EDER BRUNO FONSECA

TRACK4MINE: UMA NOVA PLATAFORMA INTELIGENTE DE COLETA, ANÁLISE E MINERAÇÃO DE DADOS E INTERAÇÕES DE USUÁRIOS NA WEB.

Monografia de Graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para a obtenção do título de Bacharel em Ciência da Computação

LAVRAS

MINAS GERAIS – BRASIL

2009

EDER BRUNO FONSECA

TRACK4MINE: UMA NOVA PLATAFORMA INTELIGENTE DE COLETA, ANÁLISE E MINERAÇÃO DE DADOS E INTERAÇÕES DE USUÁRIOS NA WEB.

Monografia de Graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para a obtenção do título de Bacharel em Ciência da Computação.

Aprovada em 27 de Novembro de 2009.

Msc. André Grützmann

Prof. Dr. Luiz Henrique Andrade Correa

Prof. Dr. André Luiz Zambalde (Co-Orientador)

Prof. Dr. Ahmed Ali Abdalla Esmin (Orientador)

LAVRAS

MINAS GERAIS – BRASIL

2009

TRACK4MINE: UMA NOVA PLATAFORMA INTELIGENTE DE COLETA, ANÁLISE E MINERAÇÃO DE DADOS E INTERAÇÕES DE USUÁRIOS NA WEB.

Resumo

Esse trabalho propõe uma ferramenta de análise e extração de conhecimento através de coleta dos dados de interações de usuários na Web. Essa ferramenta é capaz de monitorar e exibir informações de interações de usuários em *web sites*. Foram implementadas formas de visualização dos dados coletados, componentes gráficos e funcionalidades. Também foram implementados algoritmos de mineração de dados que permitissem a extração de padrões que auxiliem na tomada de decisão e na inteligência de negócio para administradores de *web sites*.

Palavras Chave: Coleta de Dados; Análise de Dados; Mineração de Dados; Coleta de Interações; Web.

Abstract

This work proposes a tool for analysis and extraction of knowledge by gathering data on user interactions on the Web. This tool is able to monitor and display information from user interactions on web sites. It has been implemented a dashboard of viewing the collected data, graphical components and features. Also, data mining algorithms has implemented that would allow the extraction of patterns to help web sites administrators in decision making and business intelligence.

Keywords: Data collection, Data Analysis, Data Mining, Interactions Collection, Web.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

1.1. CONTEXTUALIZAÇÃO E MOTIVAÇÃO

Cada vez mais a internet se afirma como um grande meio de comunicação, capaz de atingir massas, alterar hábitos e mudar conceitos e atitudes. As empresas têm percebido essa mudança no sentido da inovação e entendem-na como uma oportunidade de expor seus produtos e serviços em um meio de acesso de utilização crescente, com baixo custo e alto poder de influência (Vitor, 2008).

Ainda de acordo com Vitor (2008), através da coleta das formas de interação dos usuários com o serviço ou *site*, é possível entender e aplicar melhorias com o intuito de satisfazer melhor as necessidades dos usuários. Neste contexto, entende-se que interação é um termo geral utilizado para classificar eventos específicos emitidos por usuários em qualquer aplicação ou *site* da *web*, sob o ponto de vista da plataforma *Track4Web*.

Cada vez mais as empresas que implantam *web analytics*, os honestos vendedores, e os especialistas da indústria estão começando a admitir que trabalhar com *web analytics* direito requer muito mais que colocar as *tags* nas páginas e gerar relatórios. Em nossa indústria nasceu um princípio de o orientador que passou a nos orientar: "não é a ferramenta, mas sim como você a usa", e este é, para muitos, uma mudança dramática (Peterson et al., 2009).

Existem ferramentas de monitoramento de *web sites* disponíveis para serem utilizadas, porém, tais ferramentas apenas disponibilizam informações estatísticas, não possibilitando com isso que possam ser feitas mais inferências a partir de estudos e mineração de dados. A base de dados de tais serviços não está

acessível. Com isso, faz-se necessário criar um ambiente que realize a coleta, disponibilize uma interface de consulta e possibilite o acesso direto aos dados com o objetivo de permitir a criação de melhorias que venham a ser úteis para o administrador do *web site* (Vitor, 2008).

No capítulo 2 será apresentada a revisão bibliográfica utilizada para o embasamento teórico do presente trabalho.

1.2. OBJETIVOS

O principal objetivo deste trabalho foi aperfeiçoar a ferramenta Track4Web através do melhoramento da coleta e criação de mais funcionalidades e componentes de visualização, administração de usuário e cadastro independente e melhoramento de interface. Além disso, foram aplicadas técnicas de mineração de dados através da implementação de algoritmos específicos em análise dos dados coletados, dando a esses dados uma interpretação que ajude na inteligência de negócio.

2. REFERENCIAL TEÓRICO

2.1. MINERAÇÃO DE DADOS

Mineração de Dados extrai informações implícitas, anteriormente desconhecidas e potencialmente úteis de bases de dados. Estes conhecimentos e informações descobertos são usados por várias aplicações, incluindo análise de *marketing*, suporte à decisão, detecção de falhas e gerenciamento de negócios (Chen et al., 2003).

Segundo Rygielski et al. (2002), Mineração de Dados (*Data Mining*) é definida como uma sofisticada busca capaz de usar algoritmos estatísticos para descobrir padrões e correlações em dados. *Data Mining* descobre padrões e relações ocultas entre os dados, e é parte de um processo maior chamado *Knowledge Data Discovery* (Descoberta de Conhecimento) que descreve os passos para garantir resultados significativos.

Na visão de Fayyad & Stolorz (1997), *Knowledge Data Discovery* (KDD) se refere ao processo completo de descoberta de conhecimento relevante de dados, enquanto mineração de dados se refere a um passo particular desse processo. Mineração de Dados é a aplicação de algoritmos específicos para extração de padrões de dados, é apenas um passo complementar no processo de KDD, assim como seleção, preparação e limpeza dos dados, priorização apropriada do conhecimento e interpretação dos resultados de mineração, sendo essencial para a construção do conhecimento derivado dos dados.

Para ilustrar o processo de KDD, a Figura 1, adaptada de Han &Kamber (2006) mostra os passos que devem ser seguidos. A seguir são citados os principais passos do processo de KDD:

- Limpeza e Integração dos Dados: são realizadas operações básicas como remoção de ruídos, decisão sobre estratégias para tratamento de campos de dados perdidos e levantamento dos tipos de dados, esquemas e mapeamento de dados perdidos e desconhecidos
- Seleção e Transformação: descoberta de características úteis para representar os dados, dependendo do objetivo da tarefa; e uso de redução dimensional ou métodos de transformação para reduzir o número efetivo de variáveis debaixo de considerações.
- Mineração de Dados: pesquisa por padrões interessantes em uma forma de representação particular ou um conjunto de tais representações.
- Avaliação e Apresentação dos Padrões: visualização dos padrões extraídos, remoção de padrões redundantes ou irrelevantes e tradução dos padrões úteis em termos entendíveis pelos usuários.

