

BCC390 - Monografia I

DESENVOLVIMENTO DE UM REPOSITÓRIO DE DADOS DO FUTEBOL BRASILEIRO

Rafael Belini Souza

November 9, 2012

Tópicos

- 1 **Introdução**
- 2 **Justificativa**
- 3 **Objetivo**
- 4 **Fontes**
- 5 **Pré-processamento**
- 6 **Metodologia**
- 7 **Resultados**
- 8 **Conclusão**

Introdução

- O desenvolvimento de um repositório de dados do futebol brasileiro oferece suporte para novas pesquisas.
- A carência de uma fonte dos dados centralizada é uma justificativa para o desenvolvimento deste repositório.
- Dados coletados dos websites CartolaFC e Futpédia.
- Dados pré-processados para eliminar inconsistências.
- Será projetada e desenvolvida uma interface web para que os usuários possam resgatar os dados do sistema.

Justificativa

- Ainda há uma carência referente à pesquisas e aplicações com mineração de dados no futebol.
- Um repositório centralizado proporcionará o acesso a dados coletados de diferentes fontes.
- Será possível alavancar a pesquisa com mineração de dados no futebol.

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.

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

Seleção de atributos

Seleção de atributos é a técnica para selecionar subconjuntos de características relevantes para contruir modelos de aprendizagem robustos, processo no qual um subconjunto de M atributos é escolhido de N , obedecendo a restrição $M \leq N$. Seleção de atributos garante que a base de dados utilizada na fase de classificação tem uma qualidade melhor.

Razões para Seleção de Atributos

- Modelo simplificado
 - Mais transparente
 - Mais fácil de ser interpretado
- Classificação mais eficaz e eficiente
 - Resulta em uma acurácia melhor
 - Classificação mais precisa
- Conhecimento
 - Conhecer quais atributos são importantes pode ser útil para a aplicação

Cartola

lista de jogadores

JOGADORES	JOGOS	PREÇO		PONTUAÇÃO		CONFRONTO	SCOUT CONSOLIDADO: <input type="text" value="Atacando"/>									
		Atual	Var.	Med.	Ult.		FS	PE	A	FT	FD	FF	G	I	PP	
LAT Chiquinho <input type="button" value="Cv"/>	14	€ 2.73	0.00	0.35	0.00	X	18	42	0	0	0	3	0	0	0	
MEI Erandir <input type="button" value="Cv"/>	7	€ 2.60	0.00	1.47	0.00	X	2	6	0	0	0	0	0	1	0	
MEI Robston <input checked="" type="checkbox"/> <input type="button" value="Cv"/>	32	€ 2.50	-2.18	2.61	-4.10	X	21	105	5	2	25	27	5	3	1	
MEI Anafilson <input type="button" value="Cv"/>	26	€ 2.49	0.08	1.19	0.70	X	28	52	3	0	2	3	1	4	0	
ATA Pedro Paulo <input type="button" value="Cv"/>	12	€ 2.32	0.00	2.15	0.00	X	11	23	1	1	9	7	2	7	0	
ATA Diogo Campos <input type="button" value="Cv"/>	3	€ 2.25	0.00	1.27	0.00	X	5	4	1	0	0	0	0	2	0	
MEI William <input type="button" value="Cv"/>	20	€ 2.15	-1.09	1.68	-0.50	X	13	37	1	1	1	5	2	3	0	
ZAG Jairo <input type="button" value="Cv"/>	20	€ 2.15	0.00	0.90	0.00	X	15	26	0	0	2	0	0	1	0	
MEI Pituca <input checked="" type="checkbox"/> <input type="button" value="Cv"/>	33	€ 2.01	-1.07	0.99	1.90	X	47	73	0	0	2	6	0	1	0	
GOL Roberto <input type="button" value="Cv"/>	0	€ 2.00	0.00	0.00	0.00	X	0	0	0	0	0	0	0	0	0	
MEI ... <input type="button" value="Cv"/>	0	€ 2.00	0.00	0.00	0.00	X	0	0	0	0	0	0	0	0	0	

