Universidade Federal de Ouro Preto - UFOP

Instituto de Ciências Exatas e Biológicas - ICEB

Departamento de Computação - DECOM

DESENVOLVIMENTO DE UM REPOSITÓRIO DE DADOS DO FUTEBOL BRASILEIRO

Aluno: Rafael Belini Souza
Matricula: 07.1.4153

Orientador: Luiz Henrique de Campos Merschmann

Ouro Preto
19 de novembro de 2012

Universidade Federal de Ouro Preto - UFOP

Instituto de Ciências Exatas e Biológicas - ICEB

Departamento de Computação - DECOM

DESENVOLVIMENTO DE UM REPOSITÓRIO DE DADOS DO FUTEBOL BRASILEIRO

Relatório de atividades desenvolvidas apresentado ao curso de Bacharelado em Ciência da Computação, Universidade Federal de Ouro Preto, como requisito parcial para a conclusão da disciplina Monografia I (BCC390).

Aluno: Rafael Belini Souza
Matricula: 07.1.4153

Orientador: Luiz Henrique de Campos Merschmann

Ouro Preto
19 de novembro de 2012

Resumo

Neste trabalho de monografia é apresentada a proposta de desenvolvimento de um repositório de dados do futebol brasileiro, visando deixá-lo acessível em um ambiente web para que futuras pesquisas relacionadas à mineração de dados no esporte possam ser realizadas. A carência de uma fonte dos dados centralizada é uma justificativa para o desenvolvimento deste repositório, possibilitando os usuários ou pesquisadores realizarem suas próprias consultas para adquirir os dados de acordo com suas necessidades. Para que os dados fossem coletados, foram utilizados, como fontes, o CartolaFC [1], website onde se encontram armazenados os atributos particulares de cada jogador, e o site Futpédia [2], para a obtenção dos atributos dos jogos. Os dados coletados serão pré-processados [3] com o objetivo de eliminar inconsistências existentes nas fontes de dados. Visando deixar o sistema com fácil acesso, será projetada e desenvolvida uma interface para que os usuários possam resgatar os dados dos sistema, que será disponibilizado via servidor web.

Palavras-chave: Repositório de dados. Dados do futebol brasileiro. Pré-processamento, Desenvolvimento web.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Mineração de Dados tem atraído uma grande atenção da indústria da informação e da sociedade atual devido à disponibilidade de grandes quantidades de dados e da iminente necessidade de transformar esses dados em informação útil e conhecimento [3]. As informações e conhecimentos adquiridos podem ser usados em aplicações como análise de mercado, detecção de fraudes, fidelização de clientes, controle de produção e outras.

Com o avanço das pesquisas com bases de dados no ambiente comercial e industrial, notou-se, também, uma incrível quantidade de dados existente em diferentes domínios nos esportes. De acordo com [4], estes dados são definidos na forma de performance individual de atletas, treinamento ou decisões administrativas, jogos ou disputas e/ou quão bem um time se comporta com determinados atletas jogando juntos. O maior desafio para este domínio de aplicação não é como coletar os dados, mas ter o conhecimento de quais deverão ser coletados e como fazer o melhor uso deles.

Grandes organizações esportivas podem ser empresas multi-milionárias que arriscam gastar muito dinheiro em uma única decisão. Com esta quantidade de capital em jogo, uma decisão equivocada tem potencial para arruiná-las. No trabalho [4], os autores argumentam que a indústria dos esportes é um ambiente atrativo para aplicações de técnicas de mineração de dados, baseado na gama de riscos enfrentados por ela e pela considerável e constante necessidade de tomar boas decisões.

Ainda há poucos trabalhos relacionados com a aplicação de técnicas de mineração de dados na área de futebol. Em função disso, a proposta da criação de um repositório de dados para o futebol brasileiro surge para suprir a ausência de uma fonte de dados centralizada, auxiliando a recuperação dos devidos dados de maneira mais eficiente e limpa (adquirindo somente os dados que são realmente relevantes para uma determinada aplicação).

