos móveis/ Luis Carlos Couto de Souza. – Rio de Janeiro: UFRJ/ COPPE, 2009.

X, 64 p.: il; 29,7 cm.

Orientador: Nelson Francisco Favilla Ebecken

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2009.

Referências Bibliográficas: p. 62-64.

1. Mineração de dados da Web. 2. Metodologia de mineração de dados aplicada navegação de dispositivos móveis. I. Ebecken, Nelson Francisco Favilla. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Titulo.

AGRADECIMENTOS

Gostaria de agradecer a meus pais, Leila e Pedro, por muitas vezes terem se privado de várias coisas para proporcionarem uma melhor educação para eu e meus irmãos.

A Aline Regina e Lucas, minha família, que sempre estiveram comigo me apoiando em todos os momentos, obrigado pelo incentivo e carinho.

A Valmir Sobral pelo apoio e incentivo.

A todos os meus amigos e colegas que me auxiliaram, incentivaram e apoiaram ao longo do meu mestrado e minha dissertação

A VIVO e aos meus gestores que flexibilizaram meu período de trabalho permitindo que eu realizasse este curso e elaborasse esta dissertação.

Ao professor Nelson Ebecken, por ter me permitido estudar em uma das melhores instituições de ensino do mundo e me orientar e apoiar ao longo do curso e desta dissertação.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

METODOLOGIA DE MINERAÇÃO DE DADOS APLICADA A NAVEGAÇÃO DE DISPOSITIVOS MÓVEIS

Luis Carlos Couto de Souza

Setembro / 2009

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

O acirramento da concorrência no mercado de telefonia móvel brasileiro fez com que as operadoras buscassem diferencial em serviços de valor agregado. Dentre estes serviços destaca-se o serviço de WAP, que propicia o acesso a *web* a partir de telefones móveis. O serviço WAP baseia-se nos portais das operadoras, que por padrão são as páginas iniciais dos clientes, provendo *hyperlinks* para sites de provedores de conteúdo. O objetivo do trabalho consiste em, para o cliente, disponibilizar o acesso a informação no WAP de forma ágil e, para operadora, uma opção para incrementar a receita do serviço. O trabalho aplica técnicas de mineração de dados da *web* a fim de analisar e reorganizar o conteúdo do portal e propor um engenho baseado em filtros colaborativos para recomendação de informações da *web* que não existem no portal.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

METHODOLOGY OF DATA MINING APPLIED TO MOBILE NAVIGATION

Luis Carlos Couto de Souza

September / 2009

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

Fierce competition in the Brazilian mobile market leads carriers seek differential in value-added services. Among these services highlight the WAP service which provides web access to customers. The service is based on carrier's WAP portals providing hyperlinks to websites of content providers. The goals of this work are: for the customer streamline the browsing on the portal and for carrier an option to increase service revenue. This work apply techniques of web data mining to analyze and reorganize the content of the portal and propose a recommendation engine based on collaborative filtering to recommend information in the web that does not exist in the portal.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AAA	Authentication, Authorization and Accouting.
HITS	Hypertext Induced Topic Search
HOST	Nome pelo qual um computado é conhecido na rede
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
IANA	Internet Assigned Numbers Authority
ICTAI	International Conference on Tools with Artificial Intelligence
IETF	Internet Engineering Task Force
IP	Internet Protocol
KDD	Knowledge Discovery in
MIME	Multipurpose Internet Mail Extensions
MMS	Multimedia Messaging Service
OMA	Open Mobile Alliance
PDA	Personal Digital Assistant
RFC	Request for Comments
SVA	Serviços de valor agregado
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
XML	eXtensible Markup Language
XHTML	eXtensible HyperText Markup Language
W3C	World Wide Web Consortium
WAP	Wireless Application Protocol
WML	Wireless Markup Language
WUM	Web Usage Mining
WWW	World Wide Web

1. Introdução

O acirramento da concorrência do mercado de telefonia móvel fez com que as operadoras buscassem diferencial não só na prestação de serviços de voz, mas na prestação de serviços de valor agregado, SVA.

Serviços de valor agregado são todos os serviços que não são os de voz. É possível agrupar estes serviços em duas categorias, serviços de dados e serviços de mensagens. Dentre os serviços de valor agregado de dados vale destacar: internet, *download & streaming*, MMS, *Multimedia Messaging Service*, e WAP, *Wireless Application Protocol*.

Dentre os serviços de dados citados, o serviço de WAP destaca-se por, além de prover um meio para que dispositivos móveis acessem a WWW, *World Wide Web*, ou *Web*, oferecer uma infra-estrutura para os serviços de *download & streaming* e MMS.

O WAP, embora o nome esteja associado a uma arquitetura de protocolos, é tratado como a tecnologia que permite que clientes, dispositivos móveis com certas limitações, se comuniquem com um WAP *gateway* e este com os sites na *Web*. A arquitetura utilizada para prover este serviço varia de acordo com características regionais e costumes do público alvo.

Os *web sites*, provedores de conteúdo, devem apresentar não só um conteúdo, página, baseado em HTML, *HyperText Markup Language*, mas sim em WML, *Wireless Markup Language*, e XHTML, *eXtensible HyperText Markup Language* . Estas duas formatações por limitarem a criação e as funcionalidades de um site fazem com que as páginas consistam basicamente de uma estrutura em árvore de menus, ou seja, uma árvore de *hyperlinks*. Com isto, o primeiro acesso de um cliente é sempre direcionado

ao portal da operadora que possui a árvore de *hyperlinks* para os sites provedores de conteúdo.

O modelo de tarifação deste serviço nas operadoras brasileiras, em geral, é baseado na tarifação do tráfego gerado pelo cliente, tanto de clientes pré-pagos quanto clientes pós-pagos. Portanto, o aumento da navegação dos clientes implica diretamente no incremento da receita para as operadoras.

Um meio para alavancar o aumento da utilização do serviço está em identificar o que o cliente deseja e desta maneira incentivar sua navegação. Para atingir tal objetivo, podem ser aplicadas técnicas de mineração de dados, ou seja, a extração de conhecimento que tem como o objetivo a descoberta de dependências escondidas nos dados (HAN e KAMBER, 2001).

A área de mineração de dados focada no descobrimento de conhecimento de informações na *web* é chamada mineração da *web*. A mineração da *web* é o processo de descobrimento de informações potencialmente úteis e desconhecidas dos dados da *web* (KOSALA & BLOCKEEL, 2000).

O trabalho consiste na aplicação de técnicas de mineração de dados da *web* nos *logs* de navegação do WAP *Gateway* e do portal WAP visando personalizar e melhorar a entrega do conteúdo para o cliente a fim de incentivar e aumentar sua navegação.

1.1. Objetivo

O objetivo deste trabalho é a partir de técnicas de mineração de dados da *web*, mais especificamente mineração de uso da *web*, WUM, *Web Usage Mining*, extrair conhecimento dos *logs* de navegação de clientes do WAP *Gateway* e *logs* de navegação do portal WAP para propor uma alteração no portal WAP da operadora VIVO.

Como nem todo o tráfego WAP é direcionado ao portal da operadora, a proposta consistirá em adequar a árvore de menus considerando toda a navegação existente. As ações da proposta podem basicamente ser agrupadas em: identificar o comportamento de acessos entre os itens do portal a fim de reestruturá-lo e prover um engenho para recomendação de conteúdo a partir de padrões de navegação do portal e do WAP *Gateway*.

A identificação dos itens do portal que são acessados simultaneamente em uma mesma sessão e a segmentação de clientes propiciará a chance de organizar estes itens de forma a agrupá-los em um mesmo lugar facilitando e incentivando a navegação do cliente.

A identificação de regras de associação que possuem antecedentes como os itens do portal e consequentes como *sites* na internet permitirá recomendar itens durante a navegação do cliente no portal.

1.2. Relevância

O crescimento da *web* nas últimas décadas resultou em uma variedade de informação, produtos, serviços que fazem com que os usuários sintam-se muitas vezes desorientados. No WAP isto se agrava em virtude das limitações do manuseio dos telefones móveis e limitação dos conteúdos exibidos para os clientes.

Neste cenário, a criação e a manutenção de um *site* não se tratam de tarefas simples. Conquistar e cativar a atenção do usuário torna-se um fator primordial para o sucesso. A análise de cliques do usuário, ou seja, a análise da navegação a partir dos *logs* dos servidores surge como um meio indispensável para avaliar a eficácia de estratégias e aproveitar uma série de oportunidades na *web*.

O processo de análise dos *logs* dos servidores por intermédio de ferramenta de mineração de dados pode derivar conhecimento relevante para as organizações. Este processo, denominado de WUM, visa descobrir padrões interessantes que possam auxiliar as organizações a entender e melhor servir seus visitantes.

A importância deste trabalho está em aplicar as técnicas de WUM não apenas em um *site*, mas em ambiente WAP como um todo. Além disto, não será analisado o *log* de navegação em um site, mas sim a navegação do usuário na *web* como um todo, a partir dos *logs* de um WAP *Gateway*.

1.3. Estado da arte

A mineração de dados na *web* tem sido definida como a aplicação de técnicas de *Data Mining* em dados da *web*. Ela refere-se ao processo de descoberta de conhecimento interessante com a utilização de métodos e técnicas que permitem analisar conteúdo, estrutura e uso da *web* para a extração de informação previamente desconhecida, válida e que gera ações úteis e de grande ajuda para a tomada de decisões estratégicas.

Suas implicações estão na recuperação de informações de produtos, recuperação de opiniões de usuários, identificações de grupos de comunidades, busca de conteúdo, personalização da informação, dentre outras.

A mineração do uso da *web* é a área de mineração de dados da *web* responsável pela extração de conhecimento da utilização. Seu campo foca na recuperação de conhecimento a partir, principalmente, da utilização da *web* associada às informações dos usuários e da arquitetura do *site*. Sua aplicação está diretamente ligada ao comércio eletrônico, através da entrega customizada a partir do comportamento do usuário.

1.3.1 Descoberta de conhecimento em Bases de Dados

O descobrimento de conhecimento de base dados, em inglês *Knowledge Discovery in Databases* (KDD), é um processo interativo que possui uma série de atividades que se iniciam na coleta de informações, que em seguida são pré-processadas, mineradas para identificação de padrões que por final são avaliados e apresentados como informações úteis.

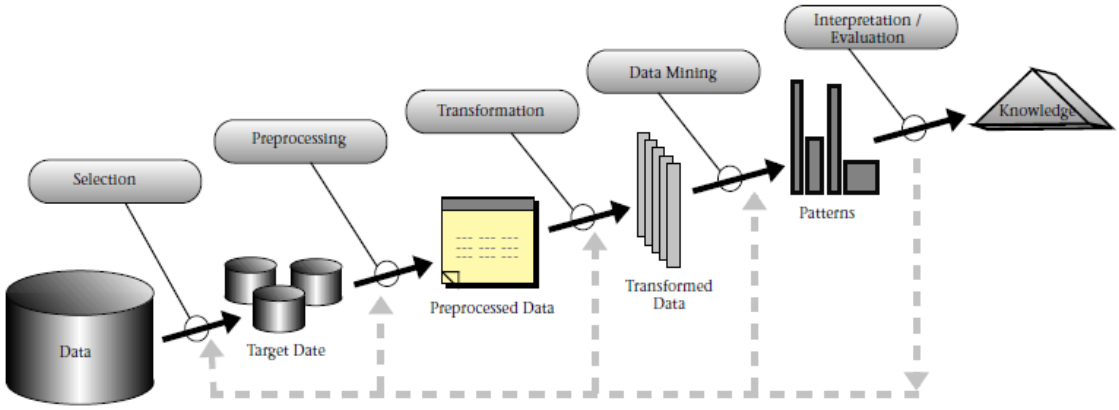

Figura 1-1 - Etapas do processo de KDD (FAYYAD *et al.*, 1996)

A mineração de dados é parte do processo de descoberta de conhecimento que visa a aplicação de métodos inteligentes para a extração de padrões. Portanto, a mineração de dados é a análise de um conjunto de dados observacionais, muitas vezes grandes, a fim de identificar relações imprevistas e resumir os dados para uma nova forma simultaneamente compreensível e útil para o proprietário. (HAND *et al.*, 2001).

Muitas pessoas tratam o termo mineração de dados como sinônimo para o processo de descobrimento de dados (HAND *et al.,* 2001). Embora este trabalho reconheça a mineração de dados como uma atividade do processo de descobrimento de conhecimento, a mineração de dados será citada como o processo de descoberta de conhecimento como um todo.

A mineração de dados pode ser aplicável para qualquer tipo de base de dados ou repositório de informação estruturada, semi-estruturada e não estruturada. Isto inclui bancos de dados relacionais, espaciais, orientados a objeto, documentos texto e até a grande rede de computadores.

1.3.2 Mineração de dados da Web

A *web* foi inventada em 1989 a partir de uma proposta feita por Tim Bernes Lee, mas que somente começou a ser trabalhada, criada de fato, a partir 1990. O incrível crescimento da *web* nas últimas décadas fez com que ela seja a maior fonte de dados acessível no mundo. Ela possui características únicas que fazem da mineração de conhecimento e informação um desafio e uma tarefa fascinante (LIU, 2007).

Em 1997 foi organizado um painel na conferência internacional de ferramenta de inteligência artificial, ICTAI, *International Conference on Tools with Artificial Intelligence*, que questionou se havia algo na mineração de dados da *web* distinto da mineração de dados.

Duas abordagens distintas foram consideradas inicialmente. A primeira, focada em processo, considerando a mineração da *web* como um sequência de tarefas. A segunda e mais aceita, focando em dados, e definindo a mineração da *web* em termo dos tipos de dados utilizados no processo de mineração (SRIVASTAVA *et al.*, 2005). A partir de então, a tarefa de mineração de dados da *web* foi denomina *Web Mining*.

Esta área de mineração de informação não estruturada foca na extração de conhecimento a partir da estrutura de *hyperlinks*, do conteúdo das páginas, e da navegação.

Embora a mineração da *web* utilize uma série de técnicas da mineração de dados, ela apresenta particularidades devido à heterogeneidade e as informações semi ou não estruturadas da *web*. A mineração da *web* pode ser classificada em três grupos de acordo com o tipo de informação tratada: mineração da estrutura, mineração do conteúdo e mineração do uso.

1.3.2.1 *Web Content Mining*

Web Content Mining, mineração de conteúdo da *web* é a área de conhecimento responsável por extrair ou minerar conhecimento a partir do conteúdo das páginas *web*. Seus métodos consistem em trabalhar diversos tipos de dados como texto, imagem, áudio, vídeo e meta dados.