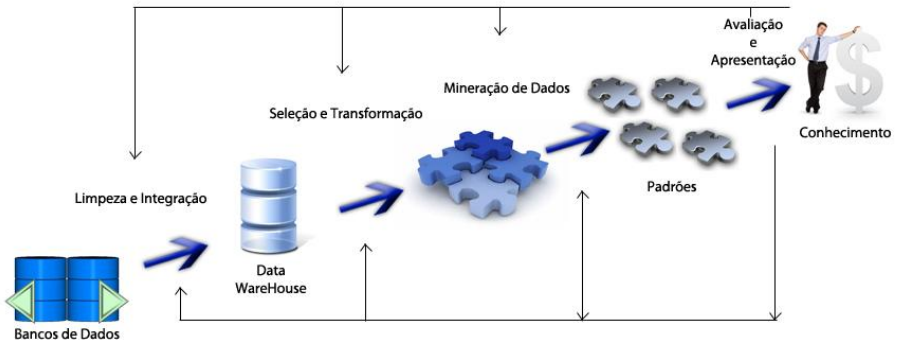

Figura 1 - Processo de *Knowledge Data Discovery* (KDD)

2.1.1. TÉCNICAS DE MINERAÇÃO DE DADOS

De acordo com Dias (2001), as técnicas de mineração de dados podem ser aplicadas a tarefas como classificação, estimativa, associação, segmentação e sumarização. Essas tarefas são descritas a seguir:

- **Classificação**

  A tarefa de classificação consiste em construir um modelo de algum tipo que possa ser aplicado a dados não classificados visando categorizá-los em classes. Um objeto é examinado e classificado de acordo com uma classe definida (Harrison, 1998 apud Dias, 2001).

  São exemplos de tarefas de classificação (Goebel e Gruenwald, 1999 apud Dias, 2001), (Mehta et al, 1996 apud Dias, 2001): classificar pedidos de créditos como de baixo, médio e alto risco; esclarecer pedidos de seguros fraudulentos; identificar a forma de tratamento na qual um paciente está mais propício a responder, baseando-se em classes

de pacientes que respondem bem a determinado tipo de tratamento médico.

- **Estimativa (ou Regressão)**

  A estimativa é usada para definir um valor para alguma variável contínua desconhecida como, por exemplo, receita, altura ou saldo de cartão de crédito (Harrison, 1998 apud Dias, 2001). Ela lida com resultados contínuos, enquanto que a classificação lida com resultados discretos. Ela pode ser usada para executar uma tarefa de classificação, convencionando-se que diferentes faixas (intervalos) de valores contínuos correspondem a diferentes classes(Dias, 2001).

  Como exemplos de tarefas de estimativa tem-se (Fayyad, 1996 apud Dias, 2001), (Harrison, 1998 apud Dias, 2001): estimar o número de filhos em uma família; estimar a renda total de uma família; estimar o valor em tempo de vida de um cliente; estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de um conjunto de diagnósticos médicos; prever a demanda de um consumidor para um novo produto.

- **Associação**

  A tarefa de associação consiste em determinar quais itens tendem a co-ocorrerem (serem adquiridos juntos) em uma mesma transação. O exemplo clássico é determinar quais produtos costumam ser colocados juntos em um carrinho de supermercado, daí o termo 'análise de *market basket*'. As cadeias de varejo usam associação para planejar a disposição dos produtos nas prateleiras das lojas ou em um catálogo, de modo que os itens geralmente adquiridos na mesma compra sejam vistos próximos entre si (Harrison, 1998 apud Dias, 2001).

- **Segmentação**

  A segmentação é um processo de partição de uma população heterogênea em vários subgrupos ou clusters mais homogêneos (Harrison, 1998 apud Dias, 2001). Na segmentação, não há classes predefinidas, os registros são agrupados de acordo com a semelhança, o que a diferencia da tarefa de classificação.

  Exemplos de segmentação: agrupar os clientes por região do país, agrupar clientes com comportamento de compra similar (Goebel e Gruenwald, 1999 apud Dias, 2001); agrupar seções de usuários Web para prever comportamento futuro de usuário (Mobasher et al, 2000 apud Dias, 2001).

- **Sumarização**

  Segundo Fayyad (1996) apud Dias (2001), a tarefa de sumarização envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. Um simples exemplo desta tarefa poderia ser tabular o significado e desvios padrão para todos os itens de dados. Métodos mais sofisticados envolvem a derivação de regras de sumarização.

A Tabela 1 mostra de uma forma resumida as técnicas de mineração de dados descritos acima.

TAREFA	DESCRIÇÃO	EXEMPLOS
Classificação	Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes	• Classificar pedidos de crédito • Esclarecer pedidos de seguros fraudulentos • Identificar a melhor forma de tratamento de um paciente
Estimativa (ou Regressão)	Usada para definir um valor para alguma variável contínua desconhecida	• Estimar o número de filhos ou a renda total de uma família • Estimar o valor em tempo de vida de um cliente • Estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de diagnósticos médicos • Prever a demanda de um consumidor para um novo produto
Associação	Usada para determinar quais itens tendem a co-ocorrerem (serem adquiridos juntos) em uma mesma transação	• Determinar quais os produtos costumam ser colocados juntos em um carrinho de supermercado
Segmentação (ou *Clustering*)	Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos	• Agrupar clientes por região do país • Agrupar clientes com comportamento de compra similar • Agrupar seções de usuários Web para prever comportamento futuro de usuário
Sumarização	Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados	• Tabular o significado e desvios padrão para todos os itens de dados • Derivar regras de síntese

Tabela 1 - Técnicas de Mineração de Dados

De acordo com Dias (2001), a escolha de uma técnica de mineração de dados a ser aplicada não é uma tarefa fácil. Segundo Harrison (1998) apud Dias, 2001, a escolha das técnicas de mineração de dados dependerá da tarefa específica a ser executada e dos dados disponíveis para análise. Harrison (1998) apud Dias (2001) sugere que a seleção das técnicas de mineração de dados deve ser dividida em dois passos: 1) traduzir o problema de negócio a ser resolvido em séries de tarefas de mineração de dados; 2) compreender a natureza dos dados disponíveis em termos de conteúdo e tipos de campos de dados e estrutura das relações entre os registros.

2.2. MINERAÇÃO DE DADOS NA WEB

Com relação à classificação de *Web Mining*, COOLEY et al. (1997) apud Onda (2006) inicialmente apresenta uma taxonomia composta de mineração de conteúdo (*Web Content Mining*) e mineração de utilização (*Web Usage Mining*). Zaïane (1999) apud Onda (2006) apresentou uma taxonomia mais completa, onde foi incluída a mineração de estrutura (*Web Structure Mining*).

Park et al. (2008) definem os elementos da taxonomia de mineração de dados na Web:

- *Web Content Mining*: mineração de textos, imagens, áudios, vídeos, metadados e *hyperlinks* em ordem para extrair regras e conceitos mais usados e sumarizar o conteúdo da *web*.
- *Web Structure Mining*: mineração de estruturas ocultas na *Web* para definir em catagorias as páginas da *Web*, mede a similaridade e a relação entre diferentes *web sites*.
- *Web Usage Mining*: mineração de dados geralmente de interações de usuários com a *web*, incluindo *logs* de acesso ao servidor *Web*, requisições de usuários e *clicks* de mouse para extrair padrões e tendências dos usuários.

Como a ferramenta Track4Web coleta dados de interações de usuários na *web*, foi realizado o estudo sobre alguns algoritmos de mineração de dados que visassem a *Web Mining*. Com o objetivo de estudar a viabilidade da aplicação foram então usados os algoritmos Apriori e AprioriAll, que serão descritos nas próximas sessão.