Futpédia

Cruzeiro		6 × 1		Atlético-MG	
	Cruzeiro		Atlético-MG	ficha do jogo	
TEC Vagner Mancini		TEC Cuca		GOLS CARTÕES RENDA, PÚBLICO e ARBITRAGEM	
GOL	Rafael	GOL	Renan Ribeiro	1x0	 Roger 9'/1tr
ZAE	Naldo	ZAD	Réver 	2x0	 Leandro Guerreiro 28'/1tr
ZAE	Léo	ZAE	Leonardo Silva 	3x0	 Anselmo Ramon 33'/1tr
ZAD	Victorino	ZAD	Werley 	4x0	 Fabrício 45'/1tr
LAD	Diego Renan 	LAD	Carlos Cesar	5x0	 Wellington Paulista 11'/2tr
MEC	Roger  	VOL	Richarlyson 	5x1	 Réver 15'/2tr
ATA	Ortigoza 	MEC	Bernard	6x1	 Everton 44'/2tr
VOL	Fabrício 	VOL	Pierre 		
VOL	Charles 	VOL	Serginho 		
ATA	Farias 	ATA	Magno Alves 		
VOL	Leandro Guerreiro  	MEC	Daniel Carvalho		
ATA	Wellington Paulista  	VOL	Fillipe Soutto		
ATA	Anselmo Ramon  	ATA	André		
VOL	Everton 				

Dificuldades encontradas

- Dados disponibilizados de diferentes maneiras
- Ausência de atributos
- Informações diferentes entres as fontes

Conceito

Pré-processamento é responsável por consolidar as informações relevantes para que sejam aplicadas nos algoritmos de mineração, por exemplo, tentando reduzir a complexidade do problema.

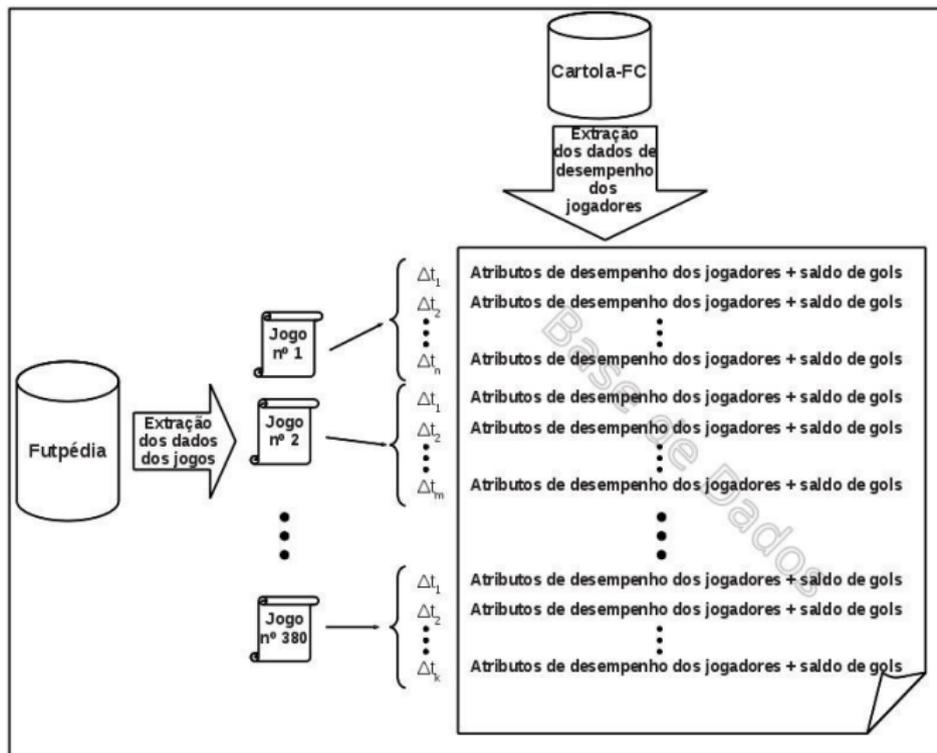
Atributos

- Remover atributos irrelevantes ou redundantes - Retirar os atributos relevantes ou mantendo os irrelevantes pode causar inconsistência para o algoritmo aplicado, resultando na descoberta de um padrão pobre
- Selecionar somente os atributos relevantes - Redução do conjunto de dados removendo os atributos irrelevantes ou redundantes

Trabalho realizado

- Ignorar instâncias
- Preencher valores manualmente
- Utilizar uma constante global para preencher valores ausentes (-1)

Metodologia



Desenvolvimento

TIME 1												TIME 2												S A L D O		
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	...		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		

