PREVISÃO DE EPIDEMIAS COM BASE EM DADOS DE REDES SOCIAIS

Denis Silva Costa[1], Giuliano Araujo Bertoti[2]
[1] [2]FATEC São José dos Campos prof. Jessen Vidal
denis.costa3@fatec.sp.gov.br, giuliano.bertoti@fatec.sp.gov.br

1. Introdução

No decorrer da história moderna várias epidemias deixaram marcas no curso da humanidade. E umas das maiores dificuldades que é encontrada, até os dias atuais, em determinadas regiões, é a constatação rápida da epidemia, para que o combate à doença possa ser feito de maneira ágil e efetiva.

Com o uso da tecnologia, e usando uma fonte de dados informal, como as redes sociais, é possível determinar, que uma região específica está sofrendo com alguma moléstia[1].

O objetivo deste trabalho é determinar a ocorrência de uma possível epidemia usando técnicas de mineração de dados em redes sociais, onde se procura por padrões textuais referentes à possíveis doenças e métricas estatísticas na determinação de um possível aumento de dados referentes as doenças pesquisadas.

2. Metodologia e materiais

Para alcançar os objetivos do trabalho foi desenvolvido um programa que armazena as informações referentes às doenças, originadas no *Twitter*, mídia social utilizada neste experimento.

A técnica de mineração de dados empregada foi a utilização do algoritmo de Mapreduce[2], que minera os dados postados pelos usuários gerando informações estatisticas.

Para melhorar e facilitar a distribuição das informações geradas como resultado deste trabalho, os pontos das possíveis epidemias serão mostrados em uma interface *web*, usando mapas. O que facilitaria aos que pesquisarão as informações a localizarem dos pontos de seu interesse.

3. Resultados

A metodologia usada neste trabalho depende da geolocalização das informações utilizadas. Neste caso específico, que o usuário habilite a geolocalização no seu perfil do Twitter. A publicação no Twitter geolocalizada ficará como ilustrado na Figura 1.

O sistema é composto de um *crawler* que, periodicamente procura por informações referentes às doenças no Twitter. Essa informação é então guardada em um banco de dados especializado em armazenar estruturas do tipo JSON[3], neste caso MongoDB[4]. JSON é o formato que será obtido nas pesquisas no Twitter.

O algoritmo de MapReduce é responsável por quantificar as citações das doenças de forma diária. Assim é possível acompanhar as alterações de maneira mais aprofundada, do que se o agrupamento fosse feito semanalmente, por exemplo.

Caso exista um crescimento em um determinado tempo, das citações à uma doença em um ponto geográfico, este ponto é então mostrado em um mapa, que será disponibilizado com uma interface web.

Todo o desenvolvimento foi realizado de forma aberta (*open source*) e pode ser visto em https://github.com/deniscostadsc/previsaodeepidemias.

Figura 1 - Publicação no Twitter geolocalizada

4. Conclusões

Com o uso das informações geradas pelo trabalho é possível identificar possíveis epidemias, com antecedência em relação aos métodos tradicionais. Isso pode melhorar o combate à várias doenças, além de diminuir o tempo de resposta dos órgão responsáveis no combate às epidemias em geral.

Em sequencia, para otimizar os resultados, pretende-se adicionar um processamento de linguagem natural[5] que poderia retirar das postagem coletadas as que se enquadrarem como falso positivo.

5. Referências

[1] R. Chunara, J. R. Andrews, J. S. Brownstein, Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak, The American Journal of Tropical Medicine and Hygiene, 2011

[2] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, Communications of the ACM, 2008

[3] D. Crockford, The application/json Media Type for JavaScript Object Notation (JSON), Internet informational RFC 4627, 2006.

[4] K. Chodorow, M. Dirolf, MongoDB: The Definitive Guide, O'Reilly Media, 2010.

[5] W. Woods, Transition Network Grammars for Natural Language Analysis, Communications of the ACM, 1970.