

Identificação de Perfis de Torcedores no Twitter

Gabriel M. R. de Paula¹, Clodoveu A. Davis. Jr²

¹Departamento de Eng. Eletrônica – Universidade Federal de Minas Gerais
Caixa Postal 209 – 30.161-970– Belo Horizonte– MG – Brazil

²Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Caixa Postal 486 – 31.270-901 – Belo Horizonte– MG – Brazil

`gabriel.mrp@ufmg.br, clodoveu@dcc.ufmg.br`

Abstract. *The goal of this work is to investigate the existence of patterns among soccer fans on Twitter, centering the analysis into the space and time characteristics of fan's posts, when related to their soccer team, proposing a relationship with the events that compose the championship.*

Resumo. *O objetivo deste trabalho é investigar a existência de padrões de torcedores no Twitter, centrando a análise nas características de espaço e tempo de mensagens de fãs, quando relacionados com a seu time de futebol, propondo uma relação com os eventos que compõem o campeonato.*

1. Proposta

O número de adeptos às redes sociais tem crescido expressivamente nos últimos anos. O *Twitter* tem sido usado como base em muitos estudos, tais como a identificação de público alvo em pesquisas de mercado. Mensagens do *Twitter*, por serem sucintas e surgirem em grande volume, são muito utilizadas para a avaliação das tendências e atitudes do coletivo ante aos fatos que lhe sejam marcantes. O presente trabalho busca identificar e caracterizar perfis de usuários interessados em futebol, com base nas mensagens postadas pro eles no *Twitter*.

Inicialmente, a busca pela identificação de perfis de torcedores se concentra nas características espaço-temporais das mensagens referentes aos clubes do campeonato brasileiro, a fim de se estabelecer relações com os eventos que compõem a competição. Tal abordagem é viabilizada pela existência de um registro temporal associado a cada mensagem e, caso assim configurado pelo usuário, pode existir também informação sobre a localização geográfica do emissor [Davis et al. 2011].

A partir dessas informações, é possível determinar aspectos como a influência do mando de campo no comportamento da torcida, assumindo uma concentração regional para cada equipe, assim como a reação do torcedor de acordo com o resultado da partida. Tais análises possibilitam que sejam identificadas características similares entre torcedores, possibilitando a detecção de padrões de comportamento dos torcedores através do *Twitter*. Para o estudo aqui apresentado, são utilizados dados obtidos por coleta automática e em tempo real de mensagens referentes ao Campeonato Brasileiro de 2012, realizada pelo projeto *Observatório do Brasileirão*¹, do Instituto Nacional de Ciência e Tecnologia para a Web (InWeb). O estudo conta com o apoio do Centro de Convergência de Novas Mídias da UFMG (CCNM), também participante do InWeb.

¹ <http://www.observatorio.inweb.org.br/>

2. Metodologia

O trabalho, do ponto de vista técnico, consiste na extração de conhecimento de um banco de dados, que no caso é formado por um conjunto estruturado de registros de mensagens. Para tal, são empregadas técnicas de Mineração de Dados. A metodologia de desenvolvimento divide-se em três etapas. Na primeira etapa realizou-se o pré-processamento dos dados, selecionando aqueles de interesse para análise. Na segunda, os dados obtidos são transformados em indicadores para o comportamento do usuário, e na terceira etapa é feita a identificação de perfis de usuários.

A quantidade de itens disponíveis no banco de dados é muito grande, da ordem de dezenas de milhões de elementos, justificando o emprego de técnicas de redução do tamanho do banco de dados. Uma das decisões tomadas com tal propósito é a de se restringir o estudo às mensagens que referenciam um time específico. Por motivos de popularidade, e de modo a minimizar a ocorrência de problemas de ambiguidade de nome, escolheu-se o Flamengo (RJ).