Nesta área de conhecimento também está incluso o campo de conteúdo de multimídia que tem por objetivo a mineração de imagens, sons e vídeos. Sua utilização está principalmente na mineração de imagens de satélites e na mineração de identificações humanas.

Através das técnicas de mineração de conteúdo é possível extrair informações de descrições de produtos, fóruns, avaliações e opiniões de usuários para uma série de finalidades.

1.3.2.2 *Web Structure Mining*

Web Structure Mining, mineração da estrutura da *web,* é a área de conhecimento responsável por extrair ou minerar conhecimento a partir da estrutura de *hyperlinks* das páginas *web*. Esta área foi inspirada pelo estudo de redes sociais.

Seus métodos tratam a ligação dos *hyperlinks* como um grafo e aplicam determinadas abordagens para relacionar os objetos detentores de tais *hyperlinks*. Duas abordagens são utilizadas, uma focando nos *hyperlinks* de entrada, prestígio, e outra nos *hyperlinks* de saída, centralidade.

As técnicas baseada na abordagem de prestígio são as mais utilizadas pelos algoritmos. Dentre estas técnicas pode-se citar Page Rank e HITS.

Através das ferramentas de mineração da estrutura da *web* é possível extrair informações para mecanismos de busca, identificar comunidades de usuários com interesses comuns, rastreamento da *web* ou *web crawlers e* ordenar prestígio de autores e obras.

1.3.2.3 *Web Usage Mining*

Web Usage Mining, mineração de dados de uso da *web*, é o campo de pesquisa focado no desenvolvimento de técnicas e ferramentas para estudo do comportamento dos usuários durante suas navegações na *web* (Borges, 2000).

A compreensão das preferências de navegação de um visitante é uma importante etapa para o estudo de qualidade, principalmente em sites de comércio eletrônico. Através dos padrões de acessos dos usuários é possível reestruturar, personalizar e adaptar a interface do site para um usuário específico ou fazer melhorias no sistema.

O processo de mineração visa analisar o comportamento de humanos e como eles interagem na internet. A partir da análise deste comportamento podem ser identificadas oportunidades de customização e personalização para melhorar o contato humano com a aplicação. Por esta razão, esta área apresenta grande interesse para o mercado de comércio eletrônico em geral.

1.3.3 Taxonomia

A seguir é apresentada a taxonomia da mineração da *web* de acordo com o tipo de dados que podem ser minerados (SRIVASTAVA et al., 2005).

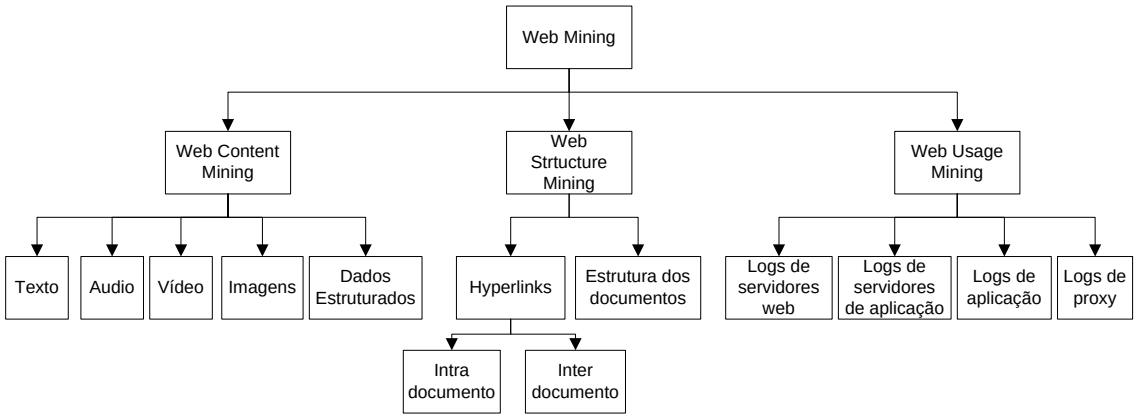

Figura 1-2 – Taxonomia de mineração de dados da *web* (SRIVASTVA *et al.*, 2005)

SRIVASTAVA (2000) classificou os dados apresentados acima em três grupos: dados de conteúdo; dados de estrutura e dados de uso. Além deste grupo, foi apresentado um quarto grupo, dados de perfil do usuário, que embora não esteja diretamente relacionado ao universo da *web* é essencial para o campo da mineração do uso da *web.*

A seguir estão descritas as classificações dos dados:

- Dados de conteúdo

Representa o dado real da página que os usuários vêem. Geralmente consiste, mas não necessariamente, de texto, imagens e gráficos.

- Dados de estrutura

Este dado descreve a organização do conteúdo. O principal tipo de estrutura são os *hyperlinks* conectando uma página a outra.

- Dados de uso

Este dado descreve os padrões de utilização da página. Ele descreve quem acessou quando e de onde veio.

- Dados do perfil do usuário

Este dado descreve as informações dos usuários que acessam o site.

1.4. Organização da dissertação

Neste capítulo foi apresentado o contexto geral do trabalho, a relevância da aplicação de uma metodologia de descoberta de conhecimento da *web* em navegações WAP, o estado da arte do campo de mineração de dados e os objetivos pretendidos.

O capítulo 2 trata da metodologia proposta e define as etapas necessárias para desenvolver o processo de descoberta de conhecimento na *web*.

No capítulo 3, é apresentado o estudo de caso, com a descrição do problema, as etapas do processo de descoberta de conhecimento na *web* utilizada para o caso real e os resultados obtidos

O capítulo 4 faz considerações quanto ao estudo de caso e sugestões de desenvolvimentos futuros que podem ser efetuados.

O capítulo 5 apresenta as considerações finais e descreve as conclusões do trabalho.

2. Metodologia

Existem diversos modelos de processo para a mineração de dados, mas todos consistem de uma fase de pré-processamento, que possuem atividades de entendimento do escopo, entendimento dos dados e pré-processamento, uma fase de mineração para o descobrimento de padrões e uma fase de análise.

Esta dissertação seguirá um modelo de fases sugerido por COOLEY (1999) que basicamente consiste em três tarefas fases: pré-processamento, descobrimento dos padrões e análise dos padrões.

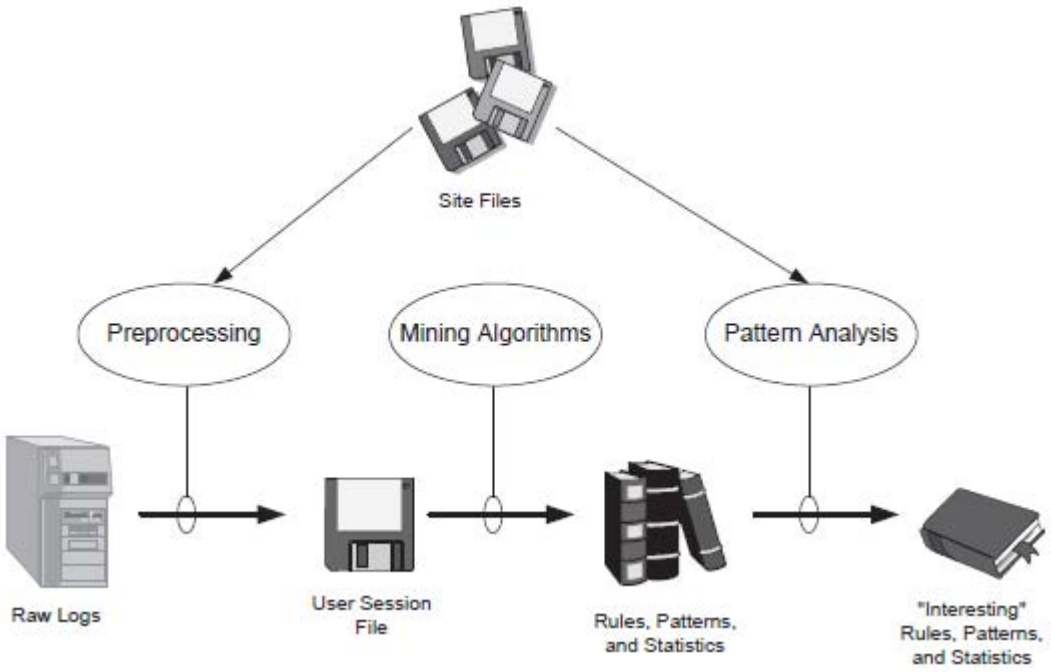

Figura 2-1 - Etapas de WUM (COOLEY, 1999)

As tarefas de pré-processamento consistem em converter o uso, o conteúdo e estrutura da informação contida em várias fontes de dados em uma abstração de dado onde seja possível a descoberta de padrões.

Para o descobrimento de padrões nos dados preparados anteriormente são utilizados métodos e algoritmos de diversos campos, do quais se pode citar: estatística, *data mining*, redes neurais, algoritmos genéticos e outros de aprendizado de máquina e reconhecimento de padrões.

Na fase de análise, os dados são analisados pelo minerador, o conhecimento descoberto passa por avaliações subjetivas de interesse e acionabilidade, e avaliados por especialistas que tentam identificar a qualidade dos padrões descobertos. Estas análises podem resultar em uma nova aplicação dos algoritmos para descoberta de conhecimento.

2.1. Pré-processamento

A fase de pré-processamento é a fase que mais exige esforço do minerador de dados. Esta fase contempla todos os aspectos para a preparação do dado. Consiste em selecionar os casos e variáveis que se deseja analisar, executar a transformação de certas variáveis e quando necessário limpar a navegação removendo transações e dados impertinentes.

2.1.1 Unificação de dados

Muitos sites por questões de disponibilidade e desempenho têm uma arquitetura balanceada com vários servidores e, ou, distribuidores de carga. Com isto a informação de navegação esta distribuída em cada um dos nós. Esta fase tem por objetivo coletar e unificar estes dados.

2.1.2 Limpeza e filtragem dos dados

Esta fase tem por objetivo remover as transações, requisições, automáticas geradas pelo próprio texto da página e que não foram explicitamente solicitados pelo usuário. Dentre estes elementos pode-se citar: imagens, sons e estilos.

Aqui também são filtradas as informações que serão tratadas nos processo seguintes. Dentre as atividades pode-se citar:

- Remover arquivos gráficos e de estilo a menos que o interesse do problema consista em analisar tal tráfego, por exemplo, para avaliação de consumo de banda;
- Extração das informações que deverão ser tratadas;
- Normalizar valores compostos quando necessário;
- Derivação do tempo para identificar o tempo de permanência em uma URL, *Uniform Resource Locator*;
- Filtragem para remoção dos acessos desnecessários realizados automaticamente.

2.1.3 Remoção de acessos de robôs

Constantemente processos automáticos estão rastreando a *web* para prover serviços de busca, comparação de preços e outras atividades de buscas exaustivas. O comportamento destes robôs difere do comportamento humano e por isso não é considerado útil para o processo de mineração. Portanto, as requisições destes processos devem ser removidas dos *logs* de navegação.

Para identificar o acesso destes robôs é possível aplicar as seguintes técnicas:

- Identificação do nome de robôs pelo parâmetro do cabeçalho HTTP, *Hypertext Transfer Protocol, user-agent*;
- Análise da quantidade de acesso por *user-agent*, por endereço IP, *Internet Protocol*, por sessão, e verificar alguma discrepância a fim de identificar robôs.

Exemplos de robôs são: o Google Bot e MSN Bot, cujas referências podem ser encontradas respectivamente em http://www.google.com/bot.html e search.msn.com/msnbot.htm

2.1.4 Identificação do usuário

A identificação do usuário é importante para identificar sua sessão no site.

Em virtude da estrutura da *web* muitas navegações ocorrem de forma anônima. Portanto, o minerador deve utilizar uma série de informações para identificação dos usuários.

Para a identificação dos usuários pode-se aplicar:

- Utilização de *cookies*, mas que está sujeito a configuração do browser do cliente;
- Utilização de *IP*, mas que está sujeito a forma de acesso com ou sem *proxy*;
- Combinação do endereço *IP*, a data de navegação, o parâmetro *user-agent* e o parâmetro *refer*;

- Para aplicações customizadas é possível a utilização de *logs* proprietários onde usuários podem ser identificador por identificadores internos da aplicação.

2.1.5 Identificação da sessão

Denomina-se sessão o período de uma navegação no site, ou seja, da entrada até a saída do usuário. Nesta atividade são determinadas as páginas que foram requisitas a ordem e a duração de visualização de cada página. Também é tentado identificar quando e em qual página o usuário deixou o site.

Como não é possível identificar a saída de um site do usuário, é necessário assumir um limite de tempo de inatividade, ou seja, sem uma nova requisição. Após este período considera-se que o usuário deixou o site.

Nesta fase também são computadas, para cada página, uma aproximação do tempo de permanência. Uma abordagem de BAGLIONI (2003) define que o tempo gasto pelo usuário é proporcional ao seu interesse pelo conteúdo da página. Portanto, a informação do tempo de permanência na página representa uma informação valiosa para o processo de mineração.

2.1.6 Complementação da navegação

Muitos usuários utilizam o voltar do navegador, que na maior parte das vezes, recupera a página armazenada localmente. Isto implica em buracos nos *logs* de navegação do servidor. As atividades desta fase consistem em, a partir do conhecimento da topologia do site, completar as transações inexistentes.

2.1.7 Integração dos dados

Esta fase tem por objetivo integrar os dados de navegação com demais dados do site e de seus usuários. Pode-se citar a associação destes dados com avaliações de realizadas pelos usuários, taxas de conversão de moedas em um determinado período, faixas etárias, faixas salariais, etc.

2.1.8 Transformação da navegação

Após as fases anteriores, os registros ainda não estão em uma forma apropriada para os algoritmos de mineração de dados. Portanto, é preferível transformar os valores das transações em uma coleção de variáveis.

Aqui são adicionadas, a partir dos dados da requisição, informações quanto a determinada característica. Geralmente estas transformações são válidas apenas para transações que possuem uma determinada representatividade no conjunto.

2.2. Mineração

Após a fase de pré-processamento os dados estão prontos para serem aplicadas as técnicas de mineração de dados.

Técnicas de clusterização podem ser utilizadas para determinar grupos de usuários com padrões semelhantes. Técnicas de associação e de análise de correlação podem identificar grupos de itens ou páginas que ocorrem juntas. Técnicas de classificação podem ser utilizadas para identificar se determinados usuários pertencem a determinada classe ou categoria pré-definida.

As técnicas a serem aplicadas dependerão do tipo do problema a ser resolvido. Problemas comuns em WUM são: a segmentação em comércio eletrônico, a criação de comunidades e a personalização o conteúdo *web* de acordo com os grupos de usuário.

