Estudo comparativo dos métodos *Wum* e *Web*sift para mineração de uso da *web* aplicado na recomendação da informação

Caio Olschowsky Borges
Universidade de Cruz Alta, UNICRUZ
E-mail: <caiogauderio@hotmail.com>

Patricia Mariotto Mozzaquatro Chicon
Universidade de Cruz Alta, UNICRUZ
E-mail: <patriciamozzaquatro@gmail.com>

Abstract

Currently, there are a large amount of data available on *web*. The extend number of data can influence in the flow of activities of developers and users, causing a overload of information. Such, this paper presents the problematic how work with large quantities of data and then extract only necessary, namely, get the personalize information according the interest of each user. In this context, the research to be developed addresses the utilization of computing methods, the mining and the *web* filtering data. The data mining can be conceptualized as the discovery and intelligent analysis of useful information of *web*. The proposed paper will integrate the *web* usage mining (usage miner) where will implemented the methods *web* usage mining (WUM) and *web* site information filter system (*web*sift) applied on extract and recommendation of information on *web*, supporting on the search process in the *web* environment.

1 Introdução

Atualmente existe uma vasta quantia de dados disponíveis na *web*, uma árvore de rede interligando cada vez mais usuários portadores de *notebooks*, *tablets*, *smartphones*, dentre outros (OLIVEIRA; SILVA, 2009). As empresas estão investindo em servidores, pela grande quantidade de informações por minuto sendo armazenadas nestes, até em nuvem. A *web* tem-se tornado um mercado de informações e continua crescendo fazendo com que as empresas aumentem os números de servidores para armazenamento e melhor organização dos dados dos usuários (COOLEY, 1999). Como solução, o trabalho proposto apresenta a utilização de técnicas computacionais, a mineração e filtragem de dados *web*. Elas tem por finalidade extrair e filtrar dados a fim de gerar conhecimento. A mineração de dados pode ser conceituada como a descoberta e análise inteligente de informações úteis da *web*. Pode-se estar interessado, por exemplo, na informação contida dentro dos documentos da *Web* (mineração de conteúdo); na informação contida entre os documentos da *Web* (mineração de estrutura) ou na informação contida na utilização ou interação com a *Web* (mineração de uso). Essas são as três categorias em que se divide a mineração na *Web* (COOLEY, 1999). A pesquisa integrou a mineração de uso da *web* (*usage miner*) onde foram implementados os algoritmos *Web Usage Mining* (WUM) e *Web* Site *Information Filter System* (*WEBSIFT*).

A pesquisa contribuiu cientificamente por meio do estudo comparativo entre dois algoritmos integrantes da mineração *web*, será apresen-

DOI: 10.18256/2359-3539/reit-imed.v1n2p3-9

tado o mais eficiente no processo de extração e busca da informação em uma base de dados, assim, outras pesquisas e projetos futuro poderão utilizá-lo.

2 Metodologia

A pesquisa aplicada no presente trabalho classifica-se como quali-quantitativa, ou seja, foram realizadas análises a fim de avaliar aspectos numéricos e qualitativos. A implementação do sistema integrou as seguintes etapas: seleção (nela que serão decididos quais conjuntos de dados que serão relevantes para que sejam obtidos resultados com informações úteis), pré-processamento (acontece a limpeza de dados e seleção dos atributos), transformação (os dados importantes que foram retirados no processo anterior são modificados de forma que a próxima etapa possa ser realizada), data-mining (nesta etapa os dados depois de transformados serão lidos e interpretados). Na etapa do data-mining irá ocorrer a aplicação dos métodos *WUM* e *WebSIFT*. A seguir são descritas as fases trabalhadas para o desenvolvimento da aplicação.

- *Fase 1* – Referencial teórico: Estudo sobre a mineração de dados; Análise e revisão bibliográfica sobre Mineração *web*, bem como a mineração de uso da *web*; Descrever o algoritmo *WUM*: conceito e funcionamento deste e alguns resultados; Estudar e analisar o algoritmo *WEBSIFT*: explicar de forma descritiva também o significado *WebSIFT*, suas fases ou etapas no funcionamento e alguns teste realizados.
- *Fase 2* – Modelagem e Desenvolvimento prático: desenvolver um sistema computacional utilizando uma linguagem de programação para realização de testes finais; Integrar uma base de dados já existente ao sistema desenvolvido, para após utilizar a mesma nos testes finais, a fim de coletar informações; Executar os dois algoritmos no sistema elaborado.
- *Fase 3* – Avaliação: Apos realização de testes com os algoritmos, avaliar o software para uso futuro; desenvolver um estudo comparativo verificando qual das técnicas após sua utilização obteve melhor eficiência na geração dos resultados.