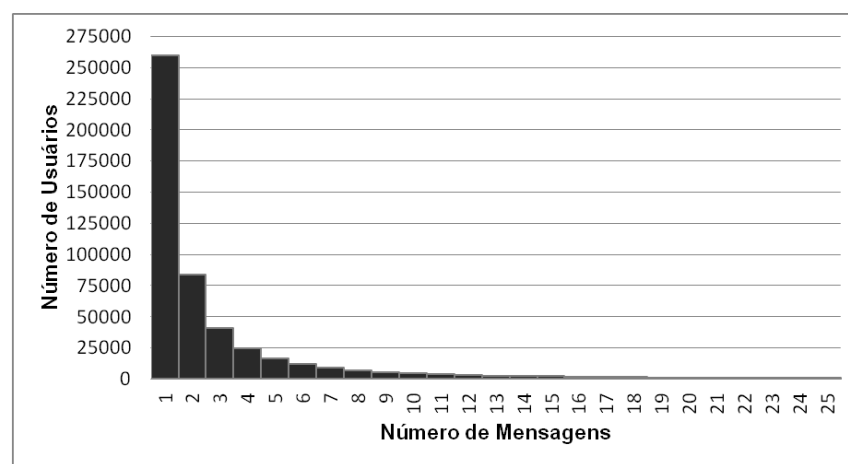


Figura 1: Distribuição dos usuários pelo número de mensagens postadas

Outra estratégia adotada para redução do volume de dados consiste na remoção de registros correspondentes a usuários que não publicam muitas mensagens a respeito do tema proposto. Como se pode perceber pela Figura 1, tal parcela é bem significativa. A figura mostra que a distribuição do número de usuários que postam certo número de mensagens segue uma função potência, ou seja, poucos usuários postam uma grande quantidade de mensagens, e existe uma grande quantidade de usuários que emitem poucas mensagens. A seleção de atributos apresentada por Han et al. (2011) é empregada a fim de se manter apenas os campos de interesse para análise.

A correspondência entre o momento e o local de postagem da mensagem e os acontecimentos do Campeonato Brasileiro é traduzida em números através de estratégias de sumarização descritiva dos dados. Estas, segundo Han et al. (2011) consistem em contagem, obtenção de valores médios e de variância. Tal procedimento corresponde à segunda etapa do trabalho. A última etapa do trabalho tem como objetivo a identificação de comportamentos de torcedores via categorização através de árvores binárias de indução. A agregação dos resultados de diversas análises seguindo esse procedimento determinará o conjunto de perfis de torcedores. A Figura 2 mostra a classificação comportamental de torcedores de acordo com a repercussão do resultado das partidas do time.

A avaliação é feita através da contagem de postagens após vitórias (PAV) e após derrotas (PAD). Caso a discrepância entre esses números fique abaixo de um limiar (LIM) define-se que não há uma dominância observada. É importante pontuar que técnicas de normalização devem ser adotadas para que usuários se mantenham equiparados.

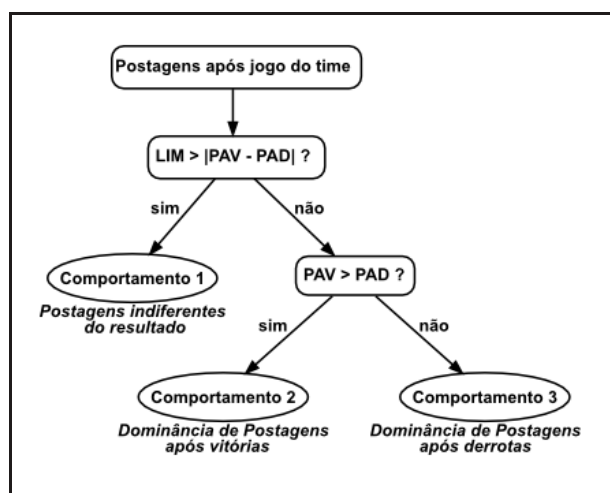


Figura 2: Identificação comportamental aplicada à repercussão após os jogos do time de interesse.

Para a validação dos perfis obtidos é necessário determinar como será feita a comprovação do tipo de torcedor identificado e o método estatístico de validação.

3. Considerações Finais

O trabalho se encontra em desenvolvimento. Pretende-se, com essa exposição, discutir as metodologias empregadas, bem como o teor conceitual do trabalho. Abre-se espaço para novas proposições de análise e para uma ampla gama de trabalhos futuros no escopo do Observatório do Brasileirão, tal como a análise da dispersão espacial dos emissores das mensagens [Backstrom et al., 2008].

Agradecimentos

Este trabalho é parcialmente apoiado pelo InWeb (CNPq processo 573871/2008-6), e por projetos CNPq (308678/2012-5) e FAPEMIG (CEX-PPM-00466-11).

Referências

- Backstrom, L., Kleinberg, J., Kumar, R., Novak, J. Spatial Variation in Search Engine Queries. Proceedings of the 17th International Conference on WWW. Beijing, China, 2008.
- Davis Jr., C. A., Pappa, G. L., Oliveira, D. R. R., Arcanjo, F. L. Inferring the Location of Twitter Messages Based on User Relationships. Transactions in GIS 15(6): 735-751, 2011.
- Han, J., Kamber, M. Data Mining Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2001.