2.2.1. MINERAÇÃO POR REGRAS DE ASSOCIAÇÃO

A mineração por regras de associação é utilizada para encontrar padrões freqüentes, isto é, associações e correlações entre conjuntos de itens. No ambiente *web*, as regras de associação são normalmente usadas para verificar as correlações entre páginas acessadas durante uma sessão. Estas regras indicam a possibilidade de relação entre páginas, mesmo se elas não estiverem diretamente conectadas(Onda, 2006).

Segundo Onda (2006), esta noção de padrão freqüente depende de duas medidas: o suporte e a confiança. A medida de suporte expressa o número mínimo de sessões para formar uma regra de associação, que são do tipo "Se-Então". Mineração de regras de associação é a descoberta de todas as regras associações provindas de um suporte e confiança mínimos pré-estabelecidos (Choa et al., 2002). Como a forma geral de uma regra é X => Y ("se X então Y"), onde X e Y são conjuntos de páginas web, um suporte de **s** para esta regra de associação significa que X e Y estão contidos em **s** das sessões. A confiança é calculada como a razão entre o suporte da regra X => Y e o número de sessões que contém X. Esta medida expressa a probabilidade que o conjunto de páginas Y seja visitado quando o conjunto X for visitado (Onda, 2006).

De acordo com Choa et al. (2002), o algoritmo Aprori se divide em duas fases. Na primeira fase, todos os *intemsets* que atendem ao suporte mínimo (frequência de *intemsets*) são gerados. Nessa fase, se um *itemset* de tamanho k é um *itemset* frequente, então todos os *intemsets* de tamanho (k-1) também será um *intemset* frequente. Do mesmo modo, se um *itemset* de tamanho k não for um *itemset* frequente, todos os *itemsets* de tamanho (k+1) não serão freqüentes.

Na segunda fase, o algoritmo gera regras de a partir de todos os *itemsets* frequentes.

Como exemplo, na Figura 2 tem-se uma tabela de regras ao qual se deseja descobrir padrões através da aplicação do algoritmo Apriori. O suporte é dado pela razão entre o número de ocorrências de um determinado elemento pelo número total de registros. Sendo assim, pode-se calcular o suporte de A de 66%, como sendo o número de ocorrências do *intemset* A dividido pelo número total de registros. A confiança do *itemset* {A,C} é dada pelo número de ocorrência de {A,C} divido pelo número de ocorrência do *itemset* A.

ID	Regras
1	A,B,C,D,E
2	A,B,E
3	A,C
4	A,C,D,B
5	E,C
6	E,A
7	C,D,A
8	D,C,E,D
9	E,D

Elementos	Suporte
A	**0.66**
B	0.33
C	**0.66**
D	**0.55**
E	**0.66**

Conjunto	Suporte
A, B	0.33
A, C	**0.44**
A, D	0.33
A, E	0.33
C, D	**0.44**
C, E	0.33
D, E	0.33

Figura 2 - O Algoritmo Apriori

Estabelecendo um suporte mínimo de 40% e uma confiança de 50%, realizando os cálculos e aplicando o algoritmo, pode-se eliminar algumas regras. Ao final. pode-se dizer que com 44% de suporte e 66% de confiança, os usuários que acessaram a página A também acessam a página C.

2.2.2. MINERAÇÃO POR PADRÕES SEQUENCIAIS

São muitas as possibilidades de extração da informação, e a mineração de dados através de padrões seqüenciais é um dos métodos mais importantes

(Chen et al., 2003). Han & Kamber (2006) definem mineração de dados por padrão sequencial como a mineração de eventos ou subsequências ordenadas que ocorrem frenquentemente.

O problema de mineração de dados sequenciais e o algoritmo AprioriAll foram introduzidos por Agrawal & Srikant (1995). De acordo com Onda (2006), a mineração de padrões seqüenciais, que é uma extensão da mineração de regras de associação, permite buscar por co-ocorrências incorporando a noção de tempo (ordem dos eventos), isto é, cliques. Desta forma podem-se descobrir quais páginas foram acessadas após um determinado conjunto de páginas. Por exemplo: 23% dos usuários do site visitam a página A, depois a página B e então a página C.

Padrões sequenciais podem ser aplicados à análise de acessos *web*, previsão de tempo, processos de produção e detecção de intrusões de rede (Han & Kamber, 2006).

Ainda de acordo com Tanasa (2005) apud Onda (2006), devido a esta restrição de ordem, esta medida de suporte geralmente é menor que o suporte das regras de associação correspondentes (com as mesmas páginas). Em *Web* Usage Mining, dependendo da quantidade de páginas do web site e do número de visitantes, os valores de suporte dos padrões seqüenciais podem chegar a 0,1%.

ID	Regras
1	A,B,C,D,E
2	A,B,E
3	A,C
4	A,C,D,B
5	E,C
6	E,A
7	C,D,A
8	D,C,E,D
9	E,D

Elementos	Suporte
A	**0.66**
B	0.33
C	**0.66**
D	**0.55**
E	**0.66**

Conjunto	Suporte
A, B	0.33
A, C	0.33
A, D	0.22
A, E	0.22
C, D	**0.44**
C, E	0.22
D, E	0.22

Figura 3 - O Algoritmo AprioriAll

Estabelecendo um suporte mínimo de 40% e uma confiança de 50%, realizando os cálculos e aplicando o algoritmo AprioriAll, pode-ser dizer que: tem-se 44% de suporte e 66% de confiança de quem acessou C, logo em seguida também irá acessar D.

2.3. A PLATAFORMA TRACK4WEB

A plataforma Track4Web foi desenvolvida como trabalho de conclusão de curso (Vitor, 2008). Esta plataforma visa a coleta de dados de sites através de registros de acesso. É uma ferramenta fracamente acoplada, ou seja, o mecanismo de coleta pode ser usado em qualquer site não sendo necessária a adaptação por parte do site analisado.

Segundo Vitor(2008), a plataforma foi desenvolvida seguindo o padrão MVC(*Model - View - Control*) de desenvolvimento, garantindo uma maior independência dos dados e da classificação e processamento (Figura 4). As tecnologias escolhidas para desenvolvimento foram o PHP, a linguagem no cliente *Java Script* e a base de dados *MySQL*. Todas as tecnologias mencionadas foram escolhidas devido ao custo de desenvolvimento, visto que todas são open source. O Mecanismo de Coleta é o componente mais importante da plataforma,

sendo o responsável por interpretar interações do usuário com qualquer *web site* e disparar eventos, que serão interpretados pela camada de controle que reside no servidor.

Figura 4 - Arquitetura MVC(*Model-View-Control*)

O Mecanismo de Análise é responsável por resgatar os dados da base de dados e disponibilizá-los de forma a permitir a análise das informações geradas com foco na tomada de decisão por parte do administrador do *web site*. Neste mecanismo não existem apenas as camadas de negócio, como acontecia no Mecanismo de Coleta, também foi criado um *dashboard* interativo para facilitar a visualização dos dados por parte do administrador como mostrado na Figura 5 (Vitor, 2008).