3 Mineração de Uso

A mineração de uso da *web* (RIGO, 2008) ou *web usage miner*, é a atividade de mineração de dados que visa a descoberta automática de padrões de comportamento de usuários, por meio de seus dados de acesso a *web*.

Diversas áreas de aplicação são identificadas na mineração de uso, como por exemplo, aplicações de reconhecimento de perfis de usuários, personalização e recomendação, melhorias e reestruturação do projeto de sites da *web*, avaliação de sites em ambientes de educação a distância, inteligência de negócios e comércio eletrônico (COUTO, 2009). As informações podem ser descobertas em navegadores, servidores e aplicações *web*.

Os dados em geral descobertos em padrões de uso de páginas da *web* como, endereço IP, páginas referenciadas, data e tempo de acesso são os mais utilizados. Os dados de uso de servidores *web*, são permanecidos em arquivos no formato CLF (*Common Log Format*). Informações de aplicações *web*, ficam normalmente no formato XML (SOUZA, 2008). A análise é feita nestes dados ajuda na detecção de uma série de normas, como demográficas, tempo médio de acesso, resultados de campanhas e estratégias de marketing direcionadas a certas mercadorias.

A subseção a seguir ilustra os métodos *Web Usage Mining* e *Web* Site *Information Filter System*.

3.1 Método Web Usage Mining e Web Site Information *Filter System*

O método *Wum* utiliza uma linguagem de mineração MINT, como objetivo especificar critérios de consulta (conteúdo, estatística, estrutura) e uma linguagem de consulta *WebMiner* (SPILIOPOULOU-FAULSTICH, 1997). *Web*Miner é uma ferramenta que fornece uma linguagem de consulta sobre um software de mineração externo para associação de regras e para padrões sequenciais (COUTO DE SOUZA, 2009), (COUTO, 2009). O Sistema de Filtragem de Informações para *Web* ou *Web Site Information Filter System* é um framework de mineração de uso que além de realizar pré-processamento e descoberta de conhecimento, usa a estrutura e o conteúdo da informação de um site, ao fim faz a identificação dos resultados interessantes da mineração de uso.

(RIGO, 2008). Projetado para realizar a extração dos registros do servidor em um formato estendido. O algoritmo de pré-processamento inclui identificação de usuários, sessões de servidor, e apresenta páginas em cache por meio do uso do campo referenciado (WANG, 2000).

Ambos os métodos citados integram as fases de seleção dos dados, pré-processamento, transformação dos dados, mineração dos dados, interpretação e finalizando com o conhecimento.

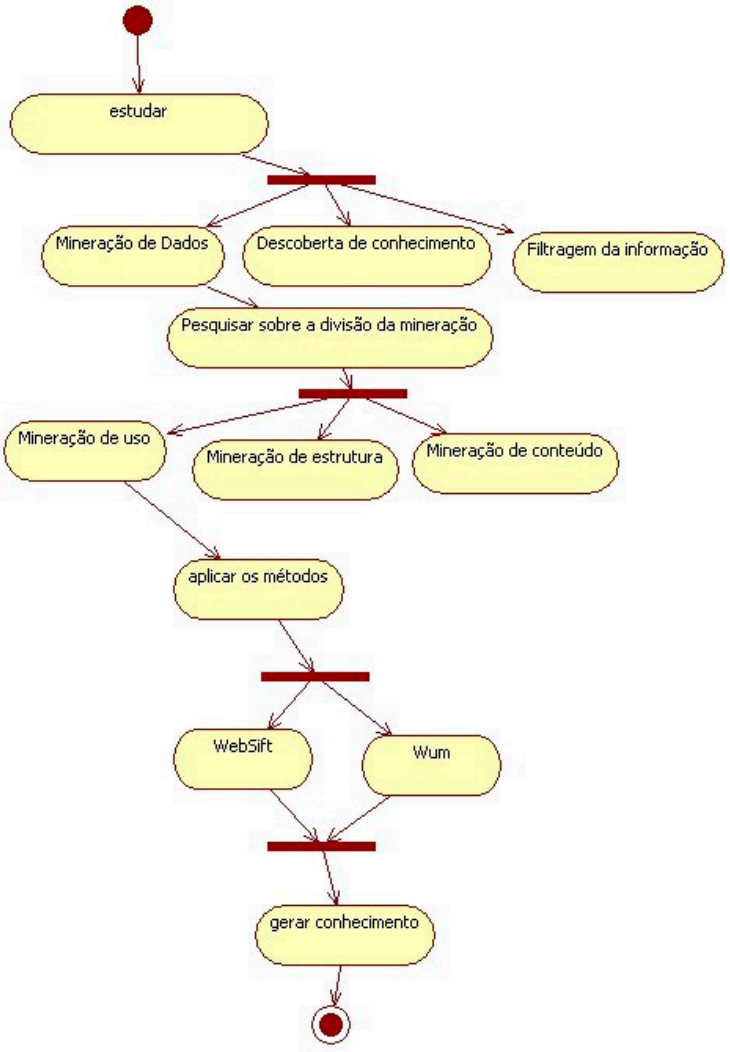
4 Sistema desenvolvido

O trabalho proposto tem por objetivo realizar um estudo comparativo entre dois métodos integrantes da mineração de uso da *web*: *Wum* e *Websift*. As subseções a seguir ilustram as etapas de desenvolvimento do sistema.

4.1 Processo de Modelagem

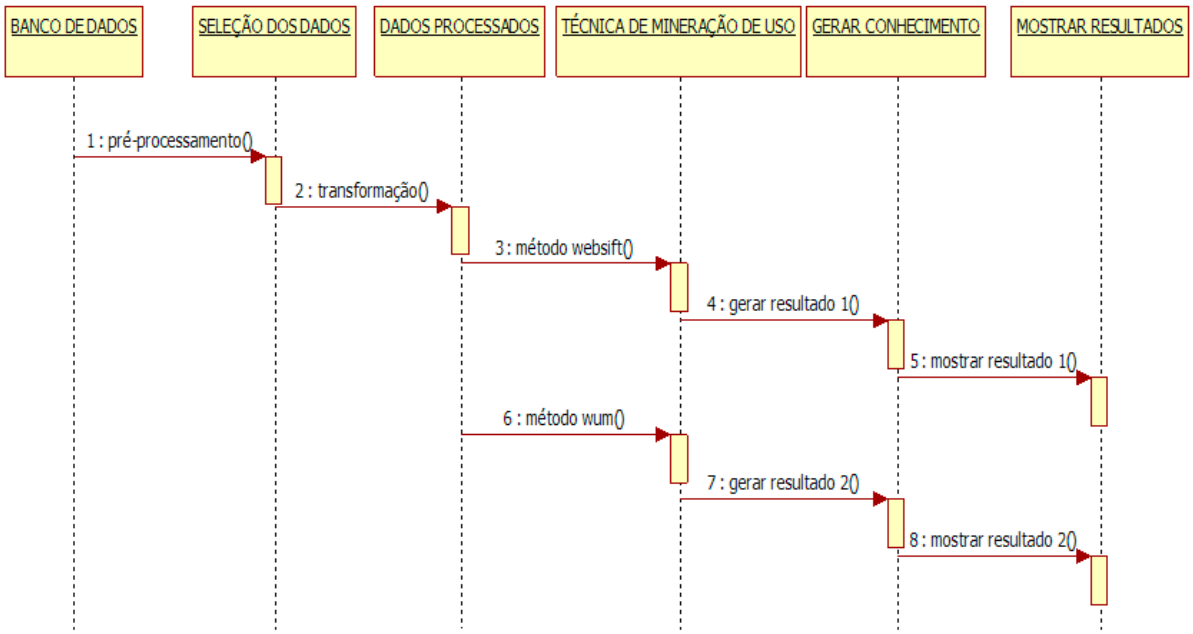
A Figura 1 demonstra a contextualização do trabalho proposto. O trabalho iniciou-se pelo estudo da Mineração de dados, o processo de descoberta de conhecimento em base de dados e a filtragem da informação. Após foi realizada análise da divisão da mineração de dados, ou seja a mineração *web*, sendo analisada a mineração de conteúdo, mineração de estrutura e mineração de uso (técnica utilizada) no presente trabalho. Já feita a escolha da técnica optou-se por implementar os métodos *Wum* e *Websift* integrantes da técnica de mineração de uso. Os métodos citados foram aplicados em uma base de dados a fim de extrair e recomendar informações, auxiliando no processo de busca, gerando assim o conhecimento.

Figura 1 - Contextualização do Estudo



A Figura 2 apresenta o diagrama de sequência ilustrando o processo de busca com a utilização dos métodos citados. Conforme ilustra a Figura 2, no primeiro momento os dados estão armazenados em um banco de dados onde ocorre o pré-processamento. Após, inicia o processo de transformação e os dados são processados. Prosseguindo, é aplicada a técnica de mineração de uso da *web* com seus métodos *Wum* e *Websift* separadamente. Após sua aplicação acontece a etapa da geração do conhecimento, isto é, os dados já foram minerados.

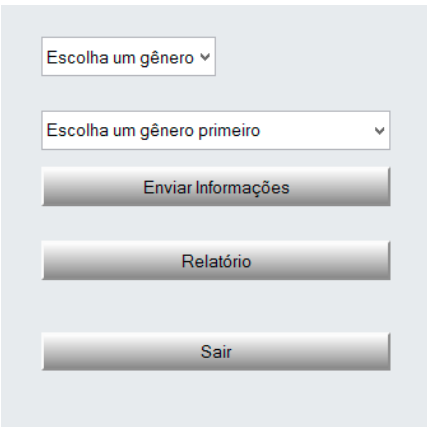
Figura 2 - Diagrama de Sequência



5 Etapas do desenvolvimento

Para o desenvolvimento do sistema foram utilizadas algumas ferramentas, são elas: *software* Eclipse onde foi feita a programação, Paint para edição de imagens, Notepad++ para auxílio na endentação e Xampp para criação do banco de dados do sistema. A Figura 3 ilustra a escolha do gênero musical após o cadastro no sistema. A amostra integrante da pesquisa incluiu 120 registros de usuários.

Figura 3 - Escolha do Gênero Musical



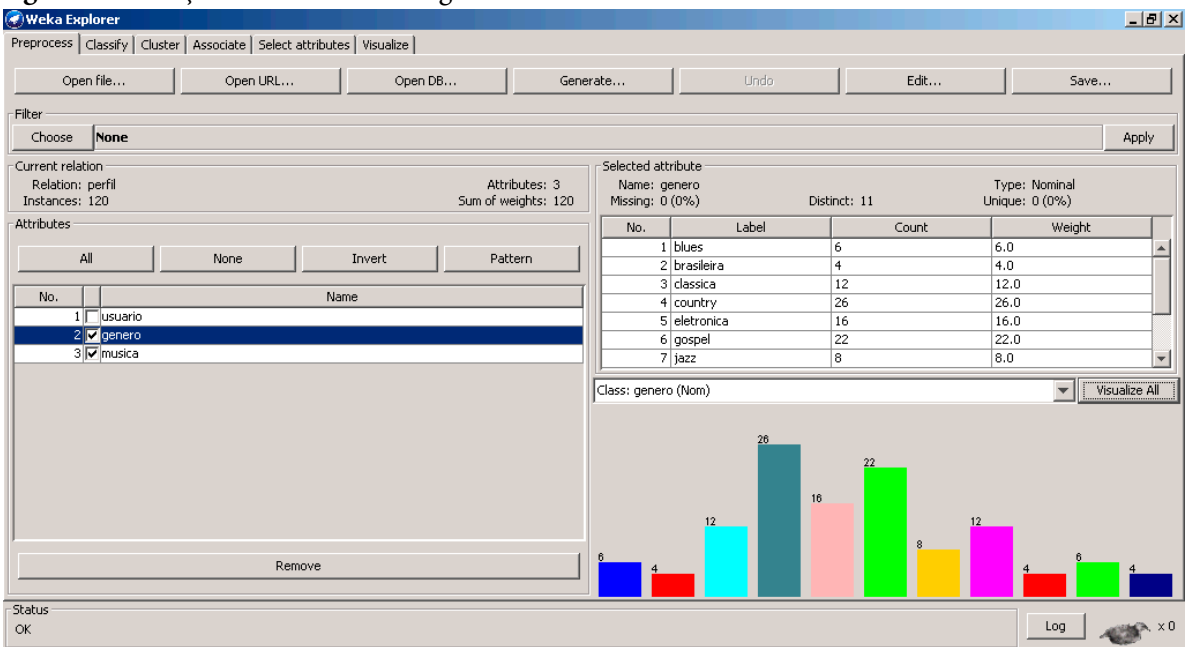
Conforme a Figura 3, a aplicação dos métodos ocorre da seguinte forma: no primeiro momento acontece a descoberta de conhecimento, ou seja, os dados passarão pelos padrões sequenciais, pelo agrupamento de páginas ou clusterização, pelas regras de associação e por fim pelo pacote de padrões de estatística. Em seguida, a técnica de filtragem de informação começa sua tarefa, filtrando os dados para após, serem visualizados e, finalizando, a técnica chama o mecanismo de conhecimento de busca.

No processo de pré-processamento, a entrada de dados é usada para traçar um arquivo de sessão do usuário, para derivar a topologia do site e classificar as páginas do mesmo. Este arquivo de sessão será transformado para o arquivo de transação e saída para próxima fase na qual é chamada de descoberta de padrões. Ambas as topologias dos sites e classificação das páginas são alimentadas dentro da filtragem de informação, que pertence ao processo de análise de padrões a faz uso de conteúdo pré-processado e informação estruturada para filtrar automaticamente os resultados da descoberta de conhecimento de algoritmos para padrões que são potencialmente interessantes. A descoberta de padrões usa os

métodos de mineração de dados *Wum* e *Websift* como mencionado anteriormente gera as regras e padrões. A descoberta de informação é então alimentada dentro de várias ferramentas de análises de padrões, que incluem a filtragem de informação, OLAP, e mecanismo de busca de conhecimento como SQL, para gerar o resultado final da mineração.

Para a geração dos resultados, foi integrada ao sistema a API do software *Weka*. Com os resultados já gerados no formato sql. Foi criada a base de dados no formato .arff, para após aplicação do métodos citados. A utilização da ferramenta justifica-se devido a mesma tornar o sistema portável, apresentar uma linguagem multiplataforma orientada a objetos (JACOMINI, 2008). A Figura 4 mostra a geração de conhecimento dos gêneros mais solicitados. Constatou-se que o gênero musical "country" (26%) obteve maior número de acessos, por segundo, o "gospel" (22%) teve segundo maior número de acessos.

Figura 4 - Geração de conhecimento gêneros mais solicitados



A subseção a seguir apresenta a implementação dos métodos *Wum* e *WebSift*.

5.1. Implementação dos métodos *Wum* e *Websift*

O método *Wum* emprega uma técnica para a descoberta de padrões de navegação por meio de uma visão agregada materializada de log da *web*. Foi aplicado o algoritmo *FilteredClassifier* integrante do método citado. Após a geração da base de dados no formato .arff, em um primeiro momento aplicou-se a regra de *use training set*. Foi selecionado um conjunto de treinamento e dividido em duas partes: cerca de sessenta (60) por cento dos dados utilizados para criar o modelo. Após, para testar a exatidão do algoritmo foi aplicada a regra *supplied test set* com os dados restantes de cerca de 40 por cento, colocando-os em um conjunto de testes e finalmente o algoritmo foi testado com a amostra total.

Na amostra de 60% o grau de exatidão foi de 47,2222% e de erro foi de 52,7778%. Apresentou um erro médio de 0,0049. Trinta e quatro instâncias foram classificadas como corretas e 38 como incorretas, totalizando 72 instâncias. O tempo médio de execução do algoritmo na amostra de *use training set* foi de 0,07 segundos. Na amostra de 40% o índice de acerto foi de 49,0566 % e de erro foi de 50,9434%. Apresentou um erro médio de 0,0047. Vinte e seis instâncias corretas e vinte e sete instâncias incorretas, totalizando 53 instâncias. O tempo médio de execução do algoritmo na amostra de *supplied training set* foi de 0,06 segundos. Finalmente, na amostra de 100% o índice de acertos foi de 41,6667% e de erros foi de 58,3333%. Apresentou um erro médio de 0,0053. Cinquenta instâncias foram classificadas como corretas e setenta como incorretas, totalizando 120 instâncias. O tempo médio de execução do algoritmo na amostra de *cross-validation* foi de 0,05 segundos. Na amostra validada, após a cria-

ção do modelo foi verificada a exatidão do algoritmo de classificação, ou seja, observou-se que o modelo construído inicialmente foi bem próximo ao conjunto de testes, o que indica que o modelo não falhará com dados desconhecidos ou quando dados futuros forem aplicados a ele.

O Método *Websift* é uma espécie de framework de mineração de uso da *web* que, além de realizar o pré-processamento e o conhecimento das descobertas, usa a estrutura e a informação do conteúdo de um site para automaticamente definir um filtro de informações confiável. Foi aplicado o algoritmo *DecisionStump* integrante do método citado.

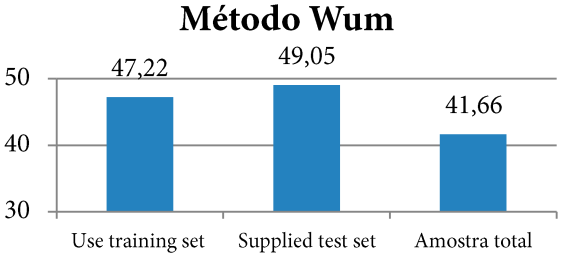
Conforme já mencionado anteriormente, também foram validadas as amostras de treinamento, teste e total. Na amostra de 60% o nível de acerto foi de 16,6667% e de erro foi de 83,3333%. Apresentou um erro médio de 0,0049. Doze instâncias foram classificadas como corretas e quarenta como incorretas, totalizando setenta e duas instâncias. O tempo médio de execução do algoritmo na amostra de *use training set* foi de 0 segundos. na amostra de 40% o nível de acertos foi de 18,8679% e de erros foi de 81,1321%. Apresentou um erro médio de 0,0067. Dez instâncias foram classificadas como corretas e quarenta e três como incorretas, totalizando cinquenta e três instâncias. O tempo médio de execução do algoritmo na amostra de *supplied test set* foi de 0,03 segundos. Finalizando, na amostra de 100% o nível de acerto foi de 18,3333% e de erro foi de 81,6667%. Apresentou um erro médio de 0,0068. Vinte e duas instâncias foram classificadas como corretas e noventa e oito como incorretas, totalizando cento e vinte instâncias. O tempo médio de execução do algoritmo na amostra de *use training set* foi de 0,06 segundos. Assim pode-se comprovar a veracidade do algoritmo.

6 Estudo comparativo entre os métodos *Wum* e *Websift*

Para avaliar o trabalho proposto utilizou-se o teste de software denominado caixa branca, também conhecido como teste estrutural. Baseada na arquitetura interna de software.

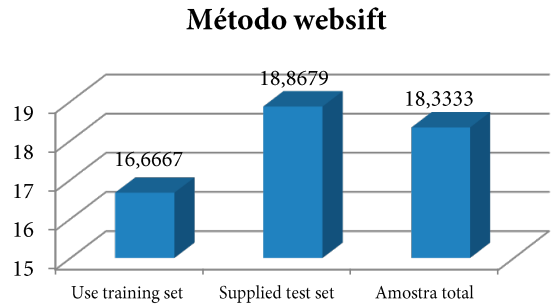
Na Figura 5, é ilustrado o gráfico feito no software *microsoft excel* com as informações dos resultados na ferramenta *Weka* do métodos *Wum*.

Figura 5 - Método *WUM* – etapa de Treinamento - Teste - Total



Conforme ilustra a Figura 5, é apresentado o percentual das amostras e números obtidos por meio de testes realizados na ferramenta de mineração *Weka*. Observa-se que na amostra de treinamento o método apresentou 47,22% de acerto, na amostra de teste 49,05% e na amostra total 41,66%. Os dados obtidos comprovam a veracidade o algoritmo aplicado, pois os valores apresentados estão bem próximos. A Figura 6 ilustra o método *Websift*.

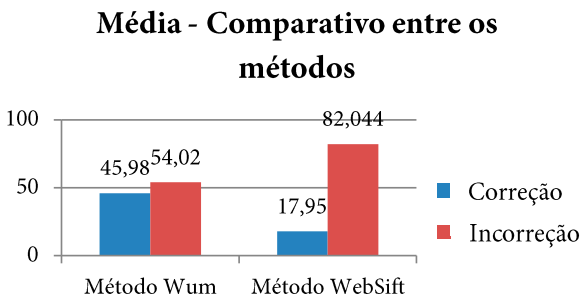
Figura 6 - Método *Web*sift



Conforme ilustra a Figura 6, é apresentado o percentual das amostras, números obtidos por meio de testes realizados na ferramenta de mineração *Weka*. Observa-se que na amostra de treinamento o método *websift* apresentou 16,6667% de acerto, na amostra de teste 18,8679% e na amostra total 18,3333%. Os dados obtidos comprovam a veracidade o algoritmo aplicado, pois os valores apresentados estão bem próximos.

A Figura 7 ilustra o gráfico com os dois métodos juntos, apresentando a média geral do grau de correção e de incorreção.

Figura 7 - Comparativo entre os métodos Wum e *Web*sift



No gráfico exposto na Figura 7 é ilustrado um comparativo de ambos os métodos.

O método *Wum* apresentou um nível de acerto de 45,98%, em contrapartida, o método *web*sift apresentou apenas 17,95% de acerto. A média de acerto foi obtida por meio do somatório das amostras de teste, treinamento e total. Nota-se claramente que os métodos *Wum* apresentou maior percentual de correção em relação ao *Websift*. O método *Wum* apresentou um grau de correção de 54,02%, em contrapartida, o método *websift* apresentou 82,044% de erro. A média de erro foi obtida por meio do somatório das amostras de teste, treinamento e total. Nota-se que o método *Web*sift apresentou maior percentual de erro em relação ao *Wum*. Nota-se que o método *Wum* se tornou melhor, por obter maior percentual de acerto em relação ao método *Web*sift.

7 Considerações Finais

Acredita-se que o estudo desenvolvido tenha alcançado seus objetivos, assim como contribuído para a evolução das pesquisas e estudos sobre métodos de mineração de dados *Wum* e *WebSift*. A principal contribuição deste trabalho foi realizar um estudo comparativo para verificar qual o método mais eficiente para a extração de informação e geração de perfis de usuários finais. Constatou-se na pesquisa, que o método WUM obteve melhor resultado na extração e recomendação de informações na *web*, auxiliando no processo de busca em ambiente *web*.

Para o desenvolvimento do sistema utilizou-se o SGBD *MySql* (Sistema Gerenciador de Banco de Dados) na a linguagem PHP (*Hypertext Preprocessor*), onde foram integrados os dois métodos citados. A aplicação recebeu o nome MGMS (*Musical Genres Mining System*), o mesmo verifica os gêneros musicais e nome das músicas que o usuário acessou. Após, o sistema gera um relatório demonstrando qual gênero e música obteve maior numero de acesso. Assim, com os resultados obtidos foi gerada a base de dados no formato .arff, a qual foi executada na ferramenta *Weka* para a geração e aplicação dos métodos.

O trabalho proposto contribuiu para as pesquisa na área da mineração *web*, mais especificamente na mineração de uso com a aplicação dos métodos *Wum* e *Websift*.

A hipótese inicial foi de que o método WUM teria melhor desempenho na predição do comportamento do usuário, durante sua interação no ambiente do que o algoritmo *Websift*. Com os testes executados pode-se comprovar a teoria com a prática. Este resultado justifica-se devido o algoritmo *WUM* utilizar primitivas de consulta SQL e suportar especificações de critérios de seleção de padrões.

Com o desenvolvimento deste trabalho surgiram possíveis novas investigações a serem feitas: Adaptar este sistema a dispositivos móveis utilizando a técnica de *bootstrap*, como também aplicá-lo em um player musical.

Referências

COOLEY, Robert; PANG-NING, Tan; SRIVASTAVA, Jaideep. *The Web Site Information Filter System*. DCCUM, Minnesota: UMN, 1999.

COUTO DE SOUZA, Luis Carlos. *Metodologia de Mineração de Dados Aplicada a Navegação de Dispositivos Móveis*. Dissertação de Mestrado do Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro COOPE/UFRJ, 2009.

JACOMINI, Diego. *Análise Da Base De Dados Dos Ingressantes*. Na Unidavi em 2008 Com A Ferramenta WEKA. Disponível em: <http://www.unidavi.edu.br/?pagina=FILE&id=56962>. Acesso em: 06 nov. 2012.

OLIVEIRA Camilo, Cássio; DA SILVA, João Carlos. *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*. UFG, Goiás, 2009.

RIGO, Sandro José. *Integração de Recursos da Web Semântica e Mineração de Uso para Personalização de Sites*. Dissertação de Doutorado em Ciência da Computação, Universidade Federal do Rio Grande do Sul – UFRGS, 2008.

SPILIOPOULOU, Myra; C. FAULSTICH, Lukas. *A Web Utilization Miner*. Berlin, 1998.

WANG, Yan. *Web Mining and Knowledge Discovery of Usage Patterns*. Universidade de Waterloo, Waterloo, Canadá, 2000.