2.1.9 Regras de associação

O uso de regras de associação foi primeiro proposto por AGRAWAL (1993), que visa encontrar relacionamentos dentre um conjunto de dados. Ele definiu a regra de associação da seguinte forma:

Sejam I= {i1, i2, ..., in}um conjunto de n-itens distintos e de uma base de dados formada por um conjunto de transações, onde cada transação T é composta por um conjunto de itens, tal que $T \subseteq I$. Uma regra de associação é uma expressão na forma A => B, onde $A \subset I$, $B \subset I$, $A \neq 0$, $B \neq 0$ e $A \cap B = 0$.

O grau de certeza de um relacionamento é medido pelo fator de suporte e de confiança. Considerando dois subconjuntos A e B, o suporte de uma relação A => B é representada pela fração de itens que possuem A e B pelo número de itens total. A confiança da relação é dada pela fração de itens que possuem A e B pelo número total de itens que possuem A. Estes fatores limitam a quantidade de regras que serão extraídas e descrevem a qualidade delas.

Existem vários tipos de regras de associação. Tais regras podem ser classificadas segundo os critérios de: tipo de valores, dimensões do dado envolvido, níveis de abstração (HAN e KAMBER, 2001).

O critério de tipo de valores refere-se a regras do tipo binário, falso e verdadeiro, e quantitativo, quando descreve relações de quantidade entre o conjunto de itens.

O critério de dimensionalidade refere-se à quantidade de dimensões do conjunto de itens, ou seja, diz respeito à quantidade de atributos dos dados estudados, que podem possuir vários ou apenas um atributo.

O critério de nível de abstração se refere à hierarquia do tipo do dado, onde dados podem possuir apenas um nível ou vários níveis de abstração.

Problemas com domínio que possuem múltiplos níveis de hierarquia, ou abstração, apresentam particularidades quanto à utilização de suporte, pois os níveis mais altos da abstração tendem a apresentar suportes altos e níveis mais específicos suporte baixos. Com isto a utilização de apenas um suporte geral provavelmente implicará ou na geração de muitas regras de associação ou no descarte de regras importantes para os itens mais específicos.

Para solucionar o problema, podem ser utilizados múltiplos suportes por níveis. Para tal, podem ser utilizadas abordagens de investigar níveis superiores apenas se os itens de níveis inferiores são frequentes ou investigar apenas níveis inferiores se níveis superiores são frequentes.

A aplicação de regras de associação em mineração de uso da *web* iniciou pelo trabalho de CHEN (1996) que definiu o conceito de sessão *web* como a quantidade máxima de cliques sucessivos. Em seguida CHEN (1998) usou algoritmos de regras de associação para encontrar relações interessantes entre sessões de usuários.

2.3. Análise

A etapa de Análise da Informação também pode ser chamada de Pós-processamento de dados e diz respeito à verificação da eficiência da aplicação dos algoritmos da etapa anterior. Em outras palavras, é o momento de avaliar se o objetivo foi cumprido da melhor forma possível, que é descobrir conhecimento novo e inovador.

Existem diversas maneiras de se avaliar a mineração como um todo, seja de forma qualitativa ou quantitativa. A utilização de métricas é considerada uma forma quantitativa, ao passo que a utilização do conhecimento de especialistas no domínio é considerada uma forma qualitativa. Os especialistas devem sempre ser consultados, em todas as etapas da Mineração, balizando a análise, ajudando a resolver situações de conflito, indicando caminhos e complementando informações.

As principais medidas quantitativas para regras de associação são suporte, confiança e tamanho das regras. O número de regras está relacionado ao critério de simplicidade da compreensão pelo especialista do negócio. A confiança está relacionada ao critério de certeza da regra. O suporte está relacionado à usabilidade da regra.

As principais medidas quantitativas para clusterização são número de agrupamentos, critérios de corte. Tais medidas implicam em identificar agrupamentos que possuam itens mais homogêneos possíveis dentro de um mesmo grupo e mais heterogêneos possíveis de elementos de outros grupos.

Através destas medidas o minerador pode direcionar o conhecimento a ser extraído antes de apresentar para o especialista de negócio. Entretanto, mesmo com estas medidas a quantidade de regras pode atingir dezenas de milhares e ainda implicar em conhecimentos óbvios. Portanto, é necessária a aplicação de medidas qualitativas.

(HAN e KAMBER, 2001) Propõe que para considerar um conhecimento interessante para os especialistas podem ser utilizados os seguintes critérios subjetivos:

- Conhecimentos inesperados se os especialistas se surpreendem os dados minerados;
- Conhecimentos esperados se os especialistas desejam confirmar uma hipótese;
- Conhecimento acionável se os especialistas podem, a partir do dado minerado, adotar medidas para o negócio;

3. Estudo de Casos

Esta seção apresenta o estudo de caso que consiste em aplicar técnicas de mineração de dados a navegação de telefones móveis da VIVO, serviço WAP, a fim de propor mecanismos e ações para facilitar a utilização do serviço para o cliente e incrementar o tráfego para operadora.

Para atingir tal objetivo o estudo propõe duas ações. A primeira consiste em analisar o tráfego do portal WAP da operadora e, se necessário, propor reorganizações nos itens do portal. A segunda ação propõe um engenho de recomendação capaz de, a partir de uma janela de navegação de um usuário no Portal, recomendar *sites* na *web* que não estejam publicados no portal.

\O estudo termina descrevendo os resultados obtidos com a aplicação de um protótipo do engenho de recomendação em uma amostra de dados real.

3.1. WAP

O WAP é uma arquitetura de protocolos para permitir que telefones móveis possam acessar a internet como se fossem navegadores em computadores pessoais. Entretanto, este termo, passou-se a identificar o serviço de dados pelo quais telefones móveis acessam a internet.

Nesta arquitetura os telefones móveis se comunicam com um componente intermediário, chamado WAP *Gateway*, cujo é responsável por comunicar-se com os telefones móveis e com os servidores da *web*.

.A evolução tecnológica do modelo da arquitetura WAP originou o WAP 2, que embora também represente uma arquitetura de protocolo, é utilizado para referir-se a terminais que suportam a tecnologia WAP 2.

Nesta evolução, embora os telefones móveis sejam capazes de se comunicarem diretamente com os servidores de conteúdo, por desenho de arquitetura as operadoras de telefonia móveis mantêm o intermédio do acesso através do WAP *Gateway*. Neste caso, o WAP *Gateway* assume basicamente o papel de um Proxy.

Para a tecnologia WAP 1 a formatação das páginas é baseada em WML. Esta formatação consiste de um arquivo bem formado onde, basicamente, o arquivo representa um "baralho" e o conteúdo exibido são as cartas. Portanto a navegação do telefone móvel através do conteúdo é realizada através da solicitação das "cartas" e quando necessário é feita a solicitação de um novo "baralho" ao servidor.

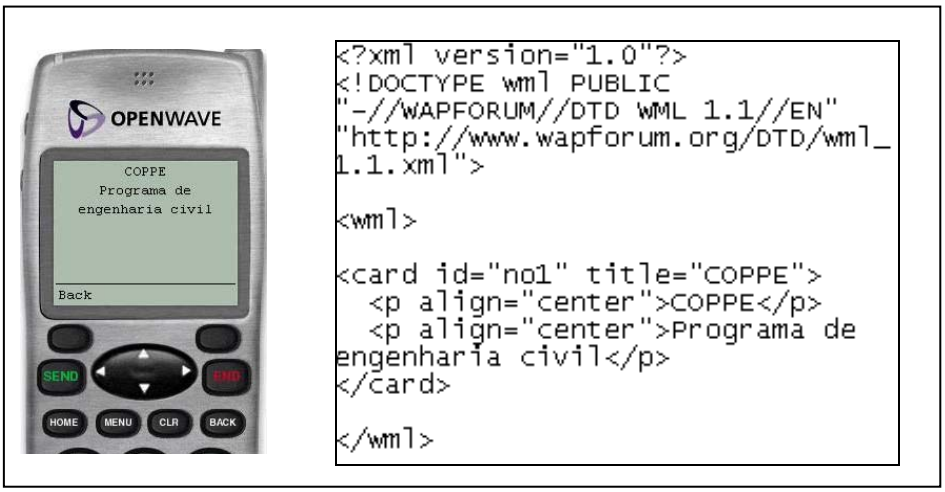

Figura 3-1 – Página com formatação WML

Em virtude da limitação da tecnologia, além das restrições do WML, existe a limitação do tamanho de uma página, que na média é de 1,5 kbytes. Isto restringe a navegabilidade dos usuários a apenas cliques.

A figura 3-1 apresenta uma página utilizando a formatação WML juntamente com a apresentação em um simulador.

Para a tecnologia WAP 2 a formatação das páginas é baseada em XHTML. Esta formatação é mais similar ao HTML e apresenta uma infinidade de recursos

gráficos se comparada ao WML. Entretanto, mesmo sem a limitação de tamanho e com os novos recursos gráficos, a navegabilidade dos usuários continua baseada em cliques.


```
<!DOCTYPE html PUBLIC "-//WAPFORUM//DTD
XHTML Mobile 1.2//EN"
"http://www.openmobilealliance.org/tech/D
TD/xhtml-mobile12.dtd">
<html
xmlns="http://www.w3.org/1999/xhtml">
<head>
        <meta http-equiv="content-type"
content="text/html; charset=utf-8" />
        <meta http-equiv="Cache-Control"
content="no-cache" />
        <meta http-equiv="Cache-Control"
content="max-age=1" />
        <title>COPPE</title>
</head>

<body>
        <p align="center">
                <h1>COPPE</h1>
                Programa de engenharia
civil
        </p>
</body>
</html>
```

Figura 3-2 – Página com formatação XHTML

Os telefones móveis mais recentes com tecnologia WAP suportam além do XHTML a interpretação do HTML. No entanto, na grande maioria, o suporte limita-se ao HTML não existindo suporte, por exemplo: *javascript*, *flash*, etc. Para estes telefones embora tenha existido um grande avanço quanto à apresentação de conteúdo, a navegabilidade para aquisição de informação é restrita devido ao tamanho da tela e as restrições do teclado.

Neste cenário, os acessos dos usuários baseiam-se principalmente em *hyperlinks* que são disponibilizados para o mesmo. Isto faz com que a grande maioria dos *sites* se baseie em uma estrutura de árvore de menus. Portanto, o refinamento de *sites* para prover o que o usuário deseja é fundamental.

3.2. Serviço WAP na operadora VIVO

A arquitetura do serviço WAP na VIVO consiste em um WAP *Gateway* realizando o intermédio entre os telefones móveis e os servidores na *internet* tanto para telefones WAP 1 quanto para telefones WAP 2.

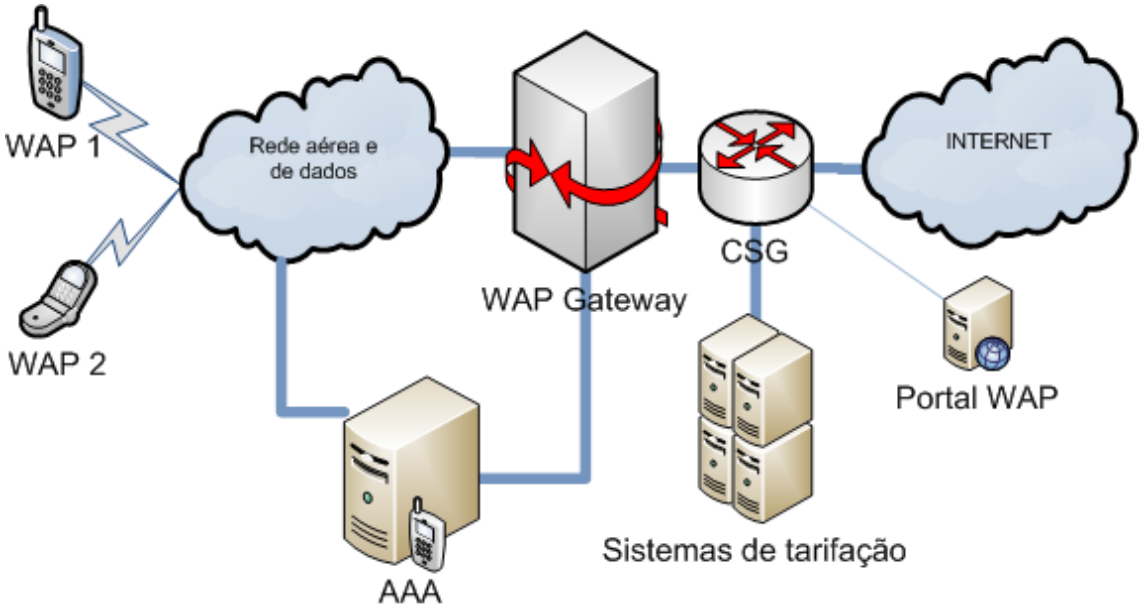

Figura 3-3 – Arquitetura do serviço WAP na VIVO

A cobrança deste serviço por parte da operadora é realizada a partir do tráfego gerado pelos clientes. Para calcular tal tráfego existe um elemento entre o WAP *Gateway* e a internet, denominado CSG, similar a um *switch*, que é responsável por aferir o que é trafegado pelo cliente e comunicar o tráfego aos sistemas de tarifação da operadora.

Para que os telefoneis possam requisitar páginas na *web*, o WAP *Gateway* autoriza a navegação a partir de mensagens recebidas por um elemento de rede responsável pela autenticação, autorização e bilhetagem dos usuários. Este elemento conhecido como AAA, em inglês, *Authentication, Authorization and Accounting*, autentica os usuários para utilizarem o serviço de dados e reporta ao WAP *Gateway* que tal usuário está autorizado a utilizar o serviço.

Para os telefones móveis WAP 1, e por padrão, para os telefones móveis WAP 2, o primeiro acesso do cliente é sempre direcionado para o portal WAP da operadora. Neste portal existem *hyperlinks*, de parceiros da operadora, que propiciam a navegação do cliente. Embora o primeiro acesso dos clientes seja direcionado para o portal da operadora, os clientes podem acessar, apesar das limitações de navegabilidade, as informações na *internet* como um navegador de um computador pessoal.

Considerando as restrições existentes nos telefones móveis e o direcionamento dos acessos iniciais para o portal da operadora, a disponibilização das informações neste portal se torna primordial para o cliente e para a operadora.

A organização da árvore de menus no portal da operadora visando entregar o conteúdo que o cliente deseja facilita a utilização da internet pelo cliente, e para operadora, implica em um incremento de receita do serviço. No entanto, embora a otimização das informações no portal da operada represente uma vantagem, o portal não pode prever quais outros sites na internet são utilizados. Sendo assim, existem dois pontos que precisam ser tratados para aperfeiçoar a entrega de informação para o cliente, e consequentemente em um incremento de receita para a operadora.