Desenvolvimento

```

1
2 INSERT INTO jogador VALUES ('1','saopaulo','MEI','caseiro','18','28','44','1','0','2','6','2','0','0','0','40','36','0','7','0','0','0','0','0','2010');
3 INSERT INTO jogador VALUES ('2','saopaulo','GOL','bosco','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
4 INSERT INTO jogador VALUES ('3','saopaulo','GOL','denis','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
5 INSERT INTO jogador VALUES ('4','saopaulo','MEI','wellington','4','5','13','0','0','1','0','0','0','0','6','9','0','1','0','0','0','0','0','0','2010');
6 INSERT INTO jogador VALUES ('5','saopaulo','LAT','thiago carleto','2','5','5','0','0','2','2','0','0','0','2','3','0','0','0','0','0','0','0','0','2010');
7 INSERT INTO jogador VALUES ('6','saopaulo','ATA','lucas gauchó','5','2','7','0','0','3','0','2','0','0','1','2','0','0','0','0','0','0','0','0','2010');
8 INSERT INTO jogador VALUES ('7','saopaulo','MEI','sergio mota','3','0','3','0','0','0','1','0','0','0','0','1','0','0','0','0','0','0','0','0','0','2010');
9 INSERT INTO jogador VALUES ('8','saopaulo','ZAG','samuel','8','6','14','0','0','0','0','0','0','0','11','16','0','2','0','0','0','0','0','0','2010');
10 INSERT INTO jogador VALUES ('9','saopaulo','MEI','ze vitor','8','6','17','0','0','1','1','0','0','0','0','13','0','1','0','0','0','0','0','0','0','2010');
11 INSERT INTO jogador VALUES ('10','saopaulo','ATA','alfredo','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
12 INSERT INTO jogador VALUES ('11','saopaulo','LAT','diogo','6','7','22','0','0','0','1','0','0','0','5','10','0','0','0','0','2','0','0','0','2010');
13 INSERT INTO jogador VALUES ('12','saopaulo','ZAG','bruno uvini','2','0','0','0','0','0','0','0','0','0','0','0','0','1','1','0','0','0','0','0','0','0','2010');
14 INSERT INTO jogador VALUES ('13','saopaulo','LAT','daniilo','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
15 INSERT INTO jogador VALUES ('14','saopaulo','GOL','leonardo','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
16
17 INSERT INTO jogador VALUES ('15','palmeiras','ATA','dinei','9','16','11','0','0','0','3','7','2','0','0','4','15','0','1','0','0','0','0','0','0','2010');
18 INSERT INTO jogador VALUES ('16','palmeiras','MEI','edinho','30','35','105','2','0','3','8','1','2','0','5','70','1','11','0','0','0','0','0','0','0','0','2010');
19 INSERT INTO jogador VALUES ('17','palmeiras','ATA','lenny','2','0','3','0','0','0','0','0','0','1','0','1','1','0','0','0','0','0','0','0','0','0','2010');
20 INSERT INTO jogador VALUES ('18','palmeiras','MEI','anselmo','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
21 INSERT INTO jogador VALUES ('19','palmeiras','MEI','joao arthur','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
22 INSERT INTO jogador VALUES ('20','palmeiras','ATA','max','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
23 INSERT INTO jogador VALUES ('21','palmeiras','ZAG','leandro amaro','6','3','4','0','0','1','0','1','0','0','3','7','0','1','0','2','0','0','0','0','0','0','2010');
24 INSERT INTO jogador VALUES ('22','palmeiras','ATA','luan','13','16','29','1','0','5','9','0','7','0','14','19','0','3','0','0','0','0','0','0','0','0','2010');
25 INSERT INTO jogador VALUES ('23','palmeiras','GOL','raphael alemão','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
26 INSERT INTO jogador VALUES ('24','palmeiras','MEI','bruno oliveira','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
27 INSERT INTO jogador VALUES ('25','palmeiras','MEI','julio cesar','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
28 INSERT INTO jogador VALUES ('26','palmeiras','MEI','bruno turco','2','0','3','0','0','0','0','0','0','0','0','2','0','0','0','0','0','0','0','0','0','0','2010');
29 INSERT INTO jogador VALUES ('27','palmeiras','LAT','luis felipe','5','2','10','0','0','1','0','0','0','0','4','1','0','1','0','1','0','0','0','0','0','0','2010');
30 INSERT INTO jogador VALUES ('28','palmeiras','ATA','patrik','18','13','20','0','0','6','6','0','0','0','9','12','0','4','0','0','0','0','0','0','0','0','2010');
31 INSERT INTO jogador VALUES ('29','palmeiras','ZAG','gualberto','2','1','0','0','0','0','0','0','0','0','0','3','0','1','1','0','0','0','0','0','0','0','2010');
32 INSERT INTO jogador VALUES ('30','palmeiras','MEI','gilsinho','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
33 INSERT INTO jogador VALUES ('31','palmeiras','GOL','carlos','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
34 INSERT INTO jogador VALUES ('32','palmeiras','GOL','fabio','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
35 INSERT INTO jogador VALUES ('33','palmeiras','ZAG','mayko','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2010');
36 INSERT INTO jogador VALUES ('34','palmeiras','MEI','jean','2','0','2','0','0','0','1','0','0','0','0','1','2','0','0','0','0','0','0','0','0','0','2010');
37