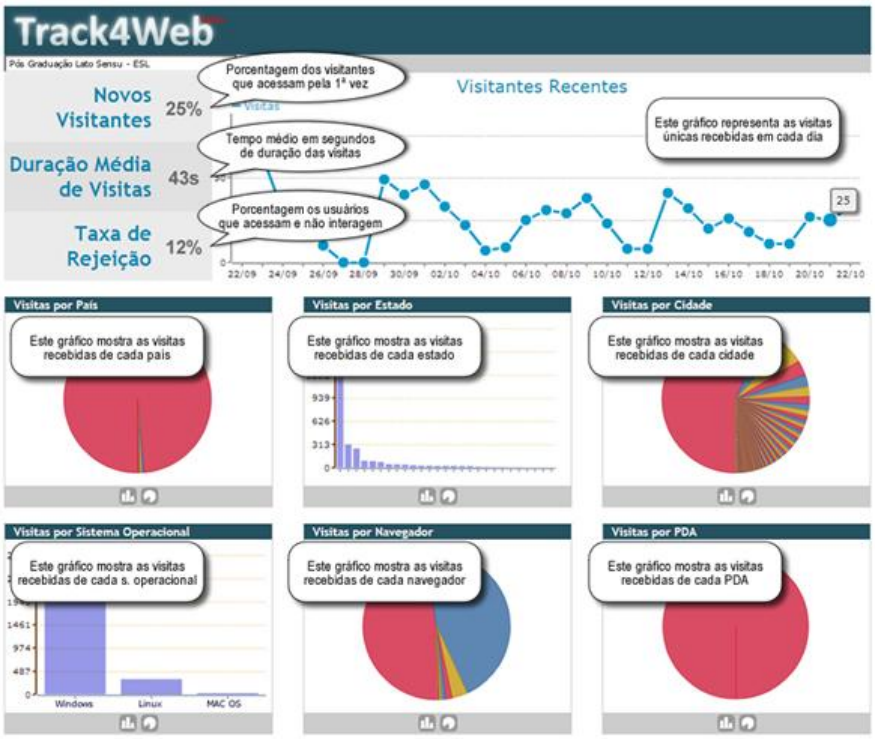

Figura 5 - Dashboard Interativo

A Figura **6**6 exibe o diagrama do mecanismo de coleta da plataforma Track4Web.

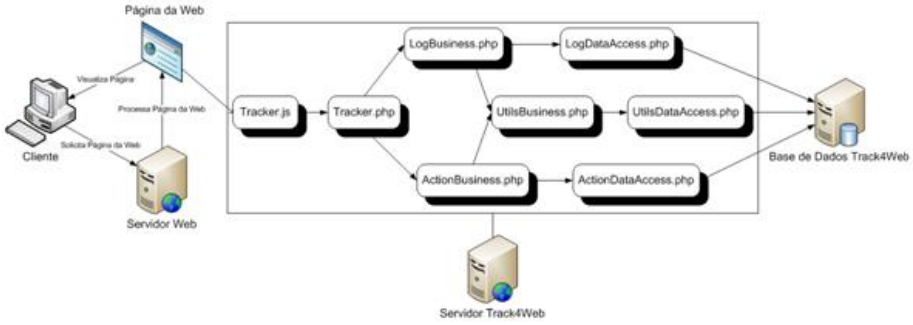

Figura 6 - Mecanismo de Coleta da Plataforma Track4Web

A Figura **7** 7 exibe o diagrama do Mecanismo de Análise desenvolvido para a plataforma.

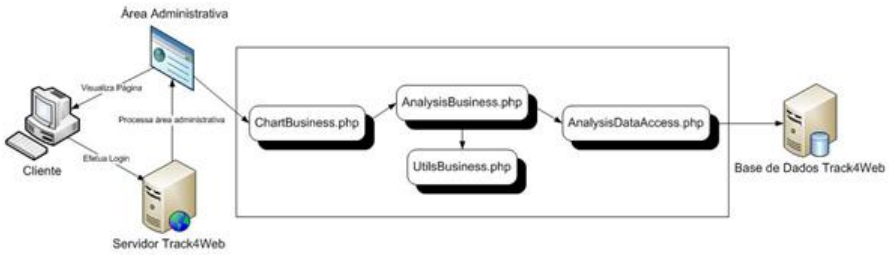

Figura 7 - Mecanismo de Análise da Plataforma Track4Web

3. PESQUISA, MATERIAIS E MÉTODOS

3.1. NATUREZA DA PESQUISA

O presente trabalho tem como principal objetivo desenvolver novas funcionalidades e implementar estudos de mineração de dados para transformar a ferramenta Track4Web em uma ferramenta mais completa, a Track4Mine. Portanto esta pesquisa é do tipo tecnológica, uma vez que se utiliza de técnicas existentes em engenharia de software para o desenvolvimento de uma ferramenta que auxiliará na tomada de decisão e na gerência e inteligência de negócios na área de análise de interação de usuários na *Web*, como por exemplo, na área de *e-commerce*.

A pesquisa pode ser classificada quanto a natureza como sendo pesquisa aplicada; quanto aos objetivos como sendo exploratória; quanto a forma de abordar o problema como pesquisa qualitativa; e quanto aos procedimentos técnicos caracteriza-se com base de pesquisa bibliográfica e documental, estudo de caso e modelagem e simulação.

O procedimento utilizado para desenvolver o presente trabalho foi o desenvolvimento experimental de um protótipo. Além disso, utilizou-se a pesquisa bibliográfica, como por exemplo, consultas a livros, artigos científicos e outros trabalhos referentes ao tema.

3.2. MATERIAIS

Esse trabalho foi desenvolvido no Laboratório de Inteligência Computacional e Sistemas Avançados (LICESA) situado no Departamento de Ciência da Computação (DCC) na Universidade Federal de Lavras (UFLA).

A plataforma de desenvolvimento do Track4Mine foi um Computador Desktop dotado da tecnologia: Processador *Intel® Core™* 2 *Duo* com 2,4 Ghz, 4 Gb de memória RAM DDR2.

O servidor hospedeiro da plataforma é um servidor virtual em uma máquina *Dell PowerEdge* 2900.

As ferramentas necessárias para o funcionamento da plataforma são:

- Servidor *Web* (preferencialmente *Apache*)
- PHP versão 5
- *MySQL versão 5.5*

A principal ferramenta de desenvolvimento foi o *Zend Development Environment* (*Integrated Development Environment*) *Zend Studio 5.5 Trial Version*, que acopla em si um ótimo *debugger* PHP. A ferramenta foi utilizada principalmente para o desenvolvimento do código das camadas de negócio. Também foi utilizado o *Adobe Dreamweaver CS3 Trial Version* como ferramenta gráfica de desenvolvimento (*Integrated Development Environment*) com um maior apelo na parte gráfica, sendo utilizado no desenvolvimento da Interface de consulta.

3.3. MÉTODOS

Este trabalho foi desenvolvido seguindo os seguintes procedimentos:

3.3.1. ESTUDO DA FERRAMENTA TRACK4WEB

Procurou-se fazer um estudo aprofundado da ferramenta. Analisou-se toda a estrutura de coleta e armazenamento, juntamente com a estrutura do banco de dados.

3.3.2. ESTUDO DE MINERAÇÃO DE DADOS

Realizou-se estudos em algoritmos de mineração de dados, analisando os que obtivessem informações relevantes e retornassem melhores resultados.

3.3.3. DESENVOLVIMENTO DA PLATAFORMA

O desenvolvimento da plataforma iniciou-se após uma rigorosa pesquisa bibliográfica e que pudesse oferecer conceitos e informações suficientes para identificar pontos a serem focalizados tanto na exibição dos gráficos quanto na aplicação de algoritmos de mineração de dados.