O primeiro ponto é analisar o portal, verificando se a estrutura da árvore de menus é ótima, e se necessário, reorganizar sua estrutura de menus. O segundo ponto é, a partir do tráfego dos clientes, considerando toda a internet, mapear recomendações para um determinado cliente a fim de potencializar sua navegação.

3.3. Reorganização do portal através de técnicas de mineração de dados

O primeiro ponto a ser tratado consiste em efetuar uma análise na ferramenta de portal para identificar melhorias na apresentação de menus. Para tal, serão recuperadas as transações dos clientes, estas transações serão pré-processadas, serão aplicadas técnicas de mineração nas sessões do cliente e, a partir do conhecimento obtido, recomendar alterações para reorganizar o portal.

3.3.1. Portal WAP

O portal WAP da VIVO consiste de uma aplicação, que a partir de informações de configuração armazenadas no banco de dados, provê dinamicamente à interface para os clientes WAP. A figura a seguir apresenta o portal exibido em um simulador WAP 2.

Figura 3-4 – Portal WAP exibido em simulador WAP 2

Este portal é responsável também por atender a terminais PDA, em inglês, *Personal Digital Assintants*, ou Assistente Pessoal Digital e terminais *SmartPhones* que representam telefones móveis com capacidade direta de acesso a internet.

A interface apresentada para o cliente, basicamente, consiste de uma árvore de menus, onde a folha representa o direcionamento para um serviço de conteúdo. A figura 3-5 exemplifica esta árvore de menus.

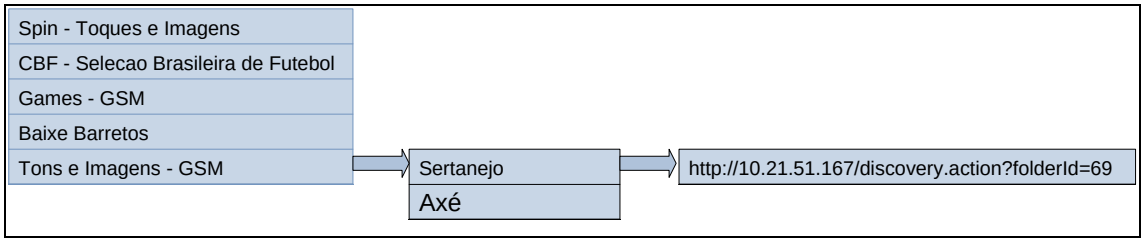

Figura 3-5 – Árvore de menus

Este portal possui uma ferramenta de administração que permite administradores, recursos especializados em marketing, gerenciarem a árvores de menus e personalizarem a visualização de tais menus de acordo com as capacidades do terminal e regional do cliente.

Para a análise efetuada a personalização de menus poderá ser desconsiderada visto que, embora os menus na árvore possam ser reorganizados, a visualização dos mesmos continuará submetida às características do cliente. Ou seja, se menus em posições diferenciadas sejam vistos em uma mesma sessão apenas para um grupo de clientes, este menus após exibidos agrupados em uma mesma pasta somente serão exibidos para o dado grupo de cliente.

No portal, além de pastas, nós de uma árvore, e *hyperlinks*, folhas de uma árvore, existem outros tipos de menus. São menus responsáveis por apresentação de texto, apresentação de caixa de busca, etc. Entretanto, como para a análise será necessário considerar a estrutura do portal, os dados coletados deverão se restringir aos tipos de menu *hyperlinks* e pastas.

A arquitetura do portal consiste em uma série de servidores que recebem as requisições através de um distribuidor de tráfego e a partir das configurações armazenada no banco de dados gera o conteúdo para o cliente.

A ferramenta possui um mecanismo de armazenamento de acessos que consiste dos servidores responsáveis por atenderem as requisições do cliente armazenar assincronamente os acessos em um banco de dados relacional central. A partir deste banco de dados, é possível a recuperação das transações.

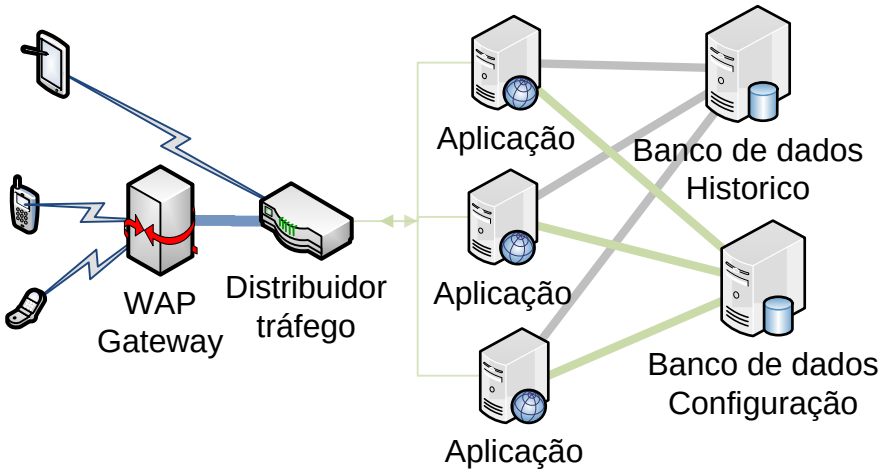

Figura 3-6 – Arquitetura do Portal WAP

Na arquitetura da figura 3-6, o atendimento a terminais PDA e *Smartphone* é diferenciado pelo distribuidor de tráfego. Esta segregação propicia que o tráfego oriundo WAP *Gateway* seja confiável se comparado ao tráfego oriundo dos terminais PDA e *Smartphone* devido à inexistência de robôs e identificação do cliente.

O armazenamento das transações destes terminais também é realizado no banco de dados. Entretanto, tais transações são armazenadas com identificador a parte dos demais clientes WAP.

3.3.2. Extração e pré-processamento das transações

Através do repositório central de dados a tarefa de pré-processamento para unificação dos dados restringiu-se a coletar os dados do servidor de banco de dados.

No banco de dados de histórico existem as informações de: código de área e identificador do cliente; data composta por hora minuto e segundo do acesso; o item e o tipo de menu acessado. Durante a extração das informações, o processo de extração limitou-se aos tipos de menus *hyperlinks* e pastas.

Foram extraídas 1.371.470 transações dos tipos pasta e *hyperlinks*. As informações foram extraídas para um arquivo texto com os seguintes parâmetros separados por ponto-e-vírgula: código de área; identificador do cliente; data com hora; minuto e segundo; tipo do menu; descrição e identificador do menu; descrição e identificador da pasta que contem o menu. Exemplo: "*21;9888870634906;20090629230000;3;CBF - Selecao Brasileira de Futebol;100354;Vivo WAP Arvore 2X;100030*"

Em virtude do WAP *Gateway* intermediar as transações garantindo a autenticidade de terminais não será necessário a remoção de acessos oriundos de robôs. Entretanto foi necessária a remoção dos acessos de PDA e *Smartphone* em virtude dos mesmos não serem identificados.

Embora tenham sido filtrados os acessos aos menus dos tipos pasta e *hyperlinks*, não houve mais atividades referentes à filtragem de acessos automáticos.

A identificação de usuários será realizada a partir de seu identificador. Sendo assim, não existirão as atividades de relacionar IP, *cookies* ou *user-agents*.

Na preparação dos dados foi identificada uma sequência de acessos replicados. Em uma análise com os recursos especialistas no WAP foi identificado que para

terminais móveis o usuário apresenta um comportamento de executar várias requisições enquanto a resposta não é apresentada. Isto se agrava com o fato da formatação WML e XHTML não ter meios para bloquear tais requisições. Sendo assim, as transações referentes a estes acessos foram removidas para não influenciarem a avaliação dos dados.

Para a obtenção da sessão, os acessos por usuários são ordenados e em seguida é calculado se a diferença de tempo entre dois acessos representa ou não uma sessão nova. Para o cálculo da diferença entre dois acessos do usuário, a data foi transformada para um número de milissegundos referente a 1 de Janeiro de 1970 0 hora, 0 minuto e 0 segundo. Com este número é calculada a diferença entre dois acessos do cliente e avaliado se esta diferença é menor que a quantidade de tempo da sessão do cliente.

Para a criação da sessão também é avaliado se a quantidade de transações existentes na sessão atende a um mínimo necessário.

Para as tarefas de pré-processamento citadas foi criado um aplicativo Java, processadorPortal.jar., responsável pelas atividades de pré-processamento descritas e geração de dois arquivos para extração de regras de associação.

Este aplicativo foi criado utilizando Java 1.6. Para o desenvolvimento foi utilizada a IDE eclipse 3.4.2.

Para a extração das regras o aplicativo gerou dois conjuntos de dados: um contendo o tráfego de pastas e *hyperlinks* e outro contendo apenas o tráfego de *hyperlinks*.

Em virtude da hierarquia do Portal, para o estudo do problema implicaria em termos que trabalhar com níveis de suporte baixos para identificar os níveis mais específicos do Portal, o que resultaria em muitas regras, ou trabalhar com nível de

suporte médio que implicaria, talvez, em perda de relacionamentos importantes. Portanto a estratégia adota foi analisar uma amostra contendo apenas os *hyperlinks* para internet e outro contendo todos os acessos a pasta e *hyperlinks*.

Foi utilizado o tempo de quinze minutos para se considerar uma nova sessão em virtude do tempo de expiração de uma sessão do portal ser de 15 minutos.

Foi utilizada uma janela mínima para sessão de quatro acessos. Este valor é oriundo de uma análise da hierarquia do Portal onde foi constatado que a maior profundidade é cinco. Devido a isto se optou por utilizar quatro acessos para se considerar uma sessão válida.

Os arquivos gerados pela aplicação são arquivos texto contendo um identificador para transação, o identificador do cliente o identificador da sessão e a descrição do menu. Exemplo: "1,2198888706,1,Promo Jogos".

3.3.3. Mineração e análise dos dados

Para a mineração dos dados foi utilizado o ambiente SAS. A ferramenta responsável pela mineração de dados denomina-se Miner.

O processo de mineração se inicia pela carga dos arquivos em uma estrutura interna localmente. Esta estrutura interna é publicada para uma biblioteca existente em um servidor. Após este passo é possível iniciar a utilização do Miner.

Durante a criação desta estrutura é possível a realização de uma análise estatísticas dos dados. Foram executadas as análises para ambos os conjuntos de dados.

Pasta e Link / Freq - Freq%	Freq	Freq %
ID_MENU		
007	28	0.00
ACSP	50	0.01
Abril.com	300	0.03
Abril.com - CDMA	16	0.00
Abril.com - Shows	80	0.01
Abril.com - Shows - CDMA	1	0.00
Academias	138	0.02
Almas Gemeas	238	0.03
Almas Gemeas - CDMA	48	0.01
Auto Market	25	0.00
Auto Market - GSM	49	0.01
Auxilio a Lista	121	0.01
Baixe Games	1131	0.13
Baixe Games de Futebol	489	0.06
Banco Real	339	0.04

Figura 3-7 – Estatísticas do arquivo com pastas e *hyperlinks*

A figura 3-7, acima, detalha a frequência dos itens do portal existentes no arquivo contendo as pastas e *hyperlinks*.

Link / Freq - Freq %	Freq	Freq %
ID_MENU		
007	24	0.01
ACSP	20	0.01
Abril.com	176	0.10
Abril.com - CDMA	10	0.01
Abril.com - Shows	56	0.03
Abril.com - Shows - CDMA	1	0.00
Almas Gemeas	146	0.08
Almas Gemeas - CDMA	30	0.02
Auto Market	18	0.01
Auto Market - GSM	36	0.02
Baixe Games	482	0.26
Baixe Games de Futebol	218	0.12

Figura 3-8 – Estatísticas do arquivo contendo *hyperlinks*

A figura 3-8, acima, detalha a frequência dos itens do portal existentes no arquivo contendo apenas os *hyperlinks*. Vale observar os baixos percentuais de frequência geral para os menus.

Após a importação dos dados e envio para o servidor central, o primeiro passo é a criação dos datasources para então iniciar a mineração com as técnicas aplicáveis.

Para este problema a primeira técnica aplicada foi a clusterização dos dados contendo os acessos a pasta e *hyperlinks* para identificar a segmentação dos usuários.

Ao utilizar um máximo de cinco clusters o resultado exibiu um cluster com 85,1% de usuários com interesses distintos e dos restantes, dois significativos: um envolvendo conteúdo erótico, 9,4% e outro envolvendo notícias e esportes, 4,2%.

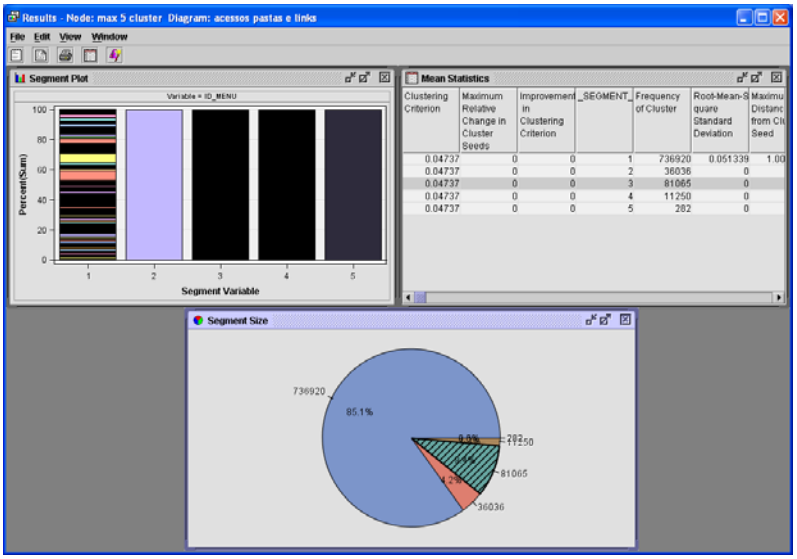

Figura 3-9 – Clusterização com cinco clusters obtidos através do SAS

Mais uma análise foi realizada, mas desta vez sem limite de cluster. Com este cenário foram identificados ao todo 42 clusters. Comparando com a clusterização anterior, o grupo de usuários com interesses distintos reduziu para 44,1%. Praticamente se mantivera, com interesses em conteúdo erótico, 9,37%, e com interesses em notícias e esportes, 4,2%. Entretanto, foi identificado mais dois cluster significativos, um com um grupo de usuários com interesse no conteúdo erótico específico do site explícita, 5,22%, e outro com interesse em portais, 4,4%.