2 Justificativa

Atualmente, mineração de dados vem sendo bastante utilizada em um grande número de áreas distintas e com diversos propósitos. Entretanto, em relação à sua utilização no esporte, em especial no futebol, ainda há uma carência significativa referente à pesquisas e aplicações. Com o desenvolvimento de um repositório para armazenar os dados do futebol brasileiro, se proporcionará o acesso direto a dados pré-processados coletados de diferentes fontes, fornecendo agilidade ao acesso às informações e permitindo que cada usuário obtenha uma base de dados específica para suas necessidades. Com isso, será possível facilitar a pesquisa com mineração de dados no futebol pelo fato de que os dados estarão disponíveis em um servidor web, fazendo com que sejam acessados a qualquer momento e de qualquer lugar.

3 Objetivos

3.1 Objetivo geral

O objetivo geral consiste em projetar e desenvolver um repositório de dados, sendo disponibilizado em um servidor web, com o intuito de fornecer os dados do futebol brasileiro para que novas pesquisas relacionadas à mineração de dados no futebol sejam realizadas.

3.2 Objetivos específicos

- Coletar dados de diferentes fontes.
- Realizar o pré-processamento das informações coletadas para alimentar um banco de dados.
- Elaboração do projeto do banco de dados.
- Projeto e implementação de um sistema web para permitir o acesso aos dados por usuários externos.

Os objetivos específicos compõem os passos para que seja possível o desenvolvimento do que foi explicitado nos objetivos gerais. A coleta dos dados é necessária para que consiga criar um repositório unificado, centralizando as fontes e dados. O pré-processamento dos dados é uma fase importante por consistir na limpeza das informações, eliminando os dados irrelevantes, inconsistências e/ou instâncias duplicadas do banco. A etapa de elaboração do projeto é a fase de planejamento do banco de dados, assim como planejamento do sistema web, de forma que este seja acessível e de fácil uso. Por último, a fase de implementação do sistema consiste em criar uma interface web que permita o acesso aos dados contidos no repositório de dados. O objetivo é que os dados possam ser acessados e extraídos de acordo com a necessidade de cada usuário. Desta maneira, este sistema possibilitará pesquisas personalizadas e, com isso, poderá colaborar com futuras pesquisas em mineração de dados no futebol.

4 Metodologia

O repositório de dados tem como objetivo facilitar o trabalho e diminuir o esforço de quem deseja realizar pesquisas voltadas para aplicação de técnicas de mineração de dados na área do futebol.

Para que a etapa de desenvolvimento seja completada, é importante percorrer algumas outras etapas anteriormente. Primeiramente, é de extrema importância que sejam estudadas as características das fontes de dados, observando a estrutura dos arquivos em que os dados são disponibilizados. Após a identificação do padrão desses arquivos, serão gerados algorítmos para extração e para realizar a ação de pré-processamento dos dados coletados, onde estes serão armezanados em um banco de dados relacional de forma apropriada. Para tanto, serão feitos o projeto e implementação do banco de dados adequadamente. Em seguida, é necessário que o banco de dados seja alimentado com os dados pré-processados. A elaboração do projeto do sistema e sua implementação compõem a próxima fase, onde será desenvolvida uma interface web que facilitará o acesso de vários usuários. A última etapa será a de implantação do sistema.

Para demonstrar a eficiência deste repositório, os testes se fazem a partir do uso prático do sistema simulando usuários distintos.

A Figura 1 ilustra a metodologia aplicada neste trabalho.