3.3.4. APLICAÇÃO DE TESTE DA PLATAFORMA

Como estudo de caso, a plataforma está sendo submetida a testes, coletando dados dos sites dos seguintes cursos:

- **Administração de Sistemas de Informação**
  Disponível em *http://www.nte.ufla.br/asi*

- **Engenharia de Software com Ênfase em Software Livre**
  Disponível em *http://www.nte.ufla.br/esl*

- **Informática em Educação**
  Disponível em *http://www.nte.ufla.br/ied*

- **Tecnologia de Redes de Computadores**
  Disponível em *http://www.nte.ufla.br/rde*

E aplicando mineração de dados, com o intuito de descobrir padrões e informações relevantes sobre os cursos de *lato sensu* a distância em que foram coletados os dados pela Track4Web . Além disso, oferece mais formas de visualização dos dados.

4. RESULTADOS E DISCUSSÃO

Esse capítulo tem como objetivo apresentar as mudanças realizadas na plataforma de coleta Track4Web, juntamente com a explicação dos algorítmos de mineração de dados aplicados.

4.1. BANCO DE DADOS

Um passo importante para o desenvolvimento de novas funcionalidades e implementação de novas rotinas foi a compreensão da base de dados. Foi feita então, uma modelagem passando para o modelo entidade-relacionamento e sentiu-se a necessidade de fazer melhorias nesse modelo e na integração dos dados. O primeiro passo, foi a modelagem usando a ferramenta *Mysql WorkBench*. Após a modelagem, obteve-se a base de dados da Figura **8**8, com integridade referencial entre as tabelas do banco.

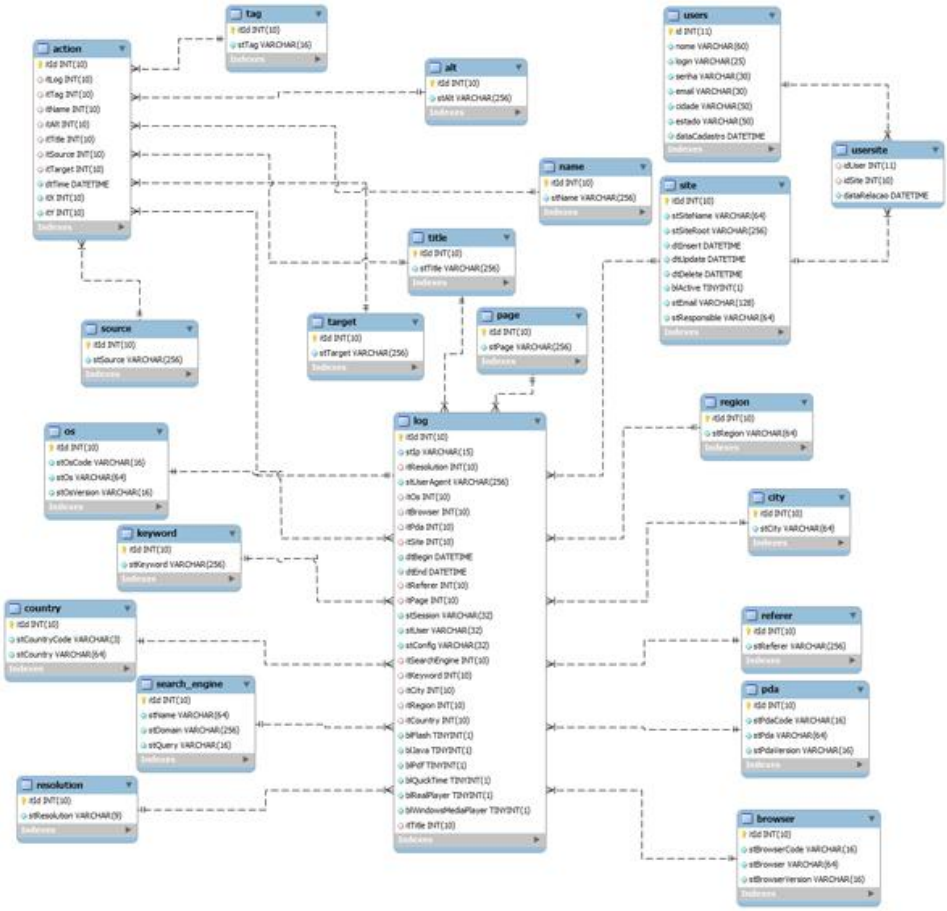

Figura 8 - Modelo Entidade-Relacionamento do Banco de Dados de Track4Mine

A garantia de integridade referencial pode reduzir bastante o armazenamento de dados redundantes, em um sistema projetado segundo os princípios teóricos da abordagem relacional, a redundância não deve existir. Além de no caso de um site que não será mais monitorado, a integridade referencial garantirá que, na exclusão do site, também serão excluídos os dados de log referentes a ele. Dessa forma, evita-se inconsistências.

4.2. INTERFACE

Na interface do Track4Mine foram modificados antigos gráficos existentes no Track4Web visando o aperfeiçoamento gráfico de exibição de dados. Além disso, a maioria dos gráficos foram rearranjados e redistribuídos para uma melhor visualização. Agrupou-se por categorias, como gráficos que indicam localidade, buscadores, conteúdo.

Na Figura 9 pode-se ver as mudanças de interface, com a criação de novos gráficos no formato de pizza, barra, linha e tabelas de dados. E a reorganização dos gráficos que já existia no Track4Web.

Figura 9 - Novos gráficos e Reorganização dos Já Existentes

Abaixo são citadas algumas das principais funcionalidades implementadas na plataforma Track4Mine:

- **Visualização de acessos agrupados por hora do dia**.

  Na Figura **10**10, agrupa-se os acessos por hora. Dessa forma, fica mas fácil perceber quais os horários críticos de acesso. O gráfico "Visitas por Hora" representa a freqüência de acessos totais agrupados por hora, e o gráfico "Visitas por Hora Hoje" representa a freqüência de acessos desde às 0(zero) hora do dia corrente agrupados por hora. Excelente para a tomada de decisão da equipe de suporte, que tem condições de usar horários de menos visitas para colocar o site ou o servidor em manutenção.

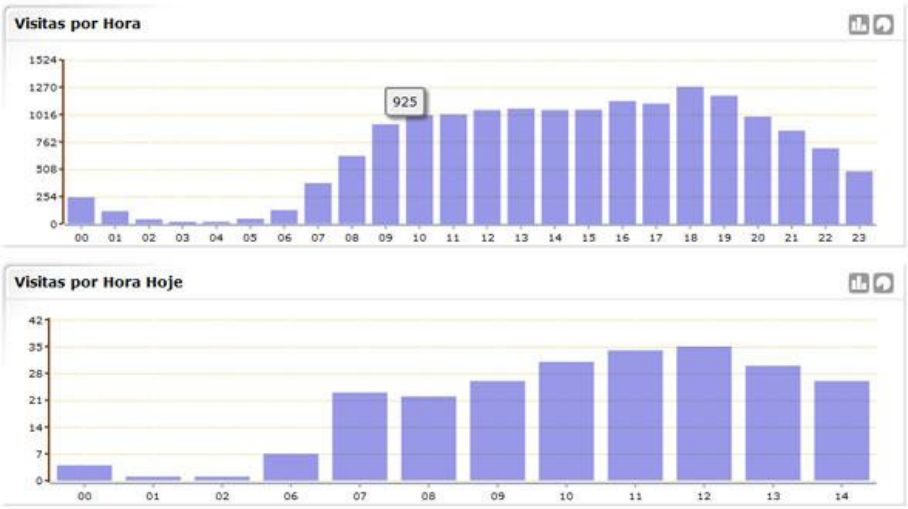

Figura 10 - Acessos por Hora

- **Alteração de perfil de usuário.**

  Criou-se uma interface de alteração de perfil pelo usuário, como pode ser visto na Figura 11. Mantendo a estrutura MVC, podemos ver na Figura 12 como foi montado a arquitetura de controle do usuário.