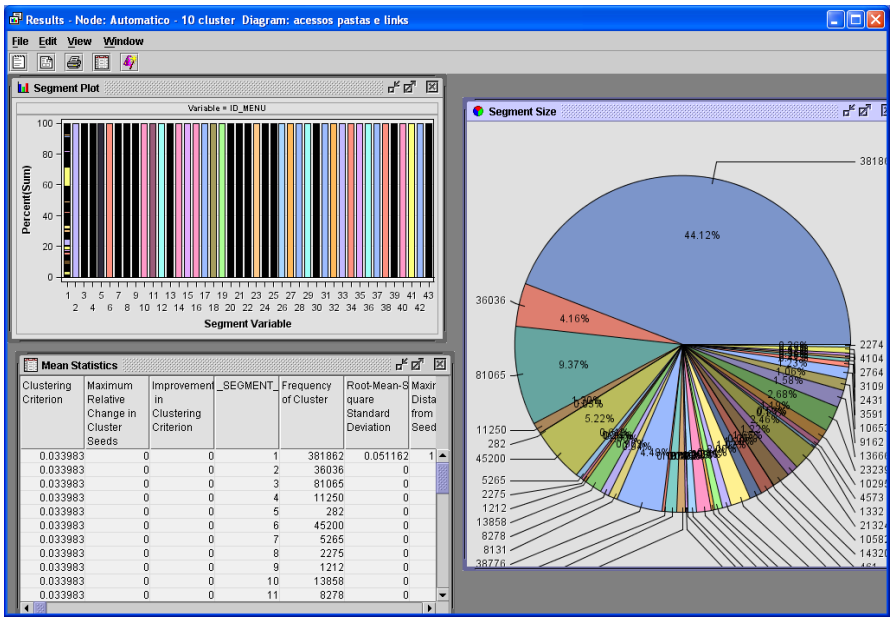

Figura 3-10 – Clusterização sem limites de clusters obtidos através do SAS

Segundo a avaliação de especialistas do WAP, a segmentação do público em busca de conteúdo erótico e notícias e esportes era esperada. Entretanto, a segmentação do público interessado em downloads através do Vivo Downloads, representado pelo cluster do Vivo Play, mostrou bastante inferior ao que se acreditava.

Após a clusterização, o próximo passo para descobrir o que o usuário acessa em conjunto foi à aplicação de regras de associação. Para a execução do algoritmo de associação foram utilizados os seguintes parâmetros quantitativos:

- Dado o baixo suporte apresentado no sumário de dados no momento de importação, foi utilizado 0,02% para suporte mínimo;
- Para confiança foi utilizado 80%;
- Foi configurado o máximo de três itens frequentes;
- Foram solicitadas no máximo 1.000.000 de regras.

Foram descobertas 279 regras de associação.

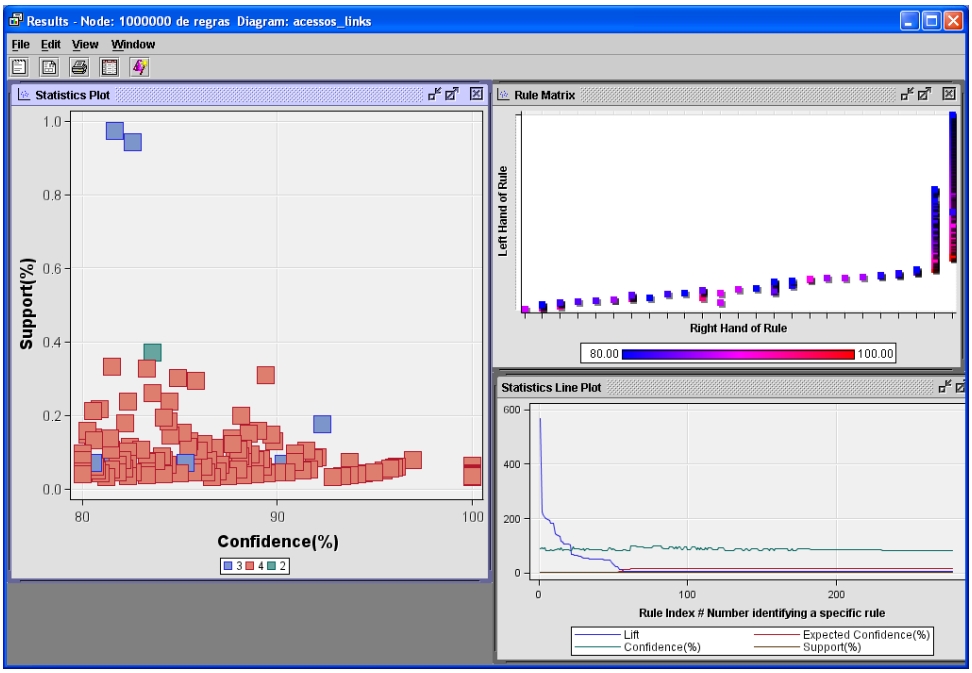

Figura 3-11 – Regras de associação obtidas através do SAS

O resumo de execução foi:

Tabela 3-1 – Resumo da extração das regras de associação

Índice	Mínimo	Máximo
Confiança esperada	0,1551741	16,81699
Confiança	80,000	100,000
Suporte	0,0315197	0,9722626
LIFT	4,757093	568,621324

A Confiança esperada identifica o número de transações consequentes dividido pelo número total de transações. O suporte é a frequência que uma dada combinação ocorre. Confiança é o percentual de casos nos quais, dado um "antecedente" ocorre um consequente. LIFT é igual à confiança dividido pela Confiança esperada.

Para identificar o grau de interesse, em conjunto com o especialista, além de suporte, confiança, e número de itens da regra, foram utilizadas as medidas qualitativas referentes ao conhecimento inesperado e acionável.

No geral, as regras apresentaram fortes relacionamentos entre: conteúdos eróticos; entre notícias e esportes, como associações entre sites com informação do clima ou jornais online. Abaixo estarão descritas as regras consideradas como inesperadas e acionáveis.

A regra "SEXY & REVISTA SET ==> PLAYBOY" obteve 85,71 de confiança, 15,34 de confiança esperada, 0,07 de suporte e 5,58 de LIFT.

Esta regra foi identificada como inesperada dado que o item "Revista Set" e os itens "Sexy" e "Playboy", além de estarem em lugares bastante distintos do menu, possuem finalidades distintas. O conteúdo "Revista Set" encontra-se no caminho da árvore "Guias/Lazer ➔ Cinema", enquanto os conteúdos "Sexy" e "Playboy" encontram-se em "Sexo/Adulto ➔ Revista".

A ação recomendável, obtida a partir do conhecimento descoberto, é incluir o menu "Revista Set" em "Sexo/Adulto ➔ Revista".

A regra "SEXY & Veja SP ==> PLAYBOY" obteve os 84,21 de confiança, 12,18 de confiança esperada, 0,04 de suporte e 6,91 de LIFT. A regra "SEXY & CONTIGO ==> PLAYBOY" obteve os 90,00 de confiança, 15,34 de confiança esperada, 0,04 de suporte e 5,87 de LIFT. Estas regras seguem também a linha da regra anterior, embora estejam em caminhos distintos no menu e possuam finalidades distintas, possuem associação. A ação será também a inclusão destes itens em um menu "Sexo/Adulto ➔ Revista"

A regra “VIVO TONS E IMAGENS & VIVO PLAY / VIDEOS & VIDEO GAMES ==> SPIN - TOQUES E IMAGENS – GSM” obteve 83,33 de confiança, 12,18 de confiança esperada, 0,04 de suporte e 6,84 de LIFT. Os menus “VIVO TONS E IMAGENS” e “VIVO PLAY / VIDEOS” encontram-se na pasta “tons e imagens” enquanto o menu “VIDEO GAMES” encontra-se na pasta “Guias e Lazer” e “SPIN - TOQUES E IMAGENS – GSM” encontra-se na página inicial.

A ação recomendável, obtida a partir do conhecimento descoberto, é adicionar os menus “VIDEO GAMES” e “SPIN - TOQUES E IMAGENS – GSM” a pasta “tons e imagens”.

3.4. Potencializar a navegação a partir de recomendações no portal WAP

O segundo ponto a ser tratado consiste em tentar potencializar a navegação dos clientes através da apresentação de conteúdos desconhecidos, no portal WAP, mas que são utilizados por clientes que navegam no portal.

Dada as limitações de formatação do WAP e as limitações de interação com os terminais quanto à digitação e visualização dos textos a busca por conteúdos é impactada. Assim sendo, a apresentação de conteúdos desconhecidos, mas consumidos por clientes com interesses comuns, se apresenta com uma possível solução para tal problema.

A aplicação de técnicas de mineração de dados para prover sistemas de recomendação foi bastante explorada. MOBASHER (2000) classificaram os sistemas de recomendação em três grupos: sistemas com filtros manuais, sistemas com filtros baseado em conteúdo e sistemas com filtros colaborativos.

Sistemas com filtros manuais são aqueles que a partir da configuração de um administrador apresentam diferentes conteúdos para o cliente. Sistemas com filtros baseado em conteúdo baseiam- nas similaridades dos conteúdos que são apresentados para perfis obtidos dos usuários. Sistemas com filtros colaborativos baseiam-se nas preferências e avaliações dos usuários a através de um mecanismo de co-relacionamento recomenda o conteúdo para o cliente.

Sistemas com filtros manuais são complexos, rígidos e caros para serem alterados. Sistemas com filtro baseado em conteúdo e filtro colaborativo propõem dinamicidade na recomendação.

Embora os dois últimos citados utilizem filtros dinâmicos, os filtros baseados em conteúdo focam em "o que o usuário está interessado", enquanto filtro colaborativo foca em "quem mais está interessado".

Vários trabalhos foram propostos para atender a filtros colaborativos. Alguns trabalhos apresentam frameworks que trabalham com recomendação *offline*, *online* e mistas com diferentes técnicas de mineração.

PERKOWITZ e ETZIONI (1998) propuseram um framework para indicação de diferentes páginas baseado em gerar a recomendação baseado em clusterização de sessões dos usuários MOBASHER *et al.* (2000) propuseram a personalização a partir de um framework que computa as recomendações, através de clusterização de sessões e páginas juntamente com associação com itens frequentes e, online utiliza o conhecimento descoberto para recomendar.

MOBASHER *et al.* (2001) propuseram um framework de recomendação baseado em regras de associação. Segundo MOBASHER *et al.* (2001) os trabalhos que propõem uma recomendação online baseados em clusterização, mesmo com o

processamento *offline*, não apresentam um desempenho bom. Os trabalhos que utilizam padrões sequenciais têm uma abrangência limitada. Os trabalhos baseados em regras de associação apresentam um bom desempenho enquanto mantêm uma boa abrangência e uma boa acurácia.

Para o problema de potencializar a navegação a partir do portal, a solução mais indicada é a utilização de um sistema baseado em filtro colaborativo e baseado em regras de associação.

3.4.1. Proposta da arquitetura do sistema de recomendação

MOBASHER *et al*. (2000) apresentaram um framework onde: o descobrimento do conhecimento era feito *offline* a partir de clusterização de sessões, de visualização de páginas e itens frequente; o site *web* personalizava o conteúdo a partir de requisições enviadas para um engenho que conhecia os padrões descobertos.

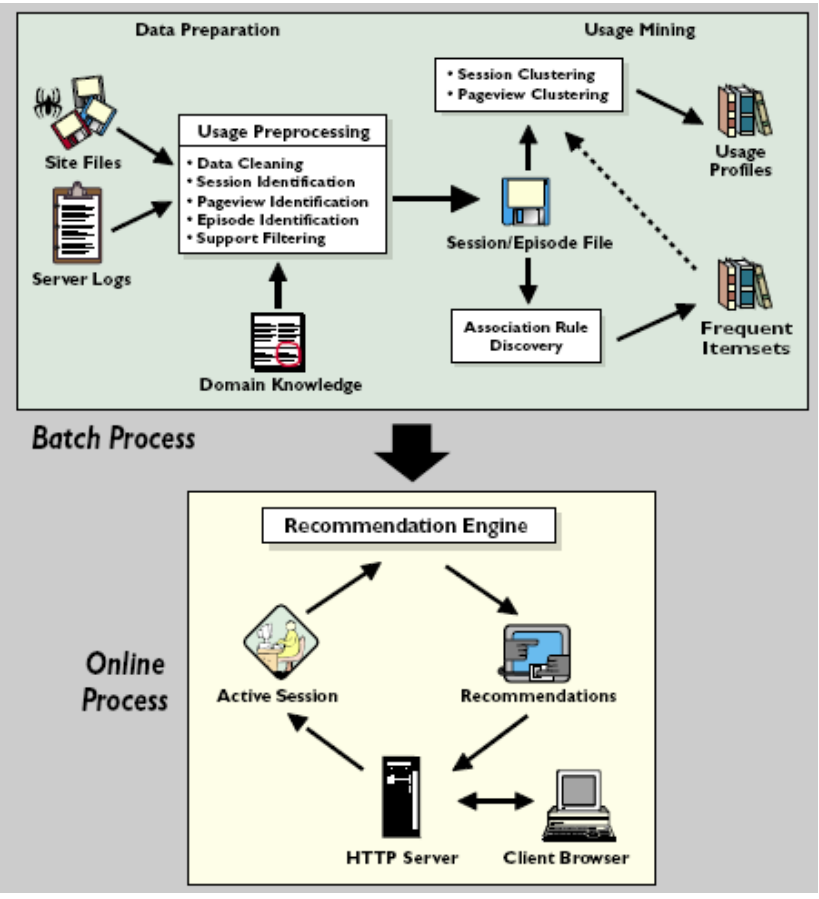

Figura 3-12 – Framework proposto por MOBASHER *et al.*(2000)

MOBASHER *et al.*(2001) apresentaram um framework onde são extraídas regras de associação *offline*, que são armazenados em uma estrutura de árvore que é lido por um componente online para recomendar o conteúdo para o cliente.

Para o desenho do sistema de recomendação será utilizado a linha de processamento *offline* com a utilização de regras de associação. No entanto, existem as seguintes preocupações:

- O tráfego a ser computado envolve o tráfego do portal e de toda a internet. Portanto, não se trata de recomendação de itens do site e sim recomendar itens que acontecem juntos no portal e na internet.
- As associações deverão ter como predecessor os acessos ao portal e sucessor os acessos a internet
- O título de um determinado acesso à internet não é conhecido e será necessário recuperar o conteúdo da página e interpretá-lo para identificação do título. Para esta recuperação, como alguns sites WAP na internet somente permitem acesso pelo WAP *Gateway*, esta recuperação implicará em efetuar as requisições através do WAP *Gateway*.