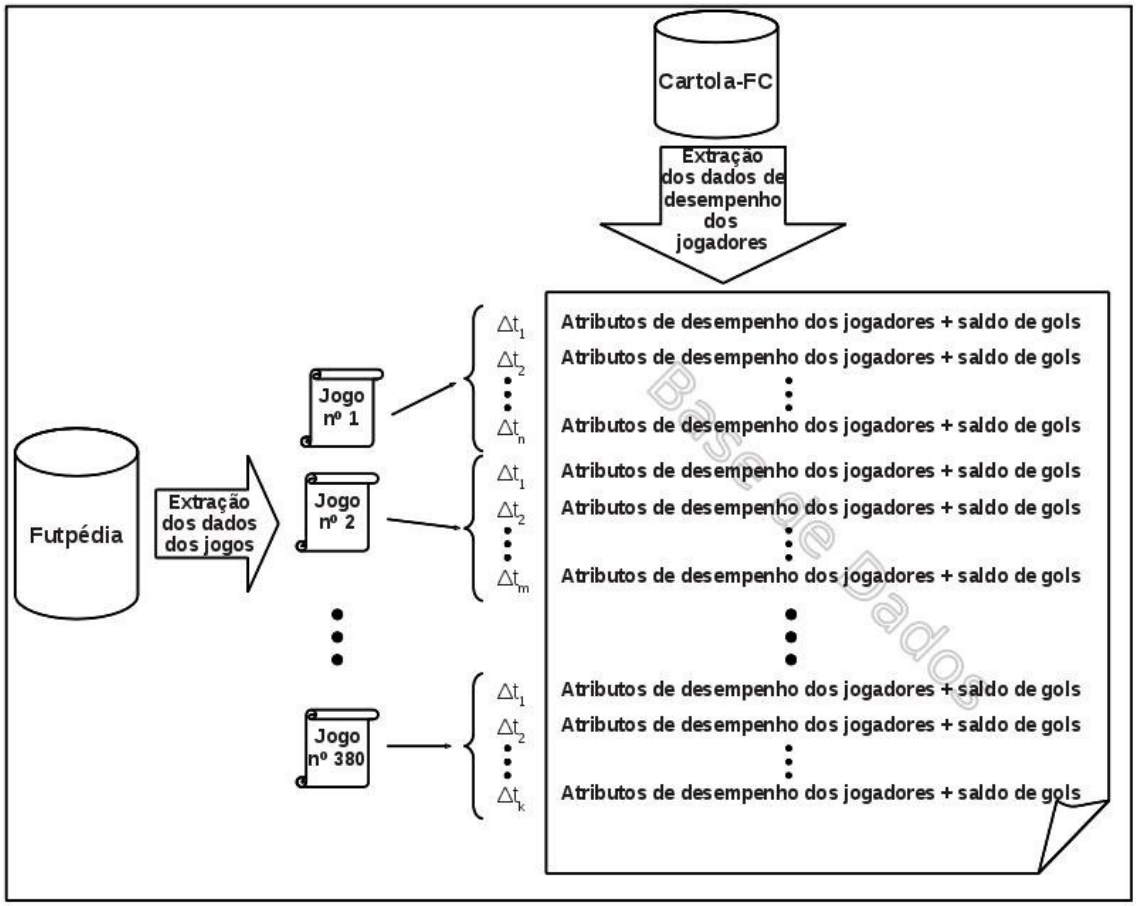

Figura 1: Metodologia

5 Desenvolvimento

Na primeira parte de desenvolvimento, ou seja, na parte de desenvolvimento da disciplina Monografia I, o objetivo era estudar a estrutura das fontes de dados citadas acima e realizar a extração dos dados necessários, aplicar a ação de pré-processamento nos dados recolhidos, estruturar um banco de dados para armazenar as informações e, por fim, gerar uma base de dados generalizada que será utilizada nas aplicações de métodos de mineração de dados e na geração de bases mais específicas e enxutas.

As Figuras 2 e 3 demonstram o formato das bases de dados utilizadas para a extração dos atributos dos jogadores e dos jogos.

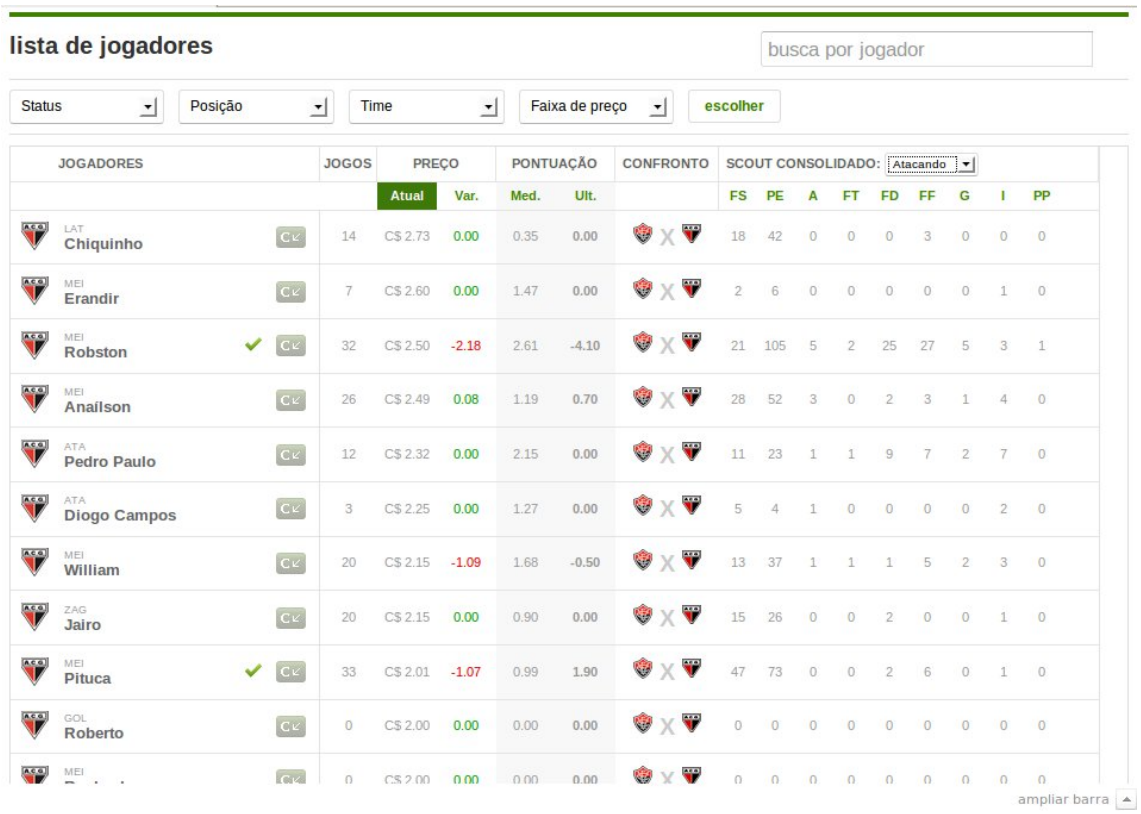

Figura 2: Atributos específicos dos jogadores disponibilizados no website CartolaFC

Primeiramente, estudando as fontes escolhidas, notou-se uma padronização na exibição dos dados que foi adotada pelas páginas em HTML. O fato de os dados serem exibidos em uma página usando a linguagem de marcação utilizada para desenvolver páginas na web (HTML) auxiliou na produção de uma algoritmo para percorrer as *tags* necessárias, identificadas pela padronização, e capturar as informações relevantes para a base. Com isso, iniciou-se o procedimento de pré-processamento. A Figura 4 demonstra a estrutura HTML das fontes de dados usadas.

O pré-processamento e o projeto do banco de dados relacional foram feitos paralelamente. O primeiro consistiu em pré-formatar os dados extraídos de uma forma que fosse possível inserí-los no banco. Cada informação foi processada, por meio de algoritmos e/ou manualmente, para que os tipos dos dados fossem equivalentes aos tipos projetados para compor a estrutura do banco.

Figura 3: Dados dos jogos disponibilizados no website CartolaFC

```
    <ul class="jogadores">
<li class="" itemprop="performers" itemscope itemtype="http://schema.org/Person/SoccerPlayer">
    <p>
        <span class="posicao">GOL</span>

        <span class="jogador" itemprop="name"> FÃ¡bio</span>



    </p>
</li>

<li class="" itemprop="performers" itemscope itemtype="http://schema.org/Person/SoccerPlayer">
    <p>
        <span class="posicao">ZAE</span>

        <span class="jogador" itemprop="name"> Gil</span>



    </p>
</li>

<li class="" itemprop="performers" itemscope itemtype="http://schema.org/Person/SoccerPlayer">
    <p>
        <span class="posicao">ZAE</span>

        <span class="jogador" itemprop="name"> LÃ©o</span>
```

Figura 4: Demonstração do padrão HTML das fontes de dados

Para gerar uma primeira base final, possuindo todos dados contidos no banco, foi desenvolvido um algoritmo para modelar esta base de uma maneira que todos os

elementos considerados relevantes, à princípio, fossem agrupados. O agrupamento destas informações foi realizado de um modo que fosse extraídas todas as escalações e saldos de gol dos times em todos os seus respectivos jogos. Assim, o arquivo texto final possui, em cada linha, o número de identificação de cada jogo, o tempo em que foi alterado a escalação (tempo de substituições, por exemplo) o saldo de gol no intervalo de tempo corrente e todos os dados de cada jogador (lembrando que o número do jogo e o tempo corrente são informações que serão retiradas da base quando esta for disponibilizada para mineração). Esta base foi gerada desta maneira para que fossem aplicados os algorítmos de mineração de dados uma primeira vez, sem que alguns elementos fossem retirados.

TIME 1												TIME 2												SALDO
GOL			DEF 1			...			ATA 5			GOL			DEF 1			...			ATA 5			
FS	PE	...	FS	PE	...	FS	PE	...	FS	PE	...	FS	PE	...	FS	PE	...	FS	PE	...	FS	PE	...	
0,29	0,11	...	1,35	2,29	...	1,54	2,97	...	-2	-2	...	0,25	0,50	...	0,85	3,65	...	1,09	2,67	...	2,8	3,17	...	NEU
0,08	0,26	...	0,29	1,53	...	0,50	1,25	...	-2	-2	...	0,17	0,33	...	0,47	1,06	...	0,50	1,50	...	1,76	2,29	...	POS
0,15	0,00	...	0,93	1,73	...	1,53	1,77	...	-2	-2	...	0,20	0,40	...	0,6	1,96	...	1,89	2,70	...	0,25	0,50	...	NEG

Figura 5: Formato padrão da base final

6 Resultados Preliminares

Como resultados preliminares, listados para serem apresentados nesta disciplina Monografia I, destaca-se o desenvolvimento de um banco de dados relacional projetado para armazenar os dados coletados de dois anos do campeonato brasileiro de futebol e uma primeira base gerada com todos os atributos possíveis para iniciar a aplicação de métodos de mineração de dados. Não foi aplicado nenhum tipo de filtro nos atributos desta primeira base.

O banco de dados será utilizado para que sejam depositados todas as informações e, com isso, para que possa ser gerada a base. A base final poderá ser usada, por exemplo, para que possa ser encontrados padrões capazes de auxiliar na predição se uma substituição durante um jogo resultará em um saldo positivo ou negativo para o time.

7 Trabalhos Futuros

Como trabalho futuro, é visado o desenvolvimento do repositório, propriamente dito, capaz de gerar uma base filtrada de acordo com a necessidade de cada usuário. Assim, um usuário que queira extrair uma base com o propósito de aplicá-la em métodos de mineração de dados, por exemplo, poderá selecionar apenas os atributos realmente relevantes à sua aplicação.

8 Cronograma de atividades

A Tabela 1 apresenta o cronograma de realização das atividades vinculadas à disciplina Monografia II.

Tabela 1: Cronograma de Atividades.

Atividades	**Dez**	**Jan**	**Fev**	**Mar**	**Abr**
Extração dos dados atuais	X				
Pré-processamento nos novos dados	X	X			
Alimentação do banco com mais dados atuais		X			
Pré-formulação de consultas SQL		X	X		
Projeto e implementação da interface web		X	X	X	
Testes do sistema				X	X
Desenvolvimento do texto da monografia				X	X
Apresentação					X

Referências

[1] Cartola fc. `http://sportv.globo.com/site/cartola-fc/`, 2011.

[2] Futpédia. `http://futpedia.globo.com/`, 2011.

[3] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Elsevier, 2 edition, 2006.

[4] Osama K. Solieman Robert P. Schumaker and Hsinchun Chen. *Sports Data Mining*. Springer, 1 edition, 2010.