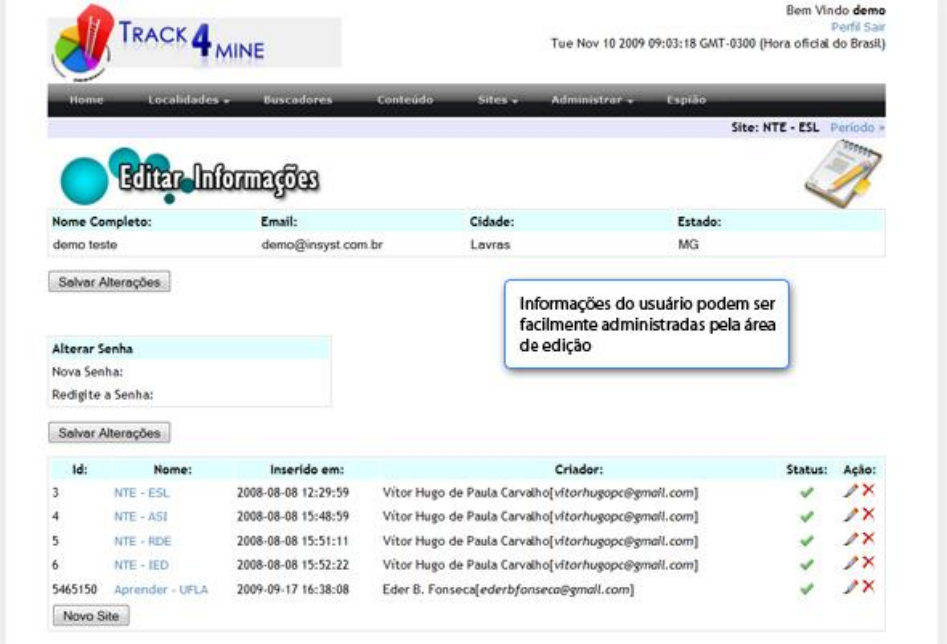

Figura 11 - Página de Perfil do Usuário

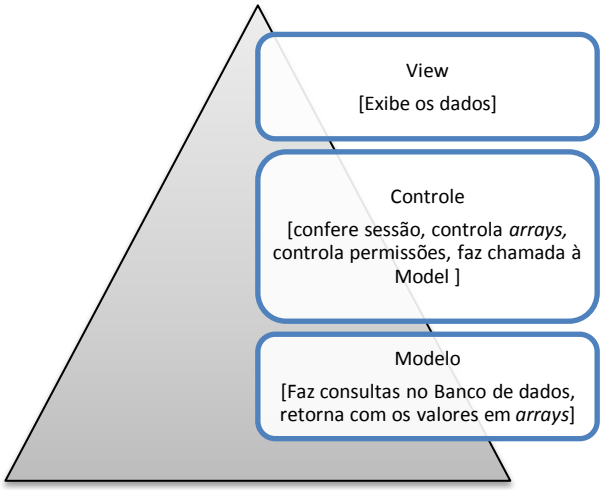

Figura 12 - Estrutura de MVC para controle do Usuário

- **Monitor de interações online em tempo real.**

  Um nova funcionalidade que merece destaque implementada no Track4Mine é o monitor de interações online em tempo real. Esse monitor consiste em uma página em AJAX que é atualizada em intervalos de tempo de 5 segundos. Dessa forma, ela permite a visualização em tempo real das navegações no site monitorado. Nessa página fornece informações como cidade, estado e país do visitante, juntamente com a hora de acesso à página, a página que originou a navegação, versão do navegador e sistema operacional e IP usado. Na Figura 13 temos uma representação do monitor.

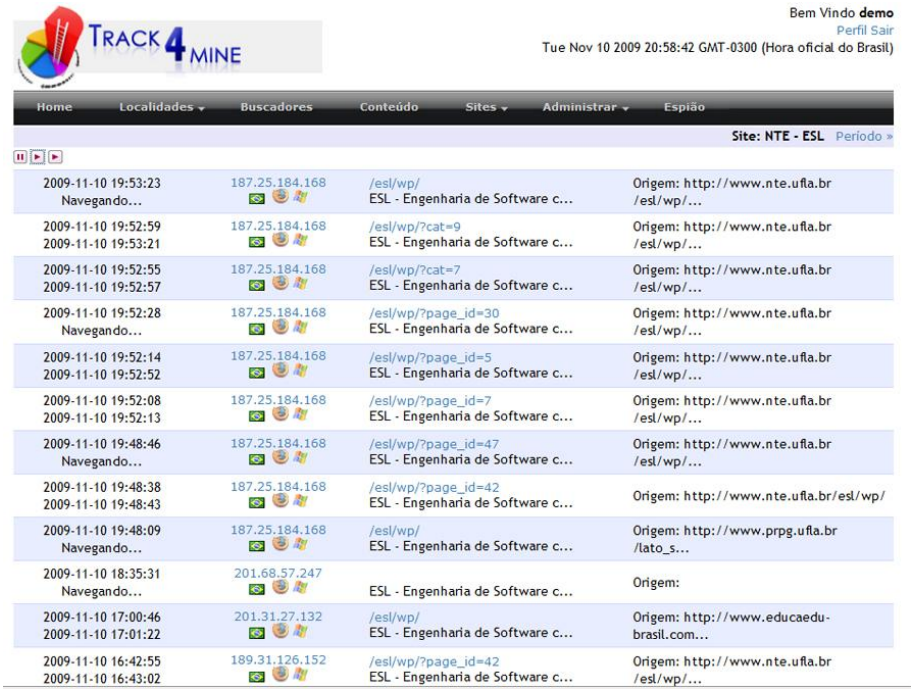

Figura 13 - Visualizador de Acessos em Tempo Real

Com essas mudanças de interface, preocupou-se em deixar a ferramenta fácil de ser usada e mais intuitiva, visto que tem interface e gráficos semelhantes à ferramentas já existentes no mercado, como por exemplo, *Google Analytics*.

4.3. MINERAÇÃO DE DADOS

O mecanismo coleta de dados através de registro de acesso implementado no Track4Web coleta dados de 4 sites(ESL, ASI, IED e RDE) e obteve mais de 270.000 registros no período de setembro de 2008 à outubro de 2009. Essa enorme quantidade de dados ultrapassava a capacidade humana de compreensão. Decidiu-se, então, aplicar algoritmos de mineração de dados para descoberta de informação implícita nessa grande base.

4.3.1. APRIORI

O primeiro passo foi a seleção e preparação dos dados. Foram comparadas as sessões que acessaram os sites. Obteve-se então uma tabela como a exibida na Tabela 3, onde o campo "stSession" representa a sessão do usuário, os demais campos representam os IDs dos sites vistados e em cada célula foi adicionado o valor 'sim' quando o usuário visitou o site referenciado pela coluna, e o valor 'não' quando o usuário não acessou o site referenciado pela coluna. A Tabela 2 define a representação numérica de cada site monitorado para o melhor entendimento da Tabela 3.

Id	Nome do Site
3	NTE – ESL
4	NTE – ASI
5	NTE – RDE

6	NTE - IED

Tabela 2 - Representação de IDs por Site

itSession	3	4	5	6
000f8be6b3136538e51c0060c49570a9	nao	nao	sim	nao
001477d67367a164507879d5e10f593a	nao	nao	nao	sim
0018d88eeaa792f53b7d859a39d236f8	nao	sim	nao	nao
002653c1c0ee699120cdb91e33f4157b	nao	nao	sim	nao
003a0763327c73a545457e0afd2c3da6	sim	nao	nao	nao
0049dbb9980a0f495f5f18ba4b63e72f	nao	nao	nao	sim
004d1d0e73005521717cdccd6edd2a44	nao	nao	sim	nao
004d6170cb3d3215de312981213b6a5d	nao	sim	nao	nao
0063d6849c42f83e59efc395c6981115	nao	nao	sim	nao
007eea8c97b6273fab29dec4fffb7b1d	nao	sim	nao	nao
0081a93224bf633c15b22019eb320939	nao	nao	sim	nao
00823d99eb3fd0f4931e1e5ab84ab2e6	nao	sim	nao	nao
0093f97860b7747f1afdb6c8e7ff7a9a	nao	sim	nao	nao
0096dc4dcf7ce9bddd015e3c558b7a4c	nao	nao	nao	sim
009d45a4ae5261462a5fd63e2c63e6b3	sim	sim	nao	nao
00ac30cc762d7ef1c5efbecf5a1b512c	nao	nao	nao	sim

Tabela 3 - Relação de Sessões e acessos do Site.

No segundo passo, através da plataforma Weka, aplicou-se o algoritmo Apriori para descobrir uma relação entre os 4 sites de pós-graduação monitorados pelo Track4Web. Obtendo os resultados representados na Tabela 4.

```
=== Run information ===
Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -
M 0.1 -S -1.0 -c -1
Relation:     QueryResult-weka.filters.
unsupervised.attribute.Remove-R2-3
Instances:    9303
Attributes:   5
              itSession
              ESL
              ASI
              RDE
              IED
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.25 (2326 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:
Size of set of large itemsets L(1): 7
Size of set of large itemsets L(2): 15
Size of set of large itemsets L(3): 7
Size of set of large itemsets L(4): 1
Best rules found:
 1. ESL=nao ASI=nao RDE=nao 2492 ==> IED=sim 2492    conf:(1)
 2. ASI=nao RDE=nao IED=sim 2510 ==> ESL=nao 2492    conf:(0.99)
 3. RDE=nao IED=sim 2563 ==> ESL=nao 2535    conf:(0.99)
```

4. ESL=sim ASI=nao 2580 ==> IED=nao 2551 conf:(0.99)
5. ASI=nao IED=sim 2555 ==> ESL=nao 2526 conf:(0.99)
6. ESL=nao ASI=nao IED=sim 2526 ==> RDE=nao 2492 conf:(0.99)
7. ESL=nao RDE=nao IED=sim 2535 ==> ASI=nao 2492 conf:(0.98)
8. ASI=nao RDE=sim 2642 ==> IED=nao 2597 conf:(0.98)
9. ASI=nao IED=sim 2555 ==> RDE=nao 2510 conf:(0.98)
10. ESL=sim 2779 ==> IED=nao 2723 conf:(0.98)

Tabela 4 - Resultado Apriori

O resultado mais importante analisado na Tabela 4 seria a regra 1, em que pode-se descrevê-la como: "Com 100% de confiança, os usuários que visitam o site IED, não visitam os sites ASI, ESL e RDE". Através desse resultado, pode-se dizer com confiança que os usuários de curso IED tem perfil diferente dos usuários dos demais cursos.

4.3.2. APRIORIALL

A implementação do algoritmo AprioriAll exigiu também uma escolha e limpeza rigorosa dos dados a serem analisados. O desafio foi descobrir um padrão de navegação de usuários, descobrir um caminho mais freqüente de navegação através das páginas do site. O site ESL foi o escolhido para a aplicação do algoritmo.

Usando a sessão como caracterização de visita única de um usuário, implementou-se uma rotina que retorna uma tabela 'visitas' que contém dois campos: um campo para a sessão do usuário, e outro campo para as páginas visitadas por este usuário, em ordem cronológica(Tabela 5). Cada número no campo 'stVisitas' representa o Id da página(Tabela 6).

stSession	stVisitas
29c5ceca9e825c8657c75a98fada39c0	25
bdd1bd37c6ddcc735db4acea101ebdb0	40,35
66d0b3eec18a341d6ad601127fe7772c	40,33,36,35
6cb538990ab7a2fa5175c3f12f721b9b	25,40
9eaaffd62d8abac7ebbdcaa9bdd7cbd6	25,40
c798ba1e15bf83b1d1562e671fbb69db	40
147e160876f177bcb1c0ab5659be1213	40,25
3d8aff9842551e206ed0ebbcdadc8e2e	25
368b01e8b45575d24927493b2d5fd5ed	25,40
8499c22df15cfa402d1c83db98233128	25
626ef96b42f55e9b41f98f3e6e6fa9b5	40
5d0acaccd0c048726eaf2665025fb108	25,42,40,25,36,35
d5b4aab0ccbeee8cc2f8d54bdbdb87d9	25,33,40
0296edc359db2a6c8f8f44baf28729dc	25
1edd93ffe927e73736652f819a9117ea	25
9c393731f06c5e1066bc79a1be893c57	25
9d91a7cac3b78d8f93f9042679e097b6	25
2cf43d443729b51c65e2cbd9c5a06187	25
fc14acf3683121dda3123e15704fa814	40,35,25,33
324e175e10c52c2812c7ba41e7081829	40,33,25,40,25,33,36,35,42,34,25,40,40,72
125170026a18ccf0aeb900b2108cfd5b	40,33,36,34,35,33,40
b536a4592f84ff195317fd1e80b55da0	40,72
c02f5bba0dff7790fcaa29283116f8e2	40,25,35,25,25,25,35,25,35
c09ea6fbdc683d9e098a21deabfda046	40,35,86,33,40,42,35
cad4f8fd964a896b60d849ab944eb148	25,34,42,36,33
c6218d1af250a6a58a82e61c31d0100d	40,33,36

Tabela 5 - Visitas em Ordem Cronológica

itId	stPage
0	
25	/esl/wp/
26	/asi/wp/
27	/rde/wp/
28	/rde/wp/?page_id=33
29	/rde/wp/?page_id=37
30	/rde/wp/?page_id=4
31	/ied/wp/?p=34
32	/ied/wp/?page_id=3
33	/esl/wp/?page_id=42
34	/esl/wp/?page_id=7
35	/esl/wp/?page_id=5
36	/esl/wp/?page_id=47
37	/ied/wp/
38	/rde/wp/?page_id=32
39	/rde/wp/?cat=4
40	/esl/wp/?page_id=9
41	/rde/wp/?page_id=35
42	/esl/wp/?page_id=43
43	/ied/wp/?page_id=28
44	/ied/wp/?page_id=4

Tabela 6 - Páginas e referências numéricas (ID).

Optou-se por não usar da mineração de dados por regras de associação pelo fato de a mesma não levar em consideração a ordem cronológica de visitas nas páginas, o que caracterizaria uma desvantagem no caso do Track4Mine. Por

exemplo, uma sessão em que a página A fosse visitada antes de uma página B seria contada como a sessão que visitou a página B e em depois visitou A.

Desse modo, a segunda parte foi a implementação do algoritmo de mineração de dados por padrões seqüenciais. Analisou-se todas os padrões de visitas e foram encontrados alguns padrões(Figura 14).

Páginas Visitadas	Suporte	Confiança
-> ESL - Enge... => ESL - Enge...	20.35%	68.36%
-> ESL - Enge... => ESL - Enge...	20.99%	70.52%
-> ESL - Enge... => ESL - Enge...	19.71%	70.46%
-> ESL - Enge... => ESL - Enge...	20.61%	73.67%
-> ESL - Enge... => ESL - Enge...	18.86%	67.43%
-> ESL - Enge... => ESL - Enge...	13.86%	81.08%
-> ESL - Enge... => ESL - Enge...	14.11%	82.58%
-> ESL - Enge... => ESL - Enge...	13.65%	79.88%
-> ESL - Enge... => ESL - Enge...	11.47%	67.12%
-> ESL - Enge... => ESL - Enge...	10.34%	85.74%
-> ESL - Enge... => ESL - Enge...	10.57%	87.66%
-> ESL - Enge... => ESL - Enge...	10.57%	87.66%

Figura 14 - Padrões Sequenciais

A Figura 14 representa padrões seqüenciais de navegação, que podem ser entendidos de maneira a analisar o comportamento de navegação de usuários. A interface retorna em links as páginas acessadas. Logo, clicando nesses links, tem-se acesso direto á página referencia.

Com a implementação do algoritmo AprioriAll pode-se verificar se os caminhos e sub-caminhos de navegação projetados pelo webmaster estão sendo realmente seguidos, ou se usuários estão descrevendo percursos diferente do esperado ao navegar num site.

5. CONCLUSÕES

5.1. CONCLUSÕES FINAIS

O objetivo desse trabalho foi melhorar a ferramenta Track4Web, transformando-a na Track4Mine, uma ferramenta composta de vários outros gráficos, formas de visualização e inteligência de negócios aplicada à sua base de dados.

Aplicando técnicas de mineração de dados nos dados coletados pela ferramenta, pôde-se conhecer padrões implícitos em sua enorme base. Com essas aplicações, pode-se então nortear administradores, desenvolvedores e suporte em atividades que vão desde um parada do site para manutenção, auxiliando na escolha de um melhor horário cujo número de acessos é menor; até a análise de comportamento para direcionar marketing, ofertas, entre outros.

Com as melhorias feitas na interface gráfica, pode-ser ter uma interface mais intuitiva e cognitiva. A ferramenta Track4Mine tem aspectos semelhantes à ferramentas onlines já existentes, como por exemplo *Google Analytics*.

O monitor em tempo real é um componente importante que pode ter aplicação efetiva em áreas de *helpdesk*, visto que fornece um acompanhamento em tempo real da navegação dos usuários no site. Assim, ficam perceptíveis as situações em que tem-se usuários com dificuldades de navegação e necessitam de ajuda. Entre outras informações, tais como, a localidade de origem e as páginas visitadas.

A aplicação do algoritmo Apriori, para descoberta de relação entre os sites monitorados, retornou informações que podem auxiliar no marketing mais

direcionado, visto que a descoberta de sites que não possuem relação implica a relação de interesse dos usuários de um site em outro. Para a aplicação nos cursos à distância, pode-se ver que não fazia sentido aplicar capital financeiro em marketing do curso IED nos demais cursos, pois os alunos desse cursos não possuem interesse nos demais.

O algoritmo AprioriAll para descoberta de padrões seqüenciais retorna informações importantes como, por exemplo, seqüências de páginas visitadas, caminhos relevantes no web site, etc. Essas informações podem auxiliar para descoberta quebras de navegação no site, página isoladas e/ou sem acesso, ou percursos de navegação não desejados.

Pode-se dizer que o objetivo principal foi alcançado, uma vez que muitas informações relevantes foram extraídas de sua enorme base de dados; e foram implementados outros gráficos, ampliando as maneiras de visualização e interpretação dos dados.

5.2. TRABALHOS FUTUROS

O desenvolvimento da ferramenta Track4Mine é uma continuidade de um trabalho já existente, continua sendo implementada e deverá gerar novas versões.

Como trabalhos futuros, pode-se implementar ainda mais funcionalidades na ferramenta, como por exemplo:

- Através da interação com o site monitorado poderá ser feito ajuste automático para atender o perfil do usuário que está navegando naquele exato momento;

- Implementar algoritmos de *clustering* para identificar o perfil do usuário em termo de banda, horário e localidade;
- Implementar algoritmos de classificação para classificar usuários de acordo com seu interesse por um determinado produto ou área no site;
- Realização de estudos e pesquisas visando a aplicação da Track4Mine como ferramenta de apoio a geração de marketing automático direcionado em sites e-commerce.

6. REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R.; SRIKANT, R.. **Mining Sequential Patterns**. International Conference on Data Engineering (ICDE). 1995.

CHEN, Y.; CHIANG, M.; KO, M.. **Discovering time-interval sequential patterns in sequence databases.** Expert Systems with Applications. 2003.

CHOA, Y. H.; KIMB, J. K.; KIMA, S. H.. **Personalized Recommender System Based on Web Usage Mining and Decision Tree Induction**. Expert Systems with Applications. 2002.

DIAS, M. M.. **Um Modelo De Formalização Do Processo De Desenvolvimento De Sistemas De Descoberta De Conhecimento Em Banco De Dados.** Tese de Doutorado. Universidade Federal de Santa Catarina. 2001.

FAYYAD, U.; STOLORZ, P.; **Data mining and KDD: Promise and challenges**. Future Generation Computer Systems. 1997.

HAN, J.; KAMBER, M.. **Data Mining, Concepts and Techniques. Morgan Kaufmann Publishers.** 2ª Edição. 124-129. 2006.

ONDA, M.. **Metodologia de Mineração de Dados para Análise do Comportamento de Navegar num Web Site.** Universidade Federal do Rio de Janeiro - Rio de Janeiro. 2006.

PARK, S.; SURESH, N. C.; JEONG, B.. **Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm.** Data & Knowledge Engineering. 2008.

PETERSON, E.; CARNEIRO, R.; FIGUEIREDO, F.; RIBEIRO, G.; TOKUNO, D.; FATALA, A.; SIMÕES, C.; VARNUM, J.; GIUNTINI, M.; MORIER, D.; HORA, C.; LOUREIRO, G.; FOLLI, A.; NARESSI, L.; TSUGI, V.; DORNELES, M.; GONÇALVES, P.; VALE, R. F.. **Web Analytics - Uma Versão Brasileira II**. Creative Commons. 2009.

RYGIELSKI, C.; WANG J.; YEN, D. C.. **Data mining techniques for customer relationship management**. Technology in Society. 2002.

Vitor Hugo de Paula Carvalho. **TRACK4WEB: Uma Plataforma Inteligente de Coleta de dados e Interações de Usuários na WEB**. 2008. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade Federal de Lavras. Orientador: Ahmed Ali Abdalla Esmin.