Dado tais requisitos, a seguinte arquitetura é proposta:

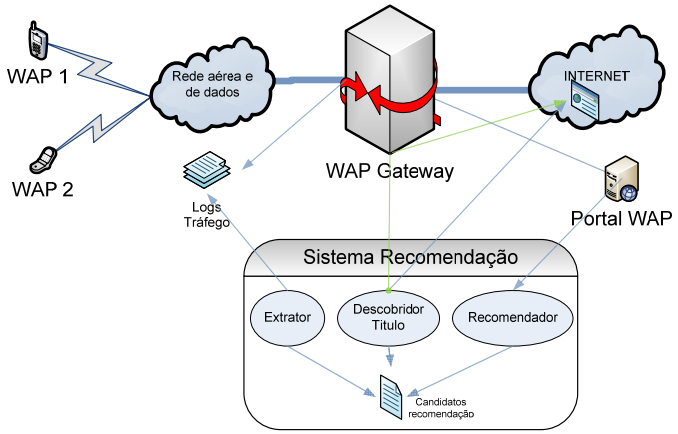

Figura 3-13 – Proposta do engenho de recomendação para o WAP

O elemento Extrator será o responsável por recuperar as transações do WAP *Gateway*, pré-processar, extrair os itens frequentes e as relações de acesso ao portal e a internet.

O Descobridor de títulos será responsável por acessar os sites na internet, recuperar os títulos e os adicionar aos respectivos sites.

O Recomendador será responsável por armazenar em memória os itens frequentes, os relacionamentos e atender as requisições por recomendação do portal.

3.4.1.1 Extrator

Este módulo será o responsável por recuperar as transações do WAP *Gateway*, processá-las e extrair os itens frequentes que serão candidatos a recomendação.

O protótipo deste módulo foi criado utilizando Java 1.6. Para o desenvolvimento foi utilizada a IDE eclipse 3.4.2.

3.4.1.1.1 Transações do WAP *Gateway*

O WAP *Gateway* armazena todas as transações que trafegam por ele em um arquivo texto com as informações separadas pelo caractere "|". Para cada transação são armazenadas trinta e três informações que tem por objetivo registrar dados do cliente, registrar dados do site na internet, registrar o ciclo da requisição e reposta do site. Para o sistema de recomendação serão utilizadas as informações descritas abaixo:

- Campo I, "*Gateway receiving request time*". Representa a data da requisição do telefone móvel. Este campo possui uma sequência de 14 caracteres expressos na seguinte formatação: ddmmaaaahh24miss, o "dd" representa o dia com dois dígitos, mm o mês com dois dígitos, "aaaa" o

ano com quatro dígitos, “hh24” a hora com dois dígitos e em vinte e quatros horas, “mi” os minutos com dois dígitos e “ss” os segundos com dois dígitos.

- Campo III, “*Mobile phone’ s number of user*”. Representa o número do telefone que efetuou a requisição. Este campo possui uma sequência numérica de dez dígitos.

- Campo IV, “*URL requested by user*”. Representa a URI do site solicitado pelo telefone móvel. Este campo é uma sequência de caracteres não limitada.

- Campo X “*Destination IP address requested by user*”. Representa o host da URI do site solicitado.Este campos é uma sequência caracteres não limitada.

- Campo XVII “*Responding state code*”. Representa o código HTTP de resposta do site requisitado.

- Campo XXIII “*Information content type*”. Representa o tipo de mídia informada pelo site.

Estas transações são escritas no arquivo texto e este arquivo é fechado por um período determinado ou por atingir um determinado tamanho. Após tal fechamento, as transações podem ser consultadas.

A seguir uma transação registrada em um arquivo é exemplificada: “*30062009235459|-10800|2198888706|21|GSM|http://termwap.vivo.com.br:80/img/layout/100923_240.gif|10.54.127.124|4|SonyEricssonK850i/R1EA Browser|termwap.vivo.com.br|672.768000|5754.880000|160|WAP2.X|GET|complete|2*

00|0|80|200.220.252.141|2009-06-30 23:54:59:210|2009-06-30 23:54:59:370|image/gif|10.23.129.134|HTTPBrowser|termwap.vivo.com.br|80|552198888706|5.620|0.657|2009-06-30 23:54:59:200|UMTS|"http://wap.sonyericsson.com/UAprof/K850iR101.xml""

3.4.1.1.2 Pré-processamento das transações

Na fase de pré-processamento das transações, é possível transpor algumas dificuldades existentes de pré-processamento em virtude das propriedades do cenário do WAP *Gateway*.

A identificação do usuário é realizada através de seu identificador. Portanto não são necessárias atividades para correlacionar IPs, *user-agents* ou informações de sessão do cliente.

Não são necessárias atividades para filtragem de acessos automáticos de robôs visto que o acesso através ao WAP *Gateway* dever ser efetuado por um telefone móvel autenticado na rede celular. No entanto, como nem todos os acessos são de cliques do usuário, tais acessos precisam ser removidos. Exemplos destes acessos são: imagens, estilos e javascript.

Para realizar tal filtragem é utilizada a informação do tipo de mídia existente no campo XXIII do arquivo de transação do WAP *Gateway*. Somente serão consideradas transações como válidas as que possuam neste parâmetro os nomes “wml”, “html”, “xhtml”,”text/plain”.

Estes conteúdos são conhecidos como MIME, *Multipurpose Internet Mail Extensions*. Eles são formatos: codificação, tipos de caracteres, etc., de mensagens

trocadas na internet e são regulados por normas mantidas pelo IANA, *Internet Assigned Numbers Authority*.

A identificação dos conteúdos a serem tratados representa o tratamento das páginas com formatação HTML, WML e XHTML. O tratamento de tais nomes de conteúdo foi obtido a partir da listagem de tipos de mídia mantida pelo IANA.

Além da remoção dos acessos automáticos realizados pelo cliente é necessário remover todos os acessos que não tiveram sucesso. Para tal, foi tratado apenas o código de resposta do HTTP 200.

O HTTP é um protocolo de aplicação utilizado para a transferência das páginas *web* na Internet. Para toda requisição efetuada a um site a resposta do site possui no cabeçalho o código de retorno.

O código 200 é a indicação de sucesso. A utilização de tal código foi a partir da análise dos códigos HTTP existentes, cujo estão descritos na RFC2608, *Request for Comments*, administrada pelo IETF, *Internet Engineering Task Force*.

Após a filtragem dos cliques válidos e com sucesso a próxima tarifa é criar as sessões, ou seja, mapear as transações de um determinado usuário que ocorreram em período de tempo. Para tal, é preciso identificar o usuário e encontrar a diferença de tempo entre dois acessos de um mesmo usuário.

A identificação do usuário é realizada a partir do campo III, que representa o seu identificador. Para o cálculo da diferença entre dois acessos do usuário, a informação do campo I é transformada para um número de milissegundos referente a 1 de Janeiro de 1970 0 hora, 0 minuto e 0 segundo. Com este número é calculada a diferença e avaliado se esta diferença é menor que a quantidade de tempo da sessão do cliente.

Para a criação da sessão, além da variável do tempo de sessão, também é utilizado o número de cliques mínimos para se considerar uma sessão válida. Assim, sessões que não possuam um mínimo de acesso são descartadas.

Como o problema reportado trata-se de correlacionar acessos ao portal com acessos a internet, devem-se ser descartadas também sessões que não possuam uma quantidade mínima de acesso ao portal. Esta quantidade deve ser no mínimo o tamanho da janela utilizada pelo componente Recomendador.

Para a identificação se uma transação é ou não do portal é utilizado o campo X que representa o endereço do HOST da URI. Como o portal pode ser acessado a partir de vários endereços de HOST distintos, o Extrator recupera uma listagem e as compara com a transação sendo analisada.

O portal utiliza uma nomenclatura para sua URL, que consiste em informar o identificador do menu requisitado pelo cliente, conforme exibido na figura 3-14. Devido a isto, o relacionamento do portal com os acessos a internet podem ser obtidos apenas a partir das transações do WAP *Gateway*. Portanto, as URL das transações do portal são substituídas pelos seus respectivos identificadores.

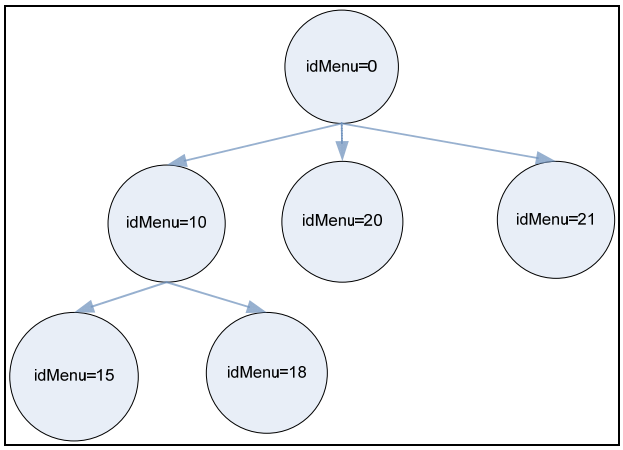

Figura 3-14 – Armazenamento das associações utilizando o identificador do menu

Em uma análise dos dados identificou-se que existem uma série de cliques seguidos em um mesmo site, assim como foi reportado na análise de transações do portal. Portanto, são removidas as transações que apresentam acessos seguidos a um mesmo endereço.

Em virtude de estar se processando todo o tráfego da internet e ser necessária a recuperação de títulos a partir do endereço, é preciso descartar as transações que representem acessos a sites dinâmicos, que podem implicar em parâmetros dinâmicos na URL, quando utilizado o método HTTP GET, conforme descrito na RFC2608.

Para remover tais transações, é configurado um suporte mínimo para as transações da internet. Assim, as transações dentre as sessões criadas que não possuam o suporte são descartadas.

Para as transações no portal também é aplicado o descarte para os menus que não possuam uma quantidade mínima de acesso. Isto é em virtude que as solicitações ao portal para menus que não existem são processadas e é apresentada uma interface indicando ao usuário para ir para a página inicial do portal.

Isto implica em que solicitações para menus inexistentes, o portal gere uma página que possuirá o código HTTP de sucesso. Portanto, transações ao portal sem um determinado suporte mínimo também são descartadas.

3.4.1.1.3 Criação do relacionamento dos acessos do portal e internet

Entre os itens do portal é criado um relacionamento a fim de identificar a sequência de navegação do cliente para posteriormente utilizar esta sequência para recomendação. Para cada item do portal é armazenada a quantidade de vezes que o mesmo aparece entre as sessões dos usuários.

Ao identificar uma sequência de acesso de um determinado tamanho, que deve ser no mínimo do tamanho da janela de recomendação utilizada no Recomendador, as transações sequentes a internet são relacionadas a tal sequência de navegação do portal.

Para o relacionamento entre os itens do portal e a internet os acessos são contabilizados para serem utilizados no cálculo do suporte e confiança para recomendação.

O componente de recomendação por ser online, necessita ter um desempenho referente a tempo de resposta e atendimento a requisições grandes. Em virtude disto, ao identificar uma sequência de acessos aos itens do portal e em seguida a internet, este acesso é computado para todos os itens da sequência envolvidos do portal. Isto é realizado para que no Recomendador, online, não seja necessário efetuar uma varredura recursiva na árvore para identificação de itens recomendados.

A associação da sequência de itens do menu ao site da internet somente ocorre se o acesso a sequência e o acesso ao site for subsequente.

Para exemplificar, suponha que exista a seguinte sequência de transações em uma sessão criada a partir do pré-processamento: idMenu=1, idMenu=4, idMenu=15, siteI, siteII, idMenu=4, idMenu=15, siteII.

Considerando um tamanho de janela igual a 3, ou seja, três acessos simultâneos ao portal, o menu idMenu=1 será relacionado com o idMenu=4 e este com o idMenu=15 e será adicionado um acesso a cada menu. Em seguida os três serão associados ao site I e será contabilizado na relação do idMenu=1 para siteI um acesso, na relação do idMenu=4 para o siteI um acesso e na relação do idMenu=15 para o siteI um acesso.

Esta quantidade no relacionamento entre o item do portal e o site da internet será utilizada para computar o suporte e a confiança no Recomendador. Estas quantidades na relação somente são computadas uma vez para sessão, ou seja, se esta sequência se repetisse na mesma sessão não seria computada.

Para as transações seguintes, idMenu=4, idMenu=15, siteII., não serão computados os relacionamentos tanto dos itens do portal quanto deste para internet visto que a sequência não possui o tamanho mínimo da janela.

Sendo assim, é possível expressar que seja S={s1, s2, s3, ..., sn} o conjunto de sessões recuperadas das transações do WAP *Gateway*. Para cada sessão, existe um conjunto de transações T={t1, t2, t3,..., tn}, onde uma dada transação pode ou não ser referente ao portal. Seja P={p1,p2,p3,..pn} os acessos referentes ao portal e I={i1, i2, i3,..., in} os acessos referentes a internet. Seja TP={p1,p2,p3, ..., pn} um subconjunto de T com as transações no portal. Seja tj o tamanho da janela. Considera-se que os menus do portal se relacionam entre si e com uma transação na internet se existe em uma sessão s uma transação t Є I que preceda o subconjunto TP, onde $|TP| \geq tj$.

Seja $QR_{(pi)}$ a quantidade de vezes que o item p do portal se relaciona com item i da internet, e seja QP a quantidade de vezes que o item do portal se relacionou com algum item da internet no conjunto S. A confiança de uma relação $p \rightarrow i$ é dado por $Conf(p \rightarrow i) = QR_{(pi)}/QP$. O suporte de uma relação $p \rightarrow i$ é dado por $Sup(p \rightarrow i) = QR(pi)/|S|$.

Para armazenar o relacionamento dos itens do portal, cujo é um grafo, e o relacionamento dos itens do portal com a internet, que se torna uma aresta com peso, a estrutura de dados descrita é criada.

A seguir é apresentada uma figura descrevendo a estrutura de armazenamento do relacionamento dos itens do portal com a internet.

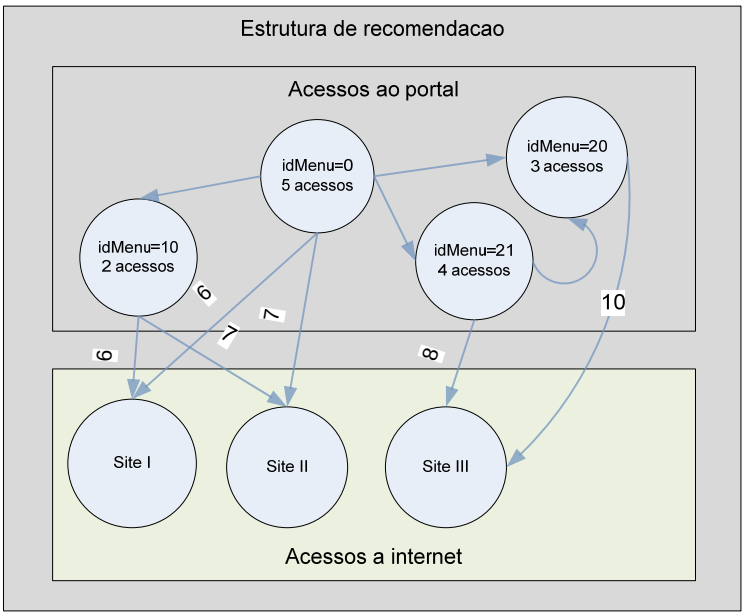

Figura 3-15 – Estrutura de relacionamentos do menu com a internet.