```

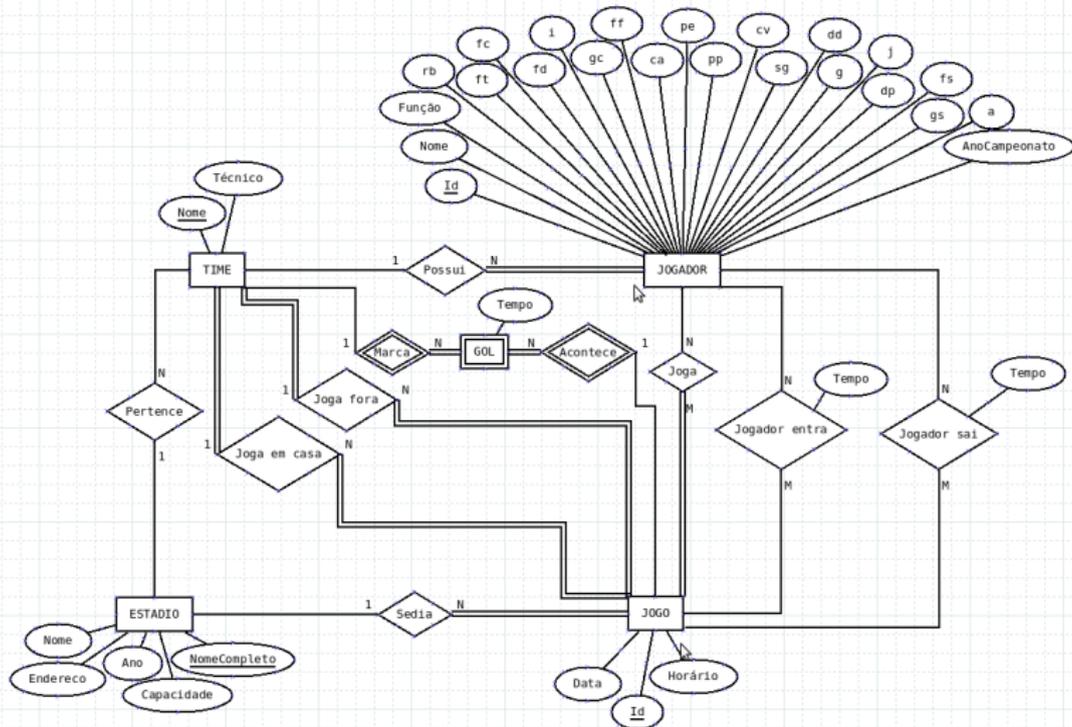
Desenvolvimento - Banco de dados

	id [PK] integer	ntime character	funcao character	nome character	j integer	fs integer	pe integer	a integer	ft integer	...	
1	1	AtleticoGO	ATA	Elias	26	58	76	5	3	...	
2	2	AtleticoGO	GOL	Márcio	35	3	12	0	0	...	
3	3	AtleticoGO	GOL	Edson	3	0	0	0	0	...	
4	4	AtleticoGO	TEC	Renê Simão	25	0	0	0	0	...	
5	5	AtleticoGO	ZAG	Welton Fel	18	...	22	28	0	1	...
6	6	AtleticoGO	LAT	Thiago Felt	25	44	82	2	1	...	
7	7	AtleticoGO	ZAG	Gílson	20	6	20	0	1	...	
8	8	AtleticoGO	MEI	Ramalho	19	29	51	0	0	...	
9	9	AtleticoGO	MEI	Rômulo	10	3	15	0	1	...	
10	10	AtleticoGO	ATA	Josiel	7	5	4	0	0	...	

Resultados preliminares

- Banco de dados projetado e alimentado.
- Base gerada para testar a consistência do banco de dados (identificou-se os jogos, os tempos em que a escalação do time fosse alterada, o saldo de gols que aquela escalação específica implicou e, em seguida, foram selecionados os atributos particulares de cada jogador).

Resultados preliminares



Conclusão

Com as seções de extração e pré-processamento dos dados já elaboradas, concentra-se os esforços para a pré-formulação das consultas no banco, partindo para o desenvolvimento do repositório.

Trabalhos futuros

Desenvolver o repositório capaz de gerar uma base filtrada de acordo com a necessidade do 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 relevantes à sua aplicação. Esta etapa incluirá a formulação de diversas consultas que permitirão ao usuário extrair diferentes bases de dados para a etapa de mineração.

Cronograma de atividades

Table: 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