Além dos relacionamentos descritos a estrutura de dados também armazena o total de sessões. A partir disto o suporte entre um item do portal e um site na internet é dado pela quantidade da aresta entre o item do portal e o site da internet, divido pelo total de sessões. A confiança é dada pela quantidade da aresta entre o item do portal e o site da internet, divido pela quantidade de acesso ao item do portal.

Após a criação desta estrutura a mesma é gravada em um arquivo texto com a seguinte formatação: @SuporteGeral indicando a quantidade de sessões existentes; @Menu indicando a lista de menus do portal; @RelMenu a matriz indicando o relacionamento entre os itens do portal, onde a diagonal principal armazena a quantidade de acessos ao item; @LinkRecomendado a lista de links e títulos da internet e @RelMenuLink a matriz indicando o relacionamento e quantidade dos itens do portal com itens da internet.

A seguir é exemplificado um arquivo criado pelo Extrator.

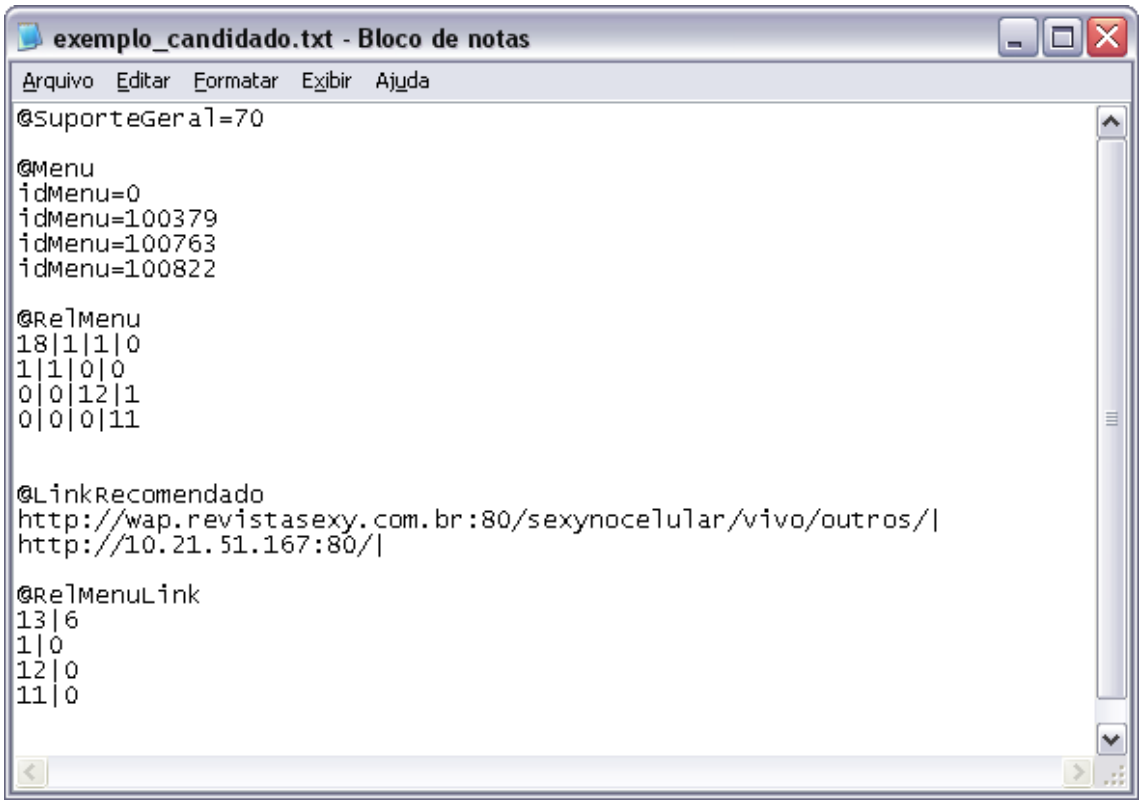

Figura 3-16 – Arquivo criado pelo Extrator.

3.4.1.2 Descobridor de títulos

Após o Extrator, é conhecido o relacionamento entre os itens do portal a fim de que a partir de uma janela de transação no portal seja possível identificar quais *hyperlinks* na internet também foram acessados. Entretanto, a exibição de um hyperlink para o usuário não faz sentido, portanto é necessária a recuperação de informação daquele hyperlink.

Como o problema restringe-se a uma prova de conceito, não será necessária a elaboração de um sistema de recuperação de informação, mas sim um sistema para tentar mapear apenas os títulos dos *hyperlinks*. Portanto, a função deste componente é a partir dos *hyperlinks* descritos pelo Extrator recuperar o título dos sites.

O protótipo deste módulo foi criado utilizando Java 1.6. Para o desenvolvimento foi utilizada a IDE eclipse 3.4.2. Foram utilizados os frameworks HTML Unit e RADIUS4J.

3.4.1.2.1 Restrições de acesso a sites WAP

Diferentemente da internet, alguns sites WAP somente estão disponíveis a partir do WAP *Gateway*, isto implica que as requisições originadas para tais sites, deverão ser realizadas através do WAP *Gateway*. Isto implica que o Descobridor necessitar enviar as requisições para a recuperação dos títulos através do WAP *Gateway*.

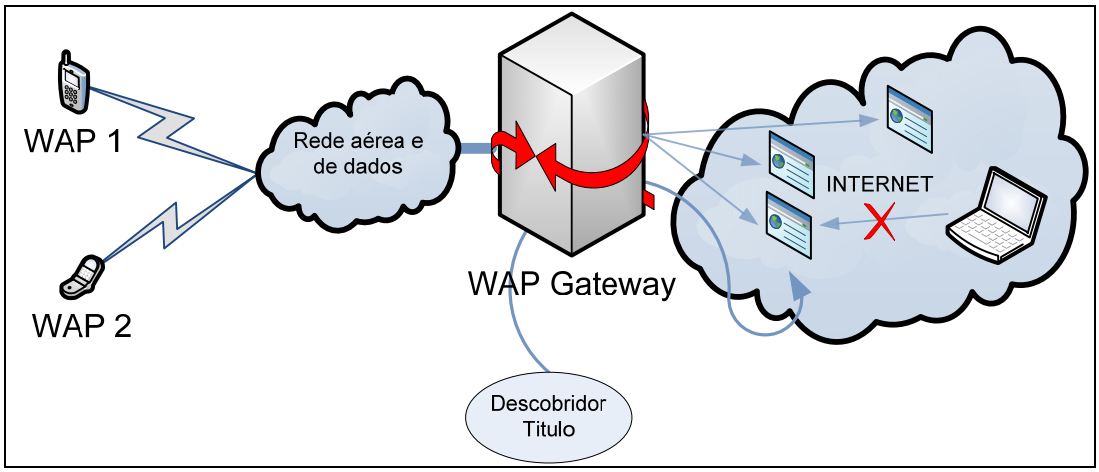

Figura 3-17 – Modelo de acesso a sites da Internet para recuperação do título.

Para realizar tais requisições é necessário que o Recuperador de títulos simule um usuário autenticado na rede celular e envie as requisições utilizando o WAP *Gateway* como um Proxy. Para tal é necessário que este Descobridor antes de enviar as requisições HTTP simule uma autenticação na rede.

Para enviar tais requisições o Recuperador deverá simular um usuário autenticado no AAA. Esta autenticação é realizada através de mensagens enviadas do

AAA para o WAP *Gateway*. Esta mensagem segue o protocolo RAIDUS, em inglês, *Remote Authentication Dial In User Service*, definido na RFC2865.

Portanto o Recuperador antes de iniciar as requisições deverá enviar uma mensagem RADIUS para o WAP *Gateway* a fim de simular a autenticação de um usuário e em seguida enviar a requisição HTTP.

3.4.1.2.2 Requisição e recuperação do título

Após o recebimento da resposta HTTP do site o Recuperador executará a recuperação do título somente se o status HTTP de requisição foi 200 indicando o sucesso.

Para as requisições com sucesso é analisado o tipo de conteúdo informado pelo site, e é realizado um processamento distinto para conteúdos WML e XHTML e HTML.

Para WML executa uma busca no conteúdo da resposta. Para tal, o WML é considerado como um XML, em inglês, *eXtensible Markup Language*, que é um padrão de formatação de arquivos textos, cujo padrão segue uma recomendação mantida pelo W3C, em inglês, *World Wide Web Consortium*.

Para eleger os elementos e atributos que serão pesquisados foi utilizada a especificação de WML mantida pelo OMA, em inglês, *Open Mobile Alliance*. Os elementos e atributos eleitos para serem pesquisados foram: o atributo “title” do elemento “card”, o valor do elemento “strong” e o valor do elemento “b”.

Primeiramente é pesquisado o elemento “*card*” para recuperar o atributo “*title*”. Se a busca falhar, é efetuada uma nova busca pelo valor do atributo “*strong*”. Senão encontrar é efetuado uma busca pelo valor do atributo “b”.

Para as páginas HTML e XHTML, a resposta também é tratada como um XML. Para eleger os atributos a serem pesquisados foram utilizadas as recomendações do HTML 4.01 e do XHTML 1.0 mantidas pelo W3C. Os elementos eleitos foram o "title", "h1", "h2", "h3", "h4", "h5" e "h6".

Primeiramente é pesquisado o valor do elemento "title". Senão encontrar é realizado uma busca pelos valores dos elementos "h1", "h2", "h3", "h4", "h5" e "h6" na respectiva ordem.

Caso o título correspondente ao hyperlink não seja encontrado o Recuperador configurará o título como vazio e o Recomendador descartará o site durante a recomendação. Ao final da recuperação de títulos o arquivo contendo a estrutura de informação é atualizado com os títulos.

A figura a seguir exemplifica o arquivo gerado pelo Recuperador.

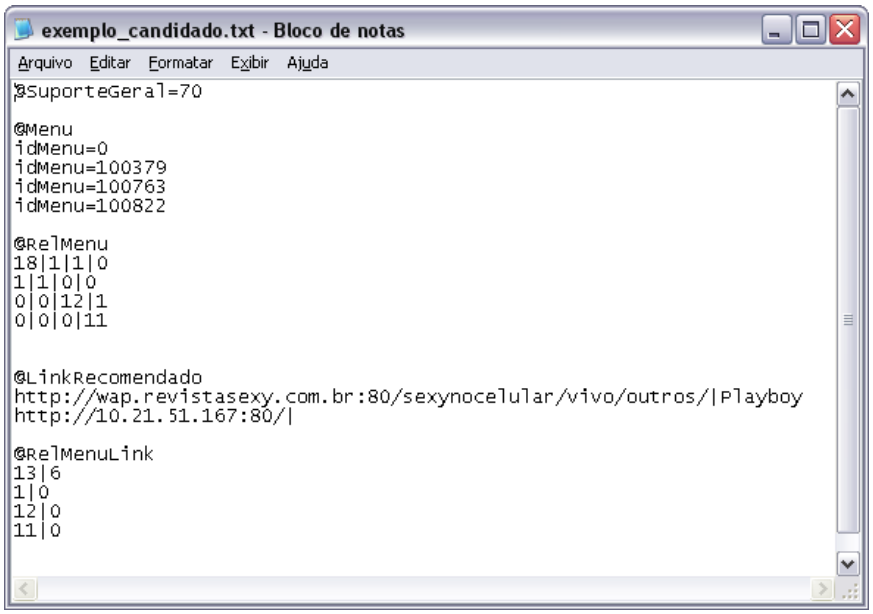

Figura 3-18 – Exemplo do arquivo criado pelo Recuperador de títulos.

3.4.1.3 Recomendador

O Recomendador é uma aplicação que disponibiliza o serviço de consulta a itens recomendáveis para o Portal. A partir do arquivo criado pelo Extrator, atualizado pelo Recuperador de títulos, o Recomendador carrega as configurações em memória e atende as requisições do Portal.

O protótipo deste módulo foi criado utilizando Java 1.6. Para o desenvolvimento foi utilizada a IDE eclipse 3.4.2.

Como foi apresentado anteriormente, associações envolvendo domínios com várias abstrações possuem particularidades que precisam ser consideradas. O Portal por se tratar de uma hierarquia de menus está incluso neste tipo de domínio. Portanto, a proposta apresentada para este problema é que o Portal ao solicitar a recomendação informe o suporte e confiança dinamicamente de acordo com o nível do último menu acessado pelo cliente. Em virtude disto, nem o Extrator e nem o Descobridor de títulos aplicaram filtragens referentes aos índices de suporte e confiança.

Para a prova de conceito, foi criado um servidor TCP com um protocolo síncrono, onde dada uma requisição do cliente o servidor devolve a resposta. Não foi previsto para este protocolo funcionalidades de autenticação, autorização, etc. Sendo assim, somente existem as requisições de: recomendação, manutenção da conexão e sair.

Após o envio do comando o servidor responde um status da resposta e a resposta. Os possíveis status são: 200 para sucesso; 304 para comando não reconhecido; 305 para argumento não informado; 306 para argumentos inválidos; e 401 para nada a recomendar.

A mensagem de requisição da recomendação, enviada pelo cliente, é uma sequência de texto, contendo as informações de suporte, confiança, número máximo de itens recomendáveis e a janela de acesso do cliente.

A requisição possui a seguinte formatação: sup=X;conf=Y;numItem=Z; janela= idMenu=A|idMenu=B|idMenu=C.

O parâmetro “sup” indica o suporte mínimo, o parâmetro “conf” indica a confiança mínima, o parâmetro “numItem” indica a quantidade máxima de itens a serem informados e o parâmetro “janela” indica a janela de navegação do cliente, onde os menus são separados pelo caractere “|”. Não existem restrições para o tamanho da janela informado. Entretanto esta janela informada deve estar coerente com a janela utilizada no Extrator.

A partir desta requisição o servidor consulta a estrutura de dados descrita anteriormente, e informa como resposta uma sequência de caracteres contendo o hyperlink do site e o seu título. Caso sejam recomendados vários itens haverá a quebra de linha indicando um novo item. Um exemplo da reposta seria:

http://www.coppe.ufrj.br/|COPPE-UFRJ Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa em Engenharia

http://www.coc.ufrj.br/|Programa de Engenharia Civil - COPPE/UFRJ

A seguir é apresentada uma figura descrevendo o servidor iniciado e uma conexão telnet aberta com o servidor e enviando as mensagens de solicitação de recomendação segundo o protocolo descrito.

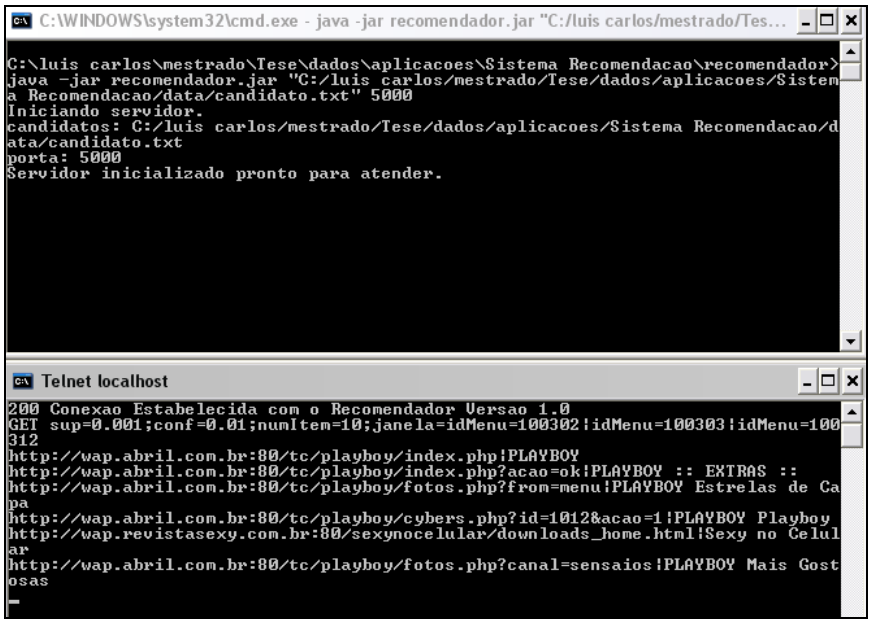

Figura 3-19 – Exibição de comunicação com o servidor de recomendação

3.4.2. Análises e Resultados

No pré-processamento das transações foi utilizada uma janela de 3 acessos no portal para se considerar uma sessão válida. Foi utilizado o parâmetro de 15 minutos para o tratamento de uma sessão. Foi utilizado o suporte de 2 acessos para os itens do portal e 40 acessos para os itens de internet.

O tráfego bruto possuía 6.950.696 transações. Após a filtragem dos dados: descarte de transações automáticas e com falha da requisição e, criação da sessão sem considerar acesso ao portal; restaram 1.278.052 referentes a 63.912 usuários distintos, que chamamos de sessões WAP válidas.

Após as demais filtragens: janela mínima de acesso ao portal e, suporte de acesso para os itens do portal e da internet; foram obtidas 133.711 transações referentes a 9.700 clientes distintos, que chamamos de sessões WAP-Portal válidas.

Ao executar o processo de associação do Extrator foram obtidas 516 URIs distintas. Ao executar a recuperação de títulos para esta lista de URI, o recuperador de

títulos obteve o título 390 URIs, ou seja, a lista a ser recomendada passou a conter 390 URIs.

Destas 390 URIs, nenhuma URI está cadastrada no Portal. Se assumir que embora as URIs não estejam diretamente cadastradas no portal, mas que seja possível alcançá-las a partir de um mesmo HOST, ainda assim existiriam 17 URIs não existente no Portal.

As 390 URIs representam 415.409 transações dentre as sessões do WAP válidas, ou seja, 32,50% das transações válidas do WAP não estão diretamente cadastradas no Portal.

Se a comparação for executada entre os HOSTs das transações recomendadas que não estão no Portal com os HOSTs das sessões WAP, o percentual é de 1,74%, 22238 acessos, representando 2,80 % dos clientes da sessão WAP. Destes 2,80% apenas 0,87% também navegaram no portal.

Na amostra analisada, a proposta de recomendação teria potencial para indicar a 14,30% dos clientes WAP um conteúdo inexistente no Portal. Isto proporcionaria o acesso ágil para o cliente a informações segundo seu interesse no portal.

Para a operadora representa um potencial de 14,30% de clientes acessarem 1,74%, ou seja, de incremento de 1,98 %.

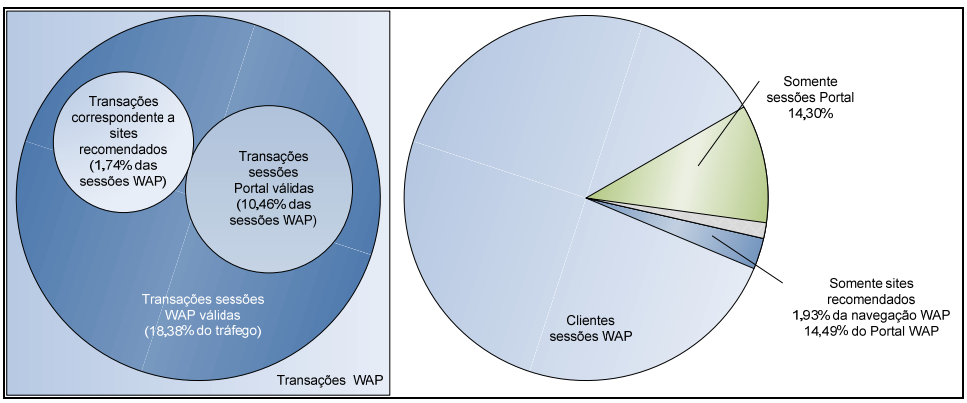

Figura 3-20 – Abrangência da recomendação em transações e clientes.

4. Trabalhos futuros

Trabalhos futuros na evolução do recuperador de títulos, módulo do sistema de recomendação, responsável pela obtenção da descrição do site a ser recomendado podem ser realizados. Podem ser adicionadas técnicas de recuperação da informação para refinar a identificação da descrição dos sites.

Trabalhos futuros na evolução do extrato, módulo do sistema de recomendação, responsável pela criação da associação de itens do portal com a internet pode ser realizadas a fim de considerar o peso de uma sequência de acesso ao em vez de apenas o último item do portal para ranquear a recomendação dos itens da internet.

Diferentemente da *web*, o WAP é um universo mais restrito. Isto implica em as operadoras procurarem parceiros de serviço para o provimento de conteúdo e com isto precisarem compartilhar receita do serviço. Portanto, é importante que o framework de recomendação possua também capacidades de filtros manuais gerenciados pelos a fim de permitir que os especialistas possam optar ou não por exibir determinadas recomendações. Com este mecanismo de filtro manual os especialistas do negócio, se necessário, poderão administrar recomendações que poderão ser realizadas.

5. Conclusão

Neste trabalho foi explicado o que é o serviço WAP, foi apresentada a arquitetura adotada na VIVO e nas operadoras brasileiras e descrita às limitações de utilização e prestação do serviço.

No cenário da operadora VIVO foi apresentada uma proposta para oferecer uma gama maior de informações e conteúdos para o cliente e para operadora uma opção para incrementar a receita do serviço.

O serviço WAP na VIVO consiste de um portal WAP, gerenciado por especialistas do marketing, que consiste em uma árvore de menus. As folhas destas árvores consistem de *hyperlinks* que apontam para sites na *web* contendo conteúdos para o WAP.

Todos os clientes, por padrão, são direcionados para o portal e a partir deste navegam para o restante da *web*. Entretanto os clientes podem digitar os endereços que desejam acessar em seu próprio telefone móvel e acessarem diretamente sem o intermédio do portal. Em virtude disto, é impossível manter no portal todas as informações desejadas pelos clientes.

A proposta consistiu em aplicar técnicas de mineração para estudar a organização da árvore do portal e propor adequações. Foi proposto também um mecanismo de recomendação de informação a partir da navegação do cliente no portal. O objetivo de tal proposta é estimular a navegação do cliente através da apresentação de menus co-relacionados e através da recomendação de informações que clientes com navegações similares no portal também acessaram na internet.

Para a análise e adequação do portal foram utilizadas técnicas de agrupamento e regras de associação. Para a recomendação foi proposto um framework baseado em

regras de associação responsável por extrair o conhecimento a partir do tráfego do WAP *Gateway* e disponibilizar tal conhecimento para ser consultado online pelo portal.

A partir da análise de uma amostra de dados foram apontadas associações recomendadas entre menus do portal. A partir desta amostra também foi calculado o potencial do framework de recomendação a atender novos clientes e aumentar o tráfego do serviço.

Segundo a última aferição da Anatel, em dezembro de 2008, existiam no Brasil 150,6 milhões de telefones celulares. Embora a evolução dos terminais aponte para dispositivos capazes de acessar a internet diretamente sem o intermédio do WAP *Gateway*, tais terminais são caros e estão restritos a uma pequena parcela da população.

Quase a maioria dos terminais existentes possui a capacidade para realizar o serviço WAP e consequentemente apta a acessar a *web*. Assim sendo, existem praticamente 150,6 milhões de potencias clientes de WAP. O estudo aqui realizado, embora focado no cenário da operadora VIVO, pode se estender a demais operadoras e prestadoras de conteúdo WAP, e consequentemente atingir a 150,6 milhões de usuários.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R., IMIELINSKY, T., SWAMI, A.(1993). "Mining assocation rules betwenn sets of items in large databases". *Proceedings of the ACM SIGMOD International Conference on management foR Data*, pp.207-216.

ANATEL. Disponível em http://www.anatel.gov.br/Portal/exibirPortalInternet.do.

BERNERS-LEE, T., HENDLER, J., LASSILA, O., 2001, "The Semantic Web", *Scientific Am.*, Vol. 279, No. 5

BAGLIONI, M., FERRARA, U., ROMEI, A., RUGGIERI, S., *et al.*,2003, "Preprocessing and mining web log data for web personalization", v. 2829, *Lecture Notes in Computer Science*, Springer, pp.237-249.

BORGES, J., 2000, *A Data Mining Model to Capture User Web Navigation Patterns*, Ph.D. thesis, Department of Computer Science, University College London, UK.

CHEN, M., PARK, J., YU, P., 1996, "Efficient data mining for path traversal patterns", *IEEE Transaction knowledge and Data Engineering*, v.10, n.2, pp. 209-221.

CHEN, M., PARK, J., YU, P., 1996, "Data mining for path traversal patterns in a Web environment", *Proceedings of the 16th International Conference on Distributed Computing Systems*, pp385-392, Hong Kong, Março.

COOLEY, R., MOBASHER, B., SRIVASTAVA, J., 1997, "Web Mining: Information and Pattern Discovery on the World Wide Web", *IEEE International Conference on Tools with Artificial Intelligence (ICTAI97)*, Newport Beach, CA, USA

COOLEY, R., TAN, P., SRIVASTAVA, J., 1999, "WebSIFT: The Web Site Information Filter System", *Web Usage Analysis and User Profiling Workshop*, San Diego, CA, USA

DILLY, R., 1995, *Data Mining - an introduction*, University of Belfast, Parallel Computer Centre - Queen's.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH P., 1996, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, v. 17, n. 3, pp. 37-54.

FIELDING, R., GETTYS, J., MOGUL, J., *et al.*, 1997, *Hypertext Transfer Protocol*, IETF.

HAN, J. E KAMBER, M., 2001, *Data Mining: Concepts and Techniques*, 1 ed. San Francisco, MorganKaufmann.

HAND, D., MANNILA, H., SMYTH, P., 2001, *Principles of Data Mining*, 1 ed.Cambridge, MIT Press.

HTML UNIT. Disponível em http://htmlunit.sourceforge.net/. RADIUS4J. Disponível em http://sourceforge.net/projects/radius4j/.

IANA, 2009, *MIME Media Types*. Dísponível em http://www.iana.org/assignments/media-types/.

KOSALA, R., BLOCKEEL, H., 2000, "Web Mining Research: A Survey", *SIGKDD Explorations*, vol. 2, n. 1, pp. 1.

LIU, B., 2007, *Web Data Mining exploring Hyperlinks, Contentes and Web Usage*, 1 ed. New York, Springer.

MARKOV, Z., LAROSE, D., 2007, *Data Mining the Web Unconvering patterns in Web Content, Structure and Usage*, John Wiley & Sons, New Jersey.

MOBASHER, B., DAI, H., LUO, T., et al., 2001, "Effective Personalization Based on Association Rule Discovery from Web Usage Data", *Communications of the ACM*, v.43, n.8, pp. 142-151.

MOBASHER, B., COOLEY, R., SRIVASTAVA, J., 2000, "Automatic Personalization Based On Web Usage Mining", Communication of ACM, Volume 43, Issue 8.

PERKOWITZ, M., ETZIONI, O., 1998, "Adaptive Web sites: automatically synthesizingWeb pages". *Proceedings of Fifteenth National Conference on Artificial Intelligence*, Madison, WI.

PEMBERTON, S.,AUSTIN, D., AXELSSON J., *et al.,* 2002, "XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition)".

PIERRAKOS, D., PALIOURAS, G., PAPATHEODOROU, C., *et al.*, 2003, "Web Usage Mining as a tool for personalization: a survey", *User Modeling and User-Adapted Interaction*, v. 13, n. 4 (Nov), pp. 311-372.

RAGGETT, D., LE HORS, A., JACOBS , I., 1999, "HTML 4.01 Specification".

RIGNEY, C., WILLENS, S., RUBENS, A., *et al*., 2000, Remote Authentication Dial In User Service, IETF.

SCIME, A., 2005, *Web Data Mining: Applications and Techniques*, 1 ed. Hershey, Idea Group Publishing.

SPILIOPOULOU, M., 2000, "Web usage mining for Web site evaluation", *Communications of the ACM*, v.43, n.8, pp. 127-134.

SRIVASTAVA, J., COOLEY, R., DESHPANDE, M., *et al*., 2000, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *SIGKDD*, Volume 1, Issues 2.

SRIVASTAVA, J., DESIKAN, P., KUMAR, V., 2005, "Web Mining – Concepts, Applications and Research Directions", v. 180, *Studies in Fuzziness and Soft Computing*, Springer, pp. 275-307.

WIRELESS APPLICATION PROTOCOL FORUM, 1998, "Wireless Application Protocol Wireless - Markup Language Specification

ZAÏANE, O., 1999, *Resource and Knowledge Discovery from the Internet and Multimedia Repositories*, Ph.D. thesis, School of Computing Science, Simon Fraser University, Canada.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo