UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

AUGUSTO DIAS PEREIRA DOS SANTOS

Descobrindo eventos locais utilizando análise de séries temporais nos dados do Twitter

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Prof. Dr. Leandro Krug Wives
Orientador

Prof. Dr. Luis Otávio Campos Álvares
Coorientador

Porto Alegre, março de 2013.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Dias Pereira dos Santos, Augusto

Descobrindo eventos locais utilizando análise de séries temporais nos dados do Twitter / Augusto Dias Pereira dos Santos. -- 2013.

71 f.

Orientador: Leandro Krug Wives. Coorientador: Luis Otávio Campos Álvares.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Informática, Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2013.

1. Redes Sociais. 2. Mineração de Dados. 3. Séries Temporais. 4. Identificação de Eventos. I. Krug Wives, Leandro, orient. II. Campos Álvares, Luis Otávio, coorient. III. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos Alexandre Netto
Vice-Reitor: Prof. Rui Vicente Oppermann
Pró-Reitor de Pós-Graduação: Prof. Vladimir Pinheiro do Nascimento
Diretor do Instituto de Informática: Prof. Luís da Cunha Lamb
Coordenador do PPGC: Prof. Luigi Carro
Bibliotecária-Chefe do Instituto de Informática: Beatriz Regina Bastos Haro

AGRADECIMENTOS

Primeiramente gostaria de agradecer a minha família pelo apoio incondicional e incentivo aos estudos, desde quando eu ainda nem sabia a importância que isso teria na minha vida e a diferença que faria para minha carreira. Por ordem cronológica a este trabalho, gostaria de agradecer ao meu primeiro orientador, Luis Otavio Álvares, por acreditar no meu potencial e nas minhas ideias ao iniciar este trabalho, assim como contribuir em grande parte com esta pesquisa. Ao Instituto de Informática, seus técnicos administrativos, professores e alunos, por todo apoio técnico e pedagógico que possibilitaram o desenvolvimento deste trabalho. Ao CPD/UFRGS, seus diretores e chefias por permitir e apoiar o minha qualificação acadêmica através deste mestrado. Aos meus colegas de mestrado pelas conversas que auxiliaram no amadurecimento deste trabalho, principalmente ao Rafael Pinto com o fornecimento e auxilio no entendimento e utilização da IGMN. Ao meu irmão mestre em Computação, Henrique Dias, pela construção colaborativa do conhecimento sobre Web crawlers, programação e análise de redes sociais. Ao meu segundo orientador, Leandro Wives, por dar continuidade à orientação, dando um rumo mais objetivo para a conclusão da pesquisa e da dissertação. Aos revisores deste texto, pelo excelente trabalho em fazer com que minha transmissão do conhecimento se tornasse compreensível, meu colega de faculdade, Christian Potter, minha mãe, Maria Zélia Dias e minha irmã, Ana Helena Dias. À minha querida namorada, Vanessa Azeredo, pelos puxões de orelha, revisão do texto e afeto incondicional. À vida, por proporcional uma viagem tão agradável pelo desconhecido, fantástico e belo.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
ARIMA	Autoregressive Integrated Moving Average
CMC	Computer-Mediated Communication
CSSN	Computer-Supported Social Network
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCA	Degree of Crowd Activity
DCBD	Descoberta de Conhecimento em Base de Dados
GIS	Geographic Information Systems
GPS	Global Positioning System
IGMN	Incremental Gaussian Mixture Network
IP	Internet Protocol
IRC	Internet Relay Chat
JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Databases
NYSE	New York Stock Exchange
RFID	Radio-Frequency IDentification
RMSE	Root-Mean-Square Error
STL	Seasonal Decomposition of Time Series by Loess
TS	Time Series
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
UTC	Universal Time Coordinated

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

O crescente uso de redes sociais gera quantidades enormes de dados que podem ser empregados em vários tipos de análises. Alguns desses dados têm informação temporal e geográfica, as quais podem ser usadas para posicionar precisamente a informação no tempo e no espaço. Nesse contexto, neste trabalho é proposto um novo método para a análise do volume massivo de mensagens disponível no Twitter, com o objetivo de identificar eventos como programas de TV, mudanças climáticas, desastres e eventos esportivos que estejam ocorrendo em regiões específicas do globo. A abordagem proposta é baseada no uso de uma rede neural para detecção de outliers em séries temporais, as quais são formadas por estatísticas coletadas em tweets localizados em diferentes divisões políticas (i.e., países, cidades). Esses outliers são usados para identificar eventos como um comportamento anormal nos dados Twitter. A efetividade do método é avaliada comparando os eventos identificados com notícias nos meios de comunicação.

Palavras-Chave: Microblogs, Análise Sócio-Geográfica, Fluxo do Twitter, Séries Temporais, Redes Neurais.

Location-Based Event Detection on Microblogs

ABSTRACT

The increasing use of social networks generates enormous amounts of data that can be employed for various types of analysis. Some of these data have temporal and geographical information, which can be used to precisely position information in time and space. In this document, a new method is proposed to analyze the massive volume of messages available in Twitter to identify events such as TV shows, climate change, disasters, and sports that are occurring in specific regions of the globe. The proposed approach is based on a neural network used to detect outliers from a time series, which is built upon statistical data from tweets located in different political divisions (i.e., countries, cities). These outliers are used to identify events as an abnormal behavior in Twitter's data. The effectiveness of the method is evaluated by comparing the events identified on the news media.

Key words: Microblogs, Socio-Geographic Analysis, Twitter Stream, Time Series, Neural Network.

1 INTRODUÇÃO

Uma rede social é um conjunto de pessoas ou grupos com algum padrão de contato ou interação entre si. Os padrões de amizade entre os indivíduos, relações de negócio entre empresas e os casamentos entre famílias são exemplos de redes que foram estudadas no passado (NEWMAN, 2003). O estudo de redes sociais, na computação, existe desde que essas eram formadas apenas pela troca de mensagens entre pessoas utilizando computadores (WELLMAN et al., 1996). Com a evolução da tecnologia, os sites de redes sociais permitem que pessoas troquem mensagens, fotos, vídeos, formem comunidades, etc. utilizando-as através de computadores, tablets, celulares, televisão entre outros dispositivos. Os dispositivos móveis permitem que as pessoas façam essa interação em tempo real, ao longo do seu dia-a-dia, e ainda indicando a posição no globo onde elas estão. Isso possibilita que essas informações geolocalizadas sejam utilizadas como uma rede de sensores para identificação de eventos naturais (SAKAKI; OKAZAKI; MATSUO, 2010) e sociais (LEE; WAKAMIYA; SUMIYA, 2011).

A identificação de eventos começou a ser estudada como parte do problema de detecção e rastreamento de tópicos (ALLAN et al., 1998). Eventos são tópicos colocados em um lugar no espaço e no tempo (ALLAN; PAPKA; LAVRENKO, 1998). As pesquisas citadas utilizavam notícias de jornais como fonte de informação, mas, nas redes sociais computacionais, a identificação de tópicos se utiliza das mensagens trocadas entre os usuários. Essas mensagens possuem informações de quando foram criadas, i.e., tempo, e algumas também incluem onde, i.e., espaço. Assim é também possível fazer a identificação de eventos utilizando as informações presentes nos dados de redes sociais (LEE; WAKAMIYA; SUMIYA, 2011).

A identificação de eventos em redes sociais computacionais pode melhorar significativamente a interação dos seus usuários na busca por conteúdo local (BECKER; NAAMAN; GRAVANO, 2011). Em casos de eventos naturais ou sociais que ameaçam a integridade de uma população, por exemplo, essa identificação pode ajudar os órgãos governamentais na busca por melhorar suas respostas a eventos de emergência (VIEWEG et al., 2010) e na criação de sistemas de alarme (SAKAKI; OKAZAKI; MATSUO, 2010).

No entanto, o grande volume de dados que as redes sociais geram é computacionalmente muito custoso de analisar. Para tanto, técnicas de mineração podem auxiliar no aumento da eficiência (tempo) do processo de descoberta de conhecimento sobre tais dados. Os algoritmos tradicionais de mineração exigem que os dados tenham relações e tipos bem definidos para a aplicação de classificação, análise de agrupamentos (*clustering*), identificação de padrões sequenciais, regras de associação, entre outros. Entretanto, os dados de redes sociais combinados com a espacialidade e temporalidade dos mesmos criam um domínio com dimensionalidade e relacionamento diferente dos esperados pelas abordagens ou mecanismos tradicionais

de mineração. Por esse motivo, novas abordagens precisam ser desenvolvidas para extrair conhecimento dos dados de redes sociais.

Nesse contexto, o objetivo deste trabalho é propor e avaliar um novo método para identificação de eventos utilizando dados de redes sociais computacionais. Cabe salientar que a definição de eventos utilizada neste trabalho é de tópicos alocados no espaço e no tempo que tenham alterado consideravelmente o volume de informações monitoradas nos dados de uma rede social computacional e que tenham visibilidade nos meios de comunicação. As principais diferenças desse método quando comparado com os existentes é que esse simplifica o processo de separação das mensagens em relação a sua origem geográfica e utiliza métodos estatísticos mais eficazes para a modelagem dos dados e posterior identificação dos eventos.

1.1 Motivação

A mineração de dados, parte do processo de descoberta de conhecimento, possibilita a descoberta de padrões interessantes, previamente desconhecidos e potencialmente úteis. Em redes sociais, esses dados podem representar o comportamento das pessoas em diversos aspectos. Este trabalho se propõe a utilizar o grande volume e a velocidade em que esses dados são gerados nas redes sociais para monitorar atividades que estejam relacionadas com o cotidiano da vida das pessoas.

Quando um conjunto de dados possui informações temporais e espaciais, são chamados de dados espaço-temporais. São poucos os trabalhos que exploram essas dimensões em redes sociais computacionais, fato que levou este trabalho a desenvolver do método aqui proposto.

1.2 Problema

A interação dos usuários em uma rede social (e com uma rede social) está ligada, eventualmente, a eventos que ocorrem no mundo real, de forma que, quando um evento do mundo real ocorre, esse é compartilhado na rede através de fotos, vídeos, comentários, conversas e identificação de presença do usuário em um local (*check-in*). Em algumas redes sociais, essas interações estão associadas a dados de posicionamento geográfico. Nesse contexto, cabe verificar se seria possível utilizar esses dados e informações para identificar quais são os eventos e em que locais eles estão ocorrendo. Se sim, é importante avaliar se isso pode ser realizado de forma automatizada e contínua. Esse tipo de identificação se torna importante para o planejamento urbano, a segurança pública e a logística na medida em que permite avaliar também acontecimentos não planejados. A busca dos usuários por conteúdos locais também é aprimorada à medida que essas informações, sobre os eventos identificados, são disponibilizadas.

1.3 Objetivos

Este trabalho tem como objetivo o desenvolvimento e a avaliação de um novo método para identificação de eventos do mundo real (fora da Internet), utilizando dados georreferenciados do Twitter, uma rede social de microblogs. O Twitter foi escolhido em razão da sua base de dados ser pública e de fácil acesso. Os métodos atuais possuem uma baixa taxa de precisão na identificação de eventos. O método proposto se utiliza de técnicas mais eficazes para a criação de um sistema mais simples de detecção, fazendo

com que a identificação retorne eventos mais relevantes, e ainda sendo possível a sua utilização em um ambiente online.

1.4 Hipótese

As hipóteses levantadas neste trabalho se originam de trabalhos semelhantes e se estendem para as evoluções que este propõe:

Hipótese 1: Através dos dados do Twitter é possível identificar eventos que ocorrem no mundo real (fora da Internet).

Nas publicações ocorridas em redes sociais como o Twitter, os usuários manifestam suas atividades, sentimentos, opiniões e demais relacionamentos com o mundo real (fora da Internet). Determinados eventos propiciam aos usuários uma boa oportunidade de se manifestar nas redes.

Hipótese 2: A utilização de análise de séries temporais sobre os dados de redes sociais aumenta a eficácia (taxa de precisão) da identificação de eventos.

Existem técnicas específicas para cada conjunto de dados, assim como para necessidade que se tem com relação à análise dos mesmos. A utilização de análise de séries temporais para dados gerados ao longo do tempo são mais indicadas do que a utilização de métodos para dados univariáveis.

1.5 Contribuições

As principais contribuições deste trabalho podem ser resumidas nos seguintes itens:

- desenvolvimento de um novo método mais eficaz para identificação de eventos em redes sociais com informação geográfica;
- incorporação de registros que não contenham latitude e longitude, mas sim marcação geográfica em formato textual;
- aplicação de redes neurais para criação de modelos das séries temporais, posterior detecção de outliers e diminuição da diferença temporal entre a ocorrência do evento e sua detecção pelo método;

1.6 Estrutura do Texto

Esta dissertação está dividida em seis capítulos. Neste capítulo, foram apresentadas as motivações que nos levaram a propor o método descrito na dissertação. O capítulo 2 apresenta uma visão geral das principais áreas de pesquisa relacionadas a este trabalho, a Descoberta de Conhecimento, as Redes Sociais e as Séries Temporais. O capítulo 3 relaciona alguns trabalhos que também desenvolvem métodos para extração de conhecimento utilizando informações geográficas obtidas em redes sociais. O capítulo 4 descreve em detalhes como o método proposto funciona, explicando cada uma das quatro fases: captura de dados, características das medidas, análise das séries temporais e descrição dos eventos. O capítulo 5 apresenta os recursos utilizados durante os experimentos. Ele também descreve a avaliação das medidas e algoritmos empregados assim como os resultados alcançados durante a avaliação do método. O capítulo 6

resume as principais contribuições deste trabalho, apresenta conclusões e uma discussão sobre trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta alguns conceitos importantes que são empregados na descoberta de conhecimento em redes sociais e utilizados ao longo do trabalho. Na seção 2.1 são descritos os conceitos gerais de descoberta de conhecimento e mineração de dados. Na seção 2.2 são apresentados alguns conceitos de redes sociais. Na seção 2.3 há uma descrição sobre a rede social Twitter. A seção 2.4 descreve detalhes sobre análise de séries temporais, conhecimento necessário para o entendimento de vários elementos deste trabalho. A seção 2.5 explica o que são dados geográficos.

2.1 Descoberta de Conhecimento e Mineração de Dados

No livro organizado por Miller e Han (2009) define-se que Descoberta de Conhecimento em Base de Dados (DCBD, em inglês KDD - Knowledge Discovery in Databases) é uma resposta ao grande volume de dados que tem sido coletado e armazenado em bancos de dados científicos ou comerciais. Com a evolução da tecnologia da informação e a adoção em massa de processos de monitoramento e controle em muitos domínios, cria-se novos dados com uma grande riqueza. Há frequentemente muitas informações nesses dados que não são percebidas pelas técnicas tradicionais de análise e consulta. O processo de descoberta tem incentivado investimentos em tecnologia para busca profunda de informações, antes escondidas, que podem ser transformadas em conhecimento para decisões estratégicas ou respostas fundamentais para pesquisas acadêmicas.

Embora o termo “mineração de dados” (*data mining*) seja utilizado popularmente como sinônimo de descoberta de conhecimento, ele é apenas um componente (o mais importante) desse longo processo. A mineração envolve transformar os dados em *informação* ou fatos sobre o domínio descrito pelo banco de dados. A descoberta é um processo de mais alto nível de obtenção de informação através da mineração e transforma essa informação em *conhecimento* (ideias e crenças sobre o domínio) pela interpretação da informação e integração com o conhecimento existente. A Figura 2.1 mostra todas as etapas da Descoberta de Conhecimento.

A DCBD é baseada na convicção de que a informação está escondida nestas grandes bases de dados na forma de *padrões de interesse*. São propriedades e relacionamentos não aleatórios que são válidos, novos, úteis e em última análise compreensíveis. *Válido* significa que o padrão é genérico o suficiente para ser aplicados a novos dados e não somente uma anomalia dos dados que foram analisados. *Novo*, por ser um padrão não trivial e também inesperado. *Útil* implica que o padrão pode levar a ações efetivas, como, por exemplo, sucessos em tomadas de decisão e investigações científicas. *Compreensível* quer dizer que o padrão deve ser simples e que humanos possam interpretá-lo sem dificuldades (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

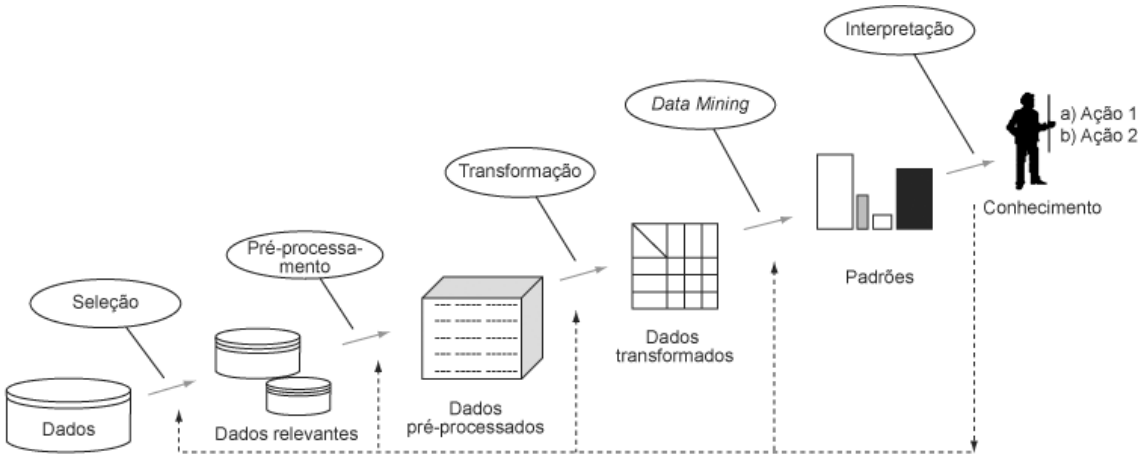

Figura 2.1: Etapas do processo de DCBD

Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, pág.10)

DCBD ultrapassa o domínio tradicional das análises estatísticas para acomodar os dados que normalmente não são favoráveis a esse tipo de análise. As análises estatísticas geralmente envolvem uma quantidade de dados numéricos pequena e limpa (sem ruído), amostrados de uma grande população, para resolver questões específicas que já estão formuladas. Muitos modelos estatísticos requerem suposições rígidas (como a independência, a estacionariedade dos processos subjacentes e a normalidade). Em contraste, os dados que estão sendo coletados e armazenados em bases de dados de grandes empresas são ruidosos, não numéricos, e possivelmente incompletos. Esses dados também são coletados de forma aberta, sem perguntas específicas em mente. A DCBD compreende princípios e técnicas de estatística, aprendizagem de máquina, reconhecimento de padrões, busca numérica e visualização científica para acomodar os novos tipos de dados e volumes de dados a ser gerado através de tecnologias de informação (MILLER; HAN, 2009).

A descoberta de conhecimento é mais fortemente indutiva do que a análise estatística tradicional. O processo de generalização da estatística está inserido no amplo processo dedutivo da ciência. Os modelos estatísticos são de confirmação, exigindo que o analista especifique um modelo, a priori, com base em alguma teoria, teste essas hipóteses e, talvez, reveja essa teoria em função dos resultados. Em contrapartida, profundamente ocultos, os padrões interessantes procurados na mineração são (por definição) difíceis ou impossíveis de especificar a priori, pelo menos com algum grau razoável de completude (MILLER; HAN, 2009). A descoberta está mais preocupada em levar os pesquisadores a formular novas previsões e hipóteses a partir de dados em oposição a testar deduções a partir de teorias através do subprocesso de indução de dados científicos (ELDER IV; PREGIBON, 1996; HAND, 1998). Assim, orienta-se que se a informação solicitada tem pouca descrição, é mais adequado utilizar DCBD do que análises estatísticas (ADRIAANS; ZANTINGE, 1997).

DCBD se encaixa mais na fase inicial do processo dedutivo quando o pesquisador formula ou modifica as teorias baseado em fatos ordenados e observações do mundo real. Nesse sentido, a descoberta está para a área da informação assim como microscópios, sensoriamento remoto e telescópios estão para as áreas atômicas, geográficas e astronômicas, respectivamente. DCBD é uma ferramenta para explorar domínios que são muito difíceis de perceber somente com as capacidades humanas.

Para buscar por uma infinidade de informações, a poderosa e focada estatística não pode competir com a ampla, mas difusa, descoberta de conhecimento. No entanto, a amplitude de análise pode lançar sombras e a descoberta não pode competir com a estatística em poder de confirmação uma vez que o padrão é conhecido (MILLER; HAN, 2009).

Alguns algoritmos de mineração de dados amplamente utilizadas são a classificação, as regras de associação, a identificação e análise de agrupamentos (*clustering*) e a identificação de padrões sequenciais. Esses grupos de algoritmos resolvem um grande número de problemas que envolvem tipos de dados específicos com relações entre os dados razoavelmente bem definidas.

Na classificação, por exemplo, observa-se um determinado número de características e suas classes respectivas que levam à criação de regras para determinar quais características representam cada classe. Dessa forma, um novo registro pode ser comparado com as regras para a identificação de sua classe. Esse é um exemplo de dados que possuem uma relação bem definida para utilização desse tipo de mineração. É comum a obrigatoriedade do uso de dados numéricos nesses algoritmos, já que a decisão se dá comparando se um valor numérico é menor, igual ou maior aos valores referência da classe.

Quando os dados não estão bem arranjados para a utilização desses algoritmos, é necessário fazer uma transformação que converta os dados para o tipo, formato e/ou relação aceita pelo algoritmo escolhido. Nem sempre esta transformação é possível, pois ocasionaria perda significativa de informação. Desta forma, é necessário criar outras técnicas que trabalhem com tipos diferentes de dados e/ou relações diferentes entre eles.

Um tipo de dado que está se tornando muito popular pelo fácil acesso a tecnologias de localização é o dado de posicionamento geográfico, obtidos através da tecnologia GPS ou outras similares. Esse tipo de dado será utilizado por este trabalho para determinar a localidade dos eventos identificados. Para extrair conhecimento desse tipo de dado os algoritmos citados nem sempre são suficientes, já que a combinação entre espacialidade e temporalidade cria um domínio, com dimensionalidade e relacionamento entre os dados, diferente dos aceitos pelos algoritmos citados. Por esta razão Miller e Han (2009) descrevem algumas novas metodologias e paradigmas envolvidos na mineração e descoberta de conhecimento em dados com informações geográficas e temporais, dentre as quais podem ser citadas: clusterização espacial multivariável, descoberta de padrões em trajetória de objetos móveis, correlação entre variáveis com dependência geográfica e exploração visual de dados espaço-temporais.

Outros tipos de informação vêm se tornando populares na internet e demandam novos algoritmos para extração de conhecimento – os dados de redes sociais. O grande diferencial desses dados é o relacionamento existente entre as informações, já que essas são associadas aos usuários e esses têm vínculos sociais. Embora não esteja no escopo deste trabalho, os vínculos sociais são de grande relevância para a mineração de dados, pois indicam o quão relacionado ao usuário estão as informações da rede, podendo, desta forma, trazer conteúdo mais relevante a ele.

A temporalidade associada às informações e também aos relacionamentos dos dados é outro aspecto que ainda não está totalmente e adequadamente coberto pelos algoritmos tradicionais de mineração. A característica de temporalidade está presente nos dados das redes sociais e é amplamente utilizada por este trabalho para ordenar os dados conforme sua ocorrência no tempo, assim como associar os eventos identificados a instantes de tempo do mundo real.

Alguns avanços na mineração dos dados de redes sociais já foram alcançados e os assuntos já bem desenvolvidos em vários grupos de pesquisa:

- análise de sentimentos;
- sistema de recomendação;
- formação de comunidades;
- propagação da informação.

Apesar desses avanços, ainda há vários tipos de informação em redes sociais em que as pesquisas ainda são muito incipientes. As redes sociais têm apresentado uma dinamicidade muito rápida, com a criação de novos tipos de conteúdo, novas formas de relacionamento, interação entre seus usuários e diversos aplicativos que aumentam a complexidade dos dados dessas redes. Mais recentemente, dados de posicionamento geográfico têm sido embutidos nas interações entre os usuários das redes sociais. Devido ao grande volume de informação e por terem características geográficas e temporais, os dados de redes sociais podem ser utilizados para criar um modelo de comportamento dos usuários no mundo real.

2.2 Redes Sociais

Uma rede social é um conjunto de pessoas ou grupos com algum padrão de contato ou interação entre si. Os padrões de amizade entre os indivíduos, relações de negócio entre empresas e os casamentos entre famílias são exemplos de redes que foram estudadas no passado (NEWMAN, 2003). O conceito de redes sociais, na computação, começou a ser formado a partir do artigo de Wellman et al. (1996), que analisa as redes sociais formadas com suporte computacional. Assim é definido que quando redes de computadores conectam pessoas assim como máquinas, essas se tornam redes sociais. A partir desta definição foram apresentados dois termos:

- Redes Sociais Suportadas por Computadores (Computer-Supported Social Networks – CSSNs);
- Comunicação Mediada por Computadores (Computer-Mediated Communication – CMC);

Apesar do estudo acima se restringir às tecnologias de e-mail e chat (IRC), mais utilizadas na época, foi possível identificar elementos que caracterizam estas redes sociais, como os objetivos de comunicação, suporte informativo e social, os tipos de relações presentes, as estruturas sociais que utilizam a computação para esta finalidade e de que forma se dá esta utilização.

No artigo de Garton, Haythornthwaite e Wellman (1997), o termo *online* é agregado a redes sociais, já que essas redes se estruturam através da Internet. O estudo apresenta os diversos aspectos que devem ser observados quando se analisam redes sociais. São eles:

- Unidade de análise: relações, laços, multiplicidade, composição;
- Características da rede: abrangência, centralidade, papéis;
- Particionamento da análise: similaridade, grupo, redes de redes.

Também são identificados que os recursos, no contexto de CMC, podem ser comunicados por via textual, gráfica, animada, áudio ou mídia de vídeo para compartilhar informação (notícias ou dados), discussões de trabalho, dar apoio

emocional ou apenas para passar o tempo com outra pessoa (HAYTHORNTHWAITE apud GARTON; HAYTHORNTHWAITE; WELLMAN, 1997).

Enquanto redes sociais online utilizam a Internet para comunicação, sites de redes sociais online utilizam a Web[1] para conectar os usuários e são geralmente administrados por empresas (p.ex., Google e Facebook). Diferente de como é organizada a Web, em sua maioria em torno de conteúdos, as redes sociais são organizadas em torno dos usuários. Os participantes entram para a rede social, publicam seus perfis pessoais assim como qualquer conteúdo, e podem se conectar a outros usuários com quem estabelecem vínculos. Isso provê a base para manter relacionamentos sociais, achar usuários com interesses similares ou localizar conteúdo e conhecimento que tenha sido publicado por outros usuários (MISLOVE et al., 2007). É importante observar que atualmente é possível a utilização de redes sociais através de aplicativos em dispositivos móveis e não somente via Web.

Os usuários de uma rede social devem se registrar no site, normalmente através de um pseudônimo. Os vínculos formados por esses usuários podem ocorrer com ou sem o consentimento do outro usuário, dependendo das regras dos sites de redes sociais. A maioria dos sites permite aos usuários criar ou se juntar a grupos de interesses específicos, para que os usuários possam compartilhar conteúdos com os participantes do grupo (MISLOVE et al., 2007).

Existe uma rede social em especial, o Twitter, que liga seus usuários através de microblogs (JAVA et al., 2007). Esses microblogs são derivados dos blogs, abreviatura de *web log*, que são diários de textos, imagens e outras mídias, assim como novidades ou conteúdos achados na internet. Alguns blogs contêm simplesmente sentimentos e atividades que seus autores executam no dia-a-dia. O termo micro simboliza a redução do diário à essência da sua mensagem em um texto curto, trazendo como grande vantagem a sua fácil utilização em dispositivos móveis.

2.3 Twitter

O recurso principal do Twitter é permitir aos seus usuários que compartilhem mensagens de até 140 caracteres com sua rede de amigos, e/ou com o mundo, respondendo à pergunta “O que está acontecendo?” (“What’s happening?”). A explosão de utilização da ferramenta se deu por três motivos principais: uma API robusta e versátil que permitiu aos desenvolvedores difundir o uso do serviço entre diversos meios (celulares, navegadores, aplicativos isolados, etc.); adoção por parte da grande mídia e celebridades como fonte de divulgação; aprimoramento da ferramenta pelos usuários.

A rede tem duas características principais: a publicação das mensagens e a funcionalidade de seguir usuários e ser seguido por eles. Ao seguir um usuário você recebe todas as mensagens que esse escreve. Da mesma forma, os usuários que seguem você recebem todas as mensagens que você escreve. Desde seu crescimento acelerado em 2007, esta rede foi alvo de diversos trabalhos para entender o seu uso ou extrair conhecimento a partir de seus dados. Alguns trabalhos estão citados abaixo e outros na seção de trabalhos relacionados (Capítulo 3, página 28).

Em um dos primeiros trabalhos (JAVA et al., 2007) que analisam o uso do Twitter, os autores estudam as propriedade topológicas e geográficas da rede. Através desse

[1] A World Wide Web, ou simplesmente Web, é a maneira de acessar informações através da Internet utilizando o protocolo HTTP.

estudo eles identificaram três principais categorias de usuários no Twitter: os que compartilham informações, os que mantêm relações de amizade e os que buscam por informações. Na primeira delas, os usuários são pessoas ou serviços automatizados que postam notícias e tendem a ter muitos seguidores. Os usuários que buscam amizade são a maioria e incluem familiares, colegas de trabalho e/ou estranhos. Quem busca informação raramente posta mensagens, mas segue outros usuários regularmente.

Java et al. (2007) também identificaram várias categorias de intenções para usar o Twitter: conversas sobre o dia-a-dia, onde usuários discutem eventos de suas vidas e seus pensamentos; compartilhamento de informações e URL; divulgação de notícias, que inclui comentar sobre eventos atuais ou agentes de notícias que postam automaticamente sobre o clima, reportagens e notícias; e conversas entre usuários, onde esses utilizam o símbolo @ para citar outros usuários em suas mensagens. Segundo esse mesmo estudo, essa última intenção é o propósito de 21% dos usuários e 12,5% das mensagens.

Kwak et al. (2010) também realizaram um estudo da topologia de rede e propagação de informação do Twitter, mostrando a relação entre os usuários, suas influências e os tópicos de tendência. Concluíram que há uma alta desigualdade na distribuição de seguidores entre os usuários da rede e uma baixa taxa dos laços de reciprocidade mostrando que o Twitter mais se assemelha com uma rede de compartilhamento de informações do que uma rede social. Kwak et al. (2010) comparam três diferentes medidas de influência: número de seguidores, page-rank e números de retweets (quando um usuário replica a mensagem de outro). Essa comparação mostrou que a lista dos usuários mais influentes difere dependendo da medida utilizada.

Analisando os tópicos de tendência (*trending topics*), palavras-chave que sofrem um grande aumento na utilização pelos usuários durante o seu período de atividade, Kwak et al. (2010) perceberam que, em sua maioria (mais de 85%), os tópicos são manchetes ou notícias que persistem como assunto na rede. Essa lista é exibida na página inicial do Twitter para indicar os 10 assuntos mais falados. Com a massiva utilização da rede social, houve uma evolução nesta listagem que já disponibiliza os tópicos de tendência por pais e até por cidade.

Huberman, Romero e Wu (2008) analisaram mais de 300 mil usuários do Twitter para entender como os usuários se relacionam. Segundo eles, as redes de seguidores e de reciprocidade não são necessariamente o melhor método de medir quem está prestando atenção em quem. Eles apontam que usuários que respondem outro usuário em suas mensagens, utilizando o símbolo @, formam uma rede de laços mais significativa. Utilizando esse método eles verificaram que a escassez de atenção e o ritmo diário de vida e trabalho das pessoas fazem com que elas interajam com aqueles poucos que realmente importam e que retribuem a sua atenção.

Como dito anteriormente, a rede social de microblogs Twitter será utilizada por este trabalho como fonte de dados para identificação de eventos. Esta opção se dá pelo fato dos dados dessa rede serem em sua maioria públicos e pela facilidade de acesso em decorrência da Twitter API. Outro motivo fundamental é o fato dessa rede permitir que seus usuários marquem a posição geográfica em que estão ao enviar as mensagens. Os dados coletados do Twitter serão processados e transformados em séries temporais para posterior detecção de outliers.

2.4 Séries Temporais e sua análise

Séries temporais podem ser definidas como uma coleção de variáveis aleatórias indexadas de acordo com a ordem em que elas são obtidas no tempo (SHUMWAY; STOFFER, 2010), ou então, uma sequência de dados, medidos em pontos sucessivos no tempo e espaçados em intervalos de tempo uniforme[2]. A óbvia correlação introduzida ao amostrar pontos adjacentes no tempo pode restringir a aplicação de vários métodos estatísticos convencionais, tradicionalmente dependentes da suposição de que estas observações adjacentes são independentes e identicamente distribuídas. A abordagem sistemática pela qual se tenta responder às questões matemáticas e estatísticas impostas por estas correlações temporais é referida como análise de séries temporais (SHUMWAY; STOFFER, 2010).

As séries temporais são utilizadas para várias aplicações: na sociologia em nascimentos, mortalidade, desemprego, crimes e divórcios; na economia em produção, taxa de juros e taxa de câmbio; na meteorologia em temperatura, cobertura de nuvens e umidade; em vendas, na estatística de produtos com defeitos produzidos por uma máquina, uma fábrica ou uma companhia inteira; em transito, no número de pessoas que usam ônibus ou o número de carros que cruzaram uma ponte; em medidas fisiológicas do batimento cardíaco, respiração e outras funções corporais; na medição do fluxo de um rio para planejamento e controle de enchentes; e em muitas outras áreas (DARLINGTON, 1990).

Uma série temporal pode tender para cima ou para baixo, como muitas séries econômicas fazem (inflação, valor de ações, etc.), ou podem flutuar em torno de uma média constante, como a temperatura do corpo humano faz. Uma série pode conter ciclos simples, como o ciclo diário da temperatura do corpo, ou pode conter vários ciclos sobrepostos. Por exemplo, a temperatura de uma cidade geralmente exibe ambos ciclos diários e anuais, enquanto a densidade do transito exibe ciclos diários, semanais, e anuais (DARLINGTON, 1990).

Os três principais objetivos da análise de séries temporais são: a previsão de valores futuros de uma série temporal, usando os valores do passado desta série ou a partir de valores de outras séries também; avaliação do impacto de um evento específico, como o efeito de uma nova lei na frequência de motoristas embriagados ou o efeito da aplicação de um pedágio em uma ponte no trafego das pontes vizinhas; e o estudo de padrões eventuais, ou seja, os efeitos das variáveis em vez de eventos na série. Isso pode necessitar de duas ou mais séries. Por exemplo, se as mudanças no desemprego sempre precederem mudanças nas taxas de crime, isso pode indicar que desemprego é uma das causas de crimes (DARLINGTON, 1990).

O primeiro passo para qualquer investigação de uma série temporal sempre envolve um exame minucioso dos dados desenhados ao longo do tempo. Após esse exame, duas abordagens existem para análise de séries temporais, as quais não são mutuamente excludentes, comumente identificadas como abordagem no domínio do tempo e abordagem no domínio de frequência. Enquanto a primeira se preocupa em analisar a correlação entre pontos adjacentes com a dependência dos valores atuais com os valores no passado, a segunda se preocupa primariamente com as características de periodicidade ou sinuosidade sistemática encontrada em grande parte dos dados. Na maioria dos casos o resultado da aplicação dessas abordagens é similar para séries

[2] http://en.wikipedia.org/wiki/Time_series/.

longas, mas o desempenho para observações curtas é melhor quando realizado no domínio do tempo (SHUMWAY; STOFFER, 2010).

As figuras abaixo mostram três séries temporais utilizadas para exercícios de modelagem (COWPERTWAIT; METCALFE, 2009, SHUMWAY; STOFFER, 2010), desenhadas no software R[3], assim como algumas considerações que podem ser feitas já pela análise visual.

A Figura 2.2 mostra o número de passageiros aéreos internacionais registrados pela Pan Am nos Estado Unidos no período de 1949 a 1960. Os dados estão disponíveis como uma série temporal no R e ilustram importantes conceitos que surgem na análise exploratória de séries temporais. É aparente que o número de passageiros está aumentando com o tempo. Em geral, as *tendências* são mudanças sistemáticas em séries temporais e que não parecem ser periódicas. A tendência desta série é um aumento linear, o modelo mais simples de tendência. Há uma clara *variação sazonal* nesta série onde valores mais altos ocorrem nos meses de Junho, Julho e Agosto, durante o verão do hemisfério norte, e valores mais baixos durante os meses de Novembro e Fevereiro. Além disso, a variação sazonal parece aumentar com a tendência. Em algumas séries pode haver *ciclos* que não correspondem períodos naturais fixos; exemplos podem incluir ciclos de negócio ou oscilações climáticas como o El Niño. Não há ciclos na série temporal de passageiros (COWPERTWAIT; METCALFE, 2009).

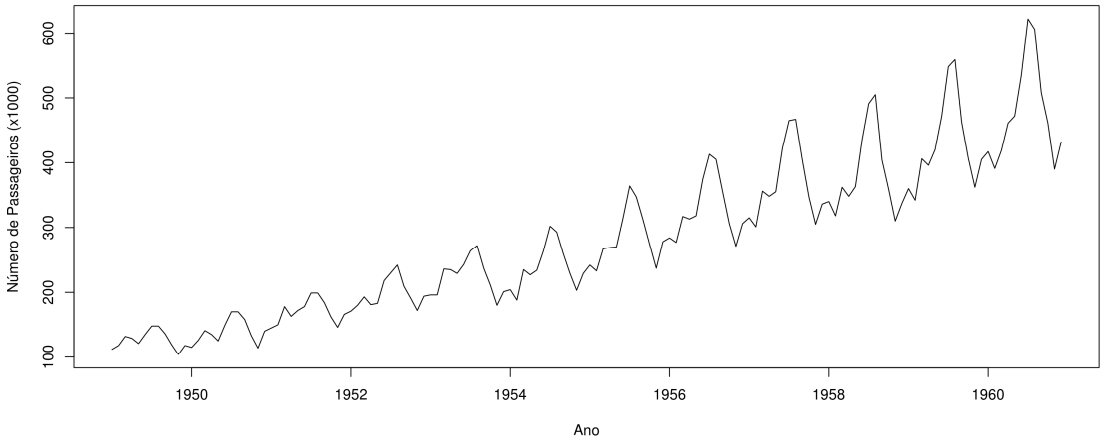

Figura 2.2: Número de passageiros aéreos internacionais nos Estados Unidos

[3] http://www.r-project.org/.

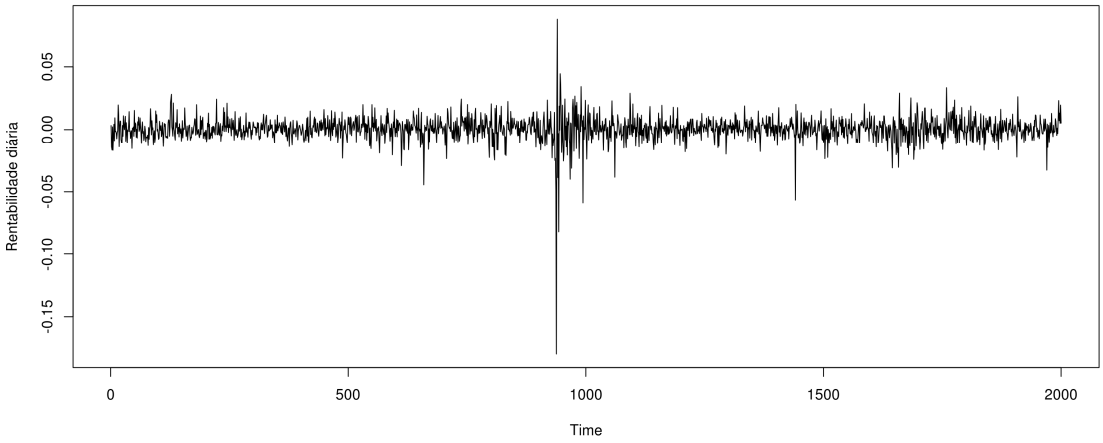

Figura 2.3: Rentabilidade diária da NYSE

Um exemplo de uma série temporal financeira, como a rentabilidade diária da NYSE (New York Stock Exchange, Bolsa de valores de NY), pode ser observado na Figura 2.3. Os dados mostram 2000 dias de transação, registrados no período de 2 de Fevereiro de 1984 a 31 de Dezembro e 1991. A grande queda da bolsa de 19 de Outubro de 1987, conhecida como Black Monday[4], pode ser vista em t = 938. A média da série aparece estável com uma taxa de rentabilidade aproximadamente zero, entretanto a volatilidade dos dados muda ao longo do tempo. Os dados mostram agrupamentos de volatilidade; isto é, altas volatilidades tendem a se agrupar. Um problema na análise desse tipo de dados financeiros é a previsão da volatilidade dos rendimentos no futuro (SHUMWAY; STOFFER, 2010).

Após a análise visual, é comum a aplicação de algoritmos para gerar modelos dos dados, utilizados para extrair mais informações da série ou para previsão de dados futuros. Na Figura 2.4 pode-se observar a utilização de dois modelos para previsão da série de passageiros da Figura 2.2. Ajustes como o do modelo Holt-Winters (HOLT; WINTERS apud COWPERTWAIT; METCALFE, 2009) são impressionantes, linha vermelha do primeiro gráfico da figura. O modelo pode ser usado para fazer previsões no futuro, como mostra a continuação dos dados após a linha vertical, sendo as linhas em azul a margem de erro do modelo. No segundo gráfico da Figura 2.4 está a previsão feita utilizando o modelo ARIMA (BROCKWELL; DAVIS, 1987), referência para previsão de séries temporais. Deve-se notar que as previsões são inteiramente baseadas nas tendências do período durante o qual o modelo foi ajustado, e seria uma previsão razoável assumir que esta tendência continue. Embora as previsões da Figura 2.4 pareçam visualmente adequadas, imprevistos podem levar a valores no futuro completamente diferentes dos apresentados no gráfico (COWPERTWAIT; METCALFE, 2009).

[4] http://en.wikipedia.org/wiki/Black_Monday_(1987)/.

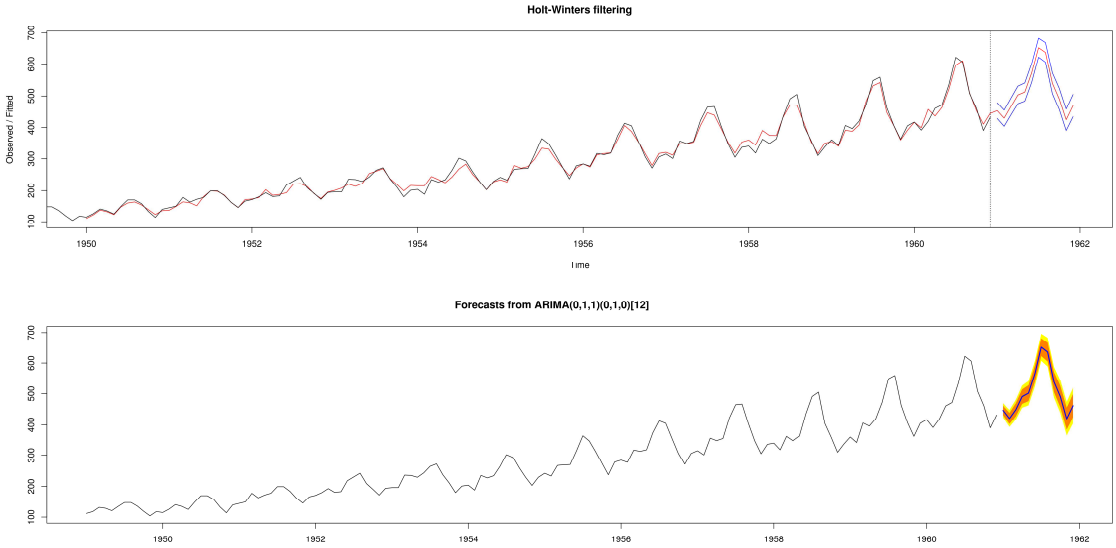

Figura 2.4: Previsão Holt-Winters e Arima aplicada a série de passageiros aéreos

A regressão linear e outros tipos de regressão são modelagens que analisam séries temporais no domínio do tempo. A regressão tem o objetivo de encontrar uma formula que possa prever cada ponto da série temporal, com precisão, a partir das entradas anteriores (DARLINGTON, 1990). O modelo (S)ARIMA (BROCKWELL; DAVIS, 1987) é exemplo de modelagem que utiliza como parte do processo a regressão. Esse modelo forma uma importante parte da abordagem Box-Jenkins (BOX; JENKINS; REINSEL, 1994) para modelagem de séries temporais. Essa sigla denota os termos para modelos autorregressivos (autoregressive, AR), integrados (integrated, I) e médias móveis (moving average, MA). A autorregressão procura explicar os valores atuais a partir dos valores passados. As médias móveis descrevem os dados calculando as médias de todos os subconjuntos consecutivos de tamanho *n* dos valores e a integração define qual ordem de diferenciação será aplicada aos dados para torná-los um ruído branco (COWPERTWAIT; METCALFE, 2009).

Usualmente o modelo é referenciado como ARIMA(*p*,*d*,*q*), onde *p*, *d* e *q* são inteiros não negativos que referenciam a ordem das partes de autorregressão, diferenciação (integração) e médias móveis. O (S) se aplica quando há uma componente sazonal na série que deve ser modelada também, nesse caso denotando como SARIMA(p, d, q)(P, D, Q), onde *P*, *D* e *Q* representam os mesmos parâmetros mas para o aspecto sazonal da série. O método consiste em quatro estágios: identificação, estimação, verificação de diagnóstico e previsão. Neste trabalho a abordagem de Box e Jenkins será referenciada genericamente como ARIMA.

Uma das formas de estimação dos parâmetros *p*, *d* e *q* é através da análise visual da série temporal em arranjos específicos, como a elaboração de um gráfico de diferenciação (SHUMWAY; STOFFER, 2010) para o parâmetro de diferenciação (*d*) e o correlograma (COWPERTWAIT; METCALFE, 2009), que pode ser usado para identificar séries temporais estacionárias para dados não sazonais que indicarão os parâmetros de *p* (autoregressão) e *q* (médias móveis). A escolha do melhor modelo deve ser feita minimizando o valor dos parâmetros *p, d* e *q* para o ajustamento da série, pois quanto maior os valores, maiores serão a quantidade de variáveis do modelo.

As redes neurais são grandes adversários dos métodos convencionais de modelagem de séries temporais. A rede neural IGMN (Incremental Gaussian Mixture Network) (ENGEL; HEINEN, 2011), por exemplo, cria e continuamente ajusta modelos

probabilísticos consistentes para todos os dados sequencialmente apresentados, após cada ponto ser apresentado e sem a necessidade de armazenar quaisquer dados do passado. Normalmente uma rede neural possui um processo de aprendizado dos dados que necessitam várias leituras sobre os dados para modelagem. O processo de aprendizagem da IGMN é guloso, ou seja, apenas uma leitura sobre os dados é necessária para se obter um modelo consistente.

Também utilizado para modelagem de séries temporais o STL (CLEVELAND et al., 1990) é um procedimento de filtragem para decomposição da série temporal em componentes de tendências, sazonalidade e resíduo. Consiste na aplicação sequencial da suavização de LOESS (LOcal regrESSion), permitindo uma análise dessas propriedades com cálculos rápidos mesmo para séries temporais longas e grande quantidade de suavização de tendência e sazonalidade.

O algoritmo STL tem uma alta eficiência (menor tempo de execução), embora o modelo gerado não seja tão aproximado aos dados reais quanto os modelos gerados pelo ARIMA. Apesar desse problema, o STL é uma ferramenta importante para decomposição da série e posterior visualização das componentes, em separado. A Figura 2.5 apresenta um exemplo de utilização do STL sobre a série de passageiros. A primeira célula apresenta a série temporal original, a segunda apresenta a sazonalidade identificada, a terceira mostra a tendência e a quarta demonstra o ruído resultante quando a sazonalidade e a tendência são subtraídas dos dados originais.

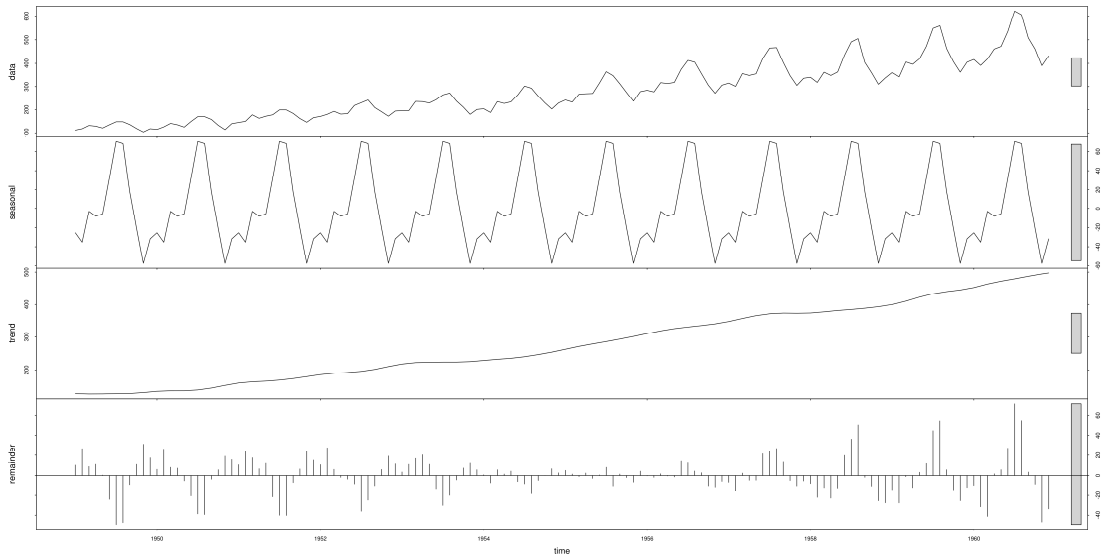

Figura 2.5: Algoritmo STL aplicado à série de passageiros

Como dito anteriormente, as regressões são frequentemente utilizadas para modelagem de séries temporais. Eventualmente alguns valores têm maior influência no resultado de uma regressão por serem muito diferente dos outros dados observados. Esse fenômeno, chamado de *outlier*, possui grande resíduo, positivo ou negativo, em relação ao modelo, e por esta razão recebe grande atenção quando ocorre (ROUSSEEUW; LEROY, 1987). *Outlier* é uma observação que diverge tanto das outras observações que desperta suspeita de que ela foi gerada por outro mecanismo (HAWKING apud BEN-GAL, 2005). Outra definição indica que uma observação distante, ou outlier, é aquela que aparenta desviar nitidamente dos outros membros da amostragem na qual ela ocorre (BARNETT; LEWIS apud BEN-GAL, 2005).

O princípio fundamental para detecção de outliers em séries temporal está em achar um modelo que melhor capture a autocorrelação existente entre os dados de uma série, usar este modelo para filtrar os dados e então aplicar métodos estatísticos tradicionais para detecção dos outliers nos resíduos obtidos, como se fossem dados univariáveis, ou seja, onde há uma só variável dependente (BEN-GAL, 2005). Um desses métodos tradicionais é o Boxplot (TURKEY apud BEN-GAL, 2005), baseado em quartis, dividindo os dados em quatro partes com o mesmo número de observações em cada uma. O primeiro e o terceiro quartil, Q_1 e Q_3, são usados para obter a medida robusta da média, (Q_1 + Q_3)/2, e o desvio padrão, Q_3 - Q_1. O método fornece também a mediana (segundo quartil Q_2), valor mínimo e valor máximo. Outliers neste método são os valores que forem maiores ou menores que 1,5 desvios padrões acima do Q_3 ou abaixo do Q_1, respectivamente.

A IGMN detecta outliers durante seu processo de modelagem dos dados. Utilizando recursos de autorregressão e agrupamento nos dados a detecção dos outliers ocorre levando em consideração a similaridade do dado com a sua vizinhança local. Um exemplo de detecção de outliers usando Boxplot, sobre o resíduo, e outro usando IGMN é exibido na Figura 2.6, pontos azuis são outliers detectados com valor acima do previsto pelo modelo e pontos vermelhos com valor abaixo do modelo. Este trabalho é baseado fortemente em análise de séries temporais e utiliza a rede IGMN e os outliers detectados por ela para identificar os eventos nos dados do Twitter. As justificativas são observadas no capítulo de Experimentos (Capítulo 5).

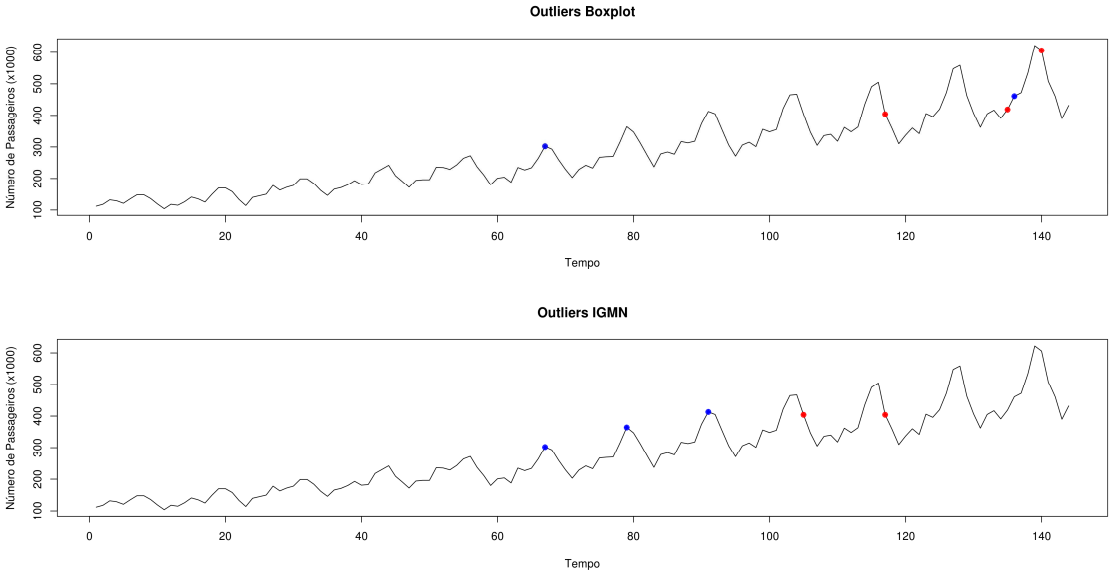

Figura 2.6: Outliers detectados utilizando Boxplot nos resíduos e IGMN na série de passageiros

2.5 Dados Geográficos

Em várias áreas do conhecimento há a necessidade de gerenciar dados geométricos, geográficos ou espaciais, ou seja, dados relacionados ao espaço. O espaço de interesse pode ser, por exemplo, a abstração bidimensional da superfície da Terra, um volume contendo o modelo do cérebro humano, ou um espaço tridimensional representando o arranjo das cadeias de moléculas de proteína (GÜTING, 1994). No contexto deste trabalho os dados espaciais gerenciados são referentes à abstração bidimensional da

superfície da Terra, representada por coordenadas de latitude e longitude ou pelo nome de regiões geográficas que nos levarão a estas coordenadas.

Para o armazenamento de dados espaciais utilizam-se sistemas de banco de dados espaciais. O que difere esses bancos dos bancos de dados relacionais, utilizado na maioria das aplicações, são tipos de dados específicos e operações para tratar e relacionar esses tipos, além dos tipos e operações tradicionais de um banco de dados relacional. Tipos de dados espaciais, ex.: ponto, linha, região, provêm abstrações fundamentais para a modelagem de estruturas de entidades geométricas no espaço assim como seus relacionamentos (*l* intersecciona *r*), propriedades (área(*r*) > 1000) e operações (intersecção(*l*,*r*) – a parte de *l* que se encontra com *r*) (GÜTING, 1994). Neste trabalho o uso de base de dados espaciais se dá para o armazenamento dos polígonos que representam os países e identificação, nesse banco, da origem das mensagens do Twitter utilizando coordenadas geográficas, latitude e longitude. Para isso foi utilizado o banco de dados PostGIS[5].

A motivação principal para uso de banco de dados espaciais são os Sistemas de Informações Geográficas (GIS – Geographic Information System). GIS prove um mecanismo conveniente para análise e visualização de dados geográficos. Dados geográficos são dados espaciais onde a estrutura básica de referência é a superfície da Terra. A riqueza de técnicas que são empacotadas em GIS é a razão por traz do seu crescimento fenomenal e suas aplicações multidisciplinares. Entre as funções mais comuns de um GIS podem ser citadas: busca de objetos geográficos ou descrições sobre eles, análise de localização, análise de terreno, análise de fluxo, distribuição, análise/estatística espacial e medições (SHEKHAR; CHAWLA, 2003). O uso de GIS para este trabalho se restringiu na visualização dos polígonos para utilização na representação dos países e visualização dos eventos identificados nos mapas, utilizando QuantumGIS[6] e Google Maps API[7].

2.6 Resumo do Capítulo

Neste capítulo foram apresentados os principais conceitos necessários para um melhor entendimento das pesquisas, desenvolvimentos e conclusões realizadas neste trabalho. A descoberta de conhecimento busca o melhor entendimento de dados que em razão do seu grande volume são de difícil análise. As redes sociais possuem este grande volume de informação e, consequentemente, muito conhecimento escondido. O Twitter, através da sua política, que por padrão torna pública as mensagens dos seus usuários, permite que seus dados possam ser coletados e analisados. A utilização dos dados georreferenciados do Twitter, separados por região geográfica, permite a criação de séries temporais que levam a identificação de eventos de causas sociais ou naturais.

[5] http://postgis.net/.
[6] http://www.qgis.org/.
[7] https://developers.google.com/maps/.

3 TRABALHOS RELACIONADOS

Neste capítulo serão apresentados os principais trabalhos na área de identificação de eventos, iniciando na seção 3.1 com a apresentação de trabalhos que definem o tópico de identificação de eventos e/ou utilizam dados textuais sem contexto geográfico para esta tarefa. Na seção 3.2 serão discutidos trabalhos que utilizam dados georreferenciados para identificação de eventos utilizando palavras-chave, e na seção 3.3, trabalhos que utilizam o volume de pessoas para essa mesma tarefa.

3.1 Identificação de Eventos

Detecção e rastreamento de eventos é um subconjunto de problemas de detecção e rastreamento de tópicos definidos por Allan et al. (1998; 2002), em uma iniciativa para investigar o estado da arte em encontrar e seguir novos eventos num fluxo de notícias. Esta problemática consiste em três tarefas: segmentar o fluxo de dados, identificar as notícias que são as primeiras a discutir sobre um determinado evento e dado um número pequeno de notícias sobre um evento, descobrir todas as notícias seguintes sobre este evento, no fluxo de notícias (ALLAN et al., 1998)

Com a grande quantidade de informações disponíveis on-line, a Internet é uma fértil fonte para este tipo de identificação de eventos e pesquisas de mineração na Web estão no campo de utilização de várias comunidades de pesquisa deste tema (KOSALA; BLOCKEEL, 2000). Nos últimos 10 anos, conteúdos gerados por usuários se tornaram dominantes em grande parte da Internet e uma Internet em tempo real surgiu para desafiar inúmeras áreas de pesquisa, particularmente recuperação de informação e mineração de dados na web (BERMINGHAM; SMEATON, 2010).

Becker, Naaman e Gravano (2011) apresentam a tarefa de identificação de eventos utilizando dados do Twitter baseado na abordagem de análise de texto e agrupamento, e mostram o número de características que devem ser consideradas para esta finalidade: temporal, social, temática e centrada no Twitter. Eles também analisam as diferentes características que podem impactar o desempenho de um sistema em tempo real para identificação de eventos. A técnica proposta para identificação de eventos oferece grande melhoria em relação a outras abordagens, mostrando que é possível identificar conteúdo de eventos do mundo real no grande volume de dado do Twitter. A utilização de dados baseados em localização na identificação de eventos é sugerida por eles como trabalho futuro.

No trabalho de Lanagan e Smeaton (2011) eles utilizaram dados do Twitter para identificar e atribuir palavras-chave a eventos esportivos e compararam o método com a identificação de eventos utilizando dados de áudio e vídeo. Analisando dois eventos esportivos eles puderam identificar os eventos com um alto grau de precisão e sucesso com uma complexidade computacional muito inferior a exigida pela análise de áudio e vídeo. Um problema dessa abordagem é a necessidade de escolher um conjunto de

palavras que representem a área de interesse, perdendo qualquer outro evento que não seja contemplado por estas palavras.

Nas seções a seguir serão analisados trabalhos que estudam a identificação de eventos utilizando como fonte de dados principalmente o Twitter. A definição de eventos utilizada por este trabalho é a de Allan, Papka e Lavrenko (1998), que diz: quando tópicos são colocados em um lugar no espaço e no tempo identificam-se eventos.

3.2 Identificação de Eventos por Palavras-Chave

Devido à característica tempo-real do Twitter, eventos naturais como terremotos podem ser identificados por mensagens de usuário antes mesmo de serem anunciados na mídia ou por órgãos responsáveis. Sakaki, Okazaki e Matsuo (2010) se propõem a utilizar os usuários do Twitter, e suas mensagens, como sensores para identificar tais eventos, levando em consideração o contexto geográfico no qual estas palavras-chave aparecem. Foram escolhidos somente usuário do Japão por ser uma região que possui grade intersecção entre usuário do Twitter e eventos de terremoto. A contribuição principal do estudo é a forma social de identificação de eventos naturais a partir do Twitter.

Para identificação do evento é realizada uma análise semântica das mensagens, filtrando somente mensagens que correspondem ao evento escolhido e relacionado ao presente momento. Cada usuário é um sensor que pode ou não identificar um evento e sua mensagem estará associada a uma posição no tempo e no espaço.

No método temporal de identificação é necessário verificar se o usuário está mencionando um evento atual ou um evento passado. Nesse aspecto o número de mensagens cresce exponencialmente já que com o passar do tempo mais pessoas falam sobre o evento. Por esta razão, falsos positivos devem ser desconsiderados do método quando o usuário estiver se referenciando a um evento do passado. Já na espacialidade pode-se considerar somente coordenadas espaciais, e, por exemplo, identificar o epicentro de um terremoto, ou soma-se também dados temporais, identificando a trajetória de um tufão, utilizando filtros de Kalman, regras Baysianas e filtros de partículas. O artigo identifica que é possível desenhar a arquitetura do alarme probabilisticamente usando o fator de falsos positivos 0,35 e levando em conta que os sensores são independentes e identicamente distribuídos (i.i.d.).

A escolha das palavras-chave para identificação de eventos como terremotos e tufões é uma questão complexa e decisiva, o que dificulta a criação de um sistema que identifique eventos de vários tipos. A dificuldade em saber se o usuário está realmente falando de um terremoto que esteja acontecendo no momento é minimizada pela quantidade de sensores existentes. Também se observa que a ocorrência desses eventos em áreas menos populosas ou com epicentro no mar dificulta a identificação da sua localidade. Nesse aspecto, os filtros de partículas foram os que apresentaram melhores resultados, tanto para epicentro quanto para trajetórias.

O sistema de aviso de terremotos foi desenvolvido por Sakaki, Okazaki e Matsuo (2010) para alertar os usuários com interesse em tais eventos. A precisão foi de 96% e os usuários recebiam as mensagens que poderiam demorar apenas 20s após a ocorrência do evento, ao contrário dos avisos de TV que poderiam demorar até 6min. A posição e trajetória dos eventos identificados em comparação aos fenômenos reais são exibidas na Figura 3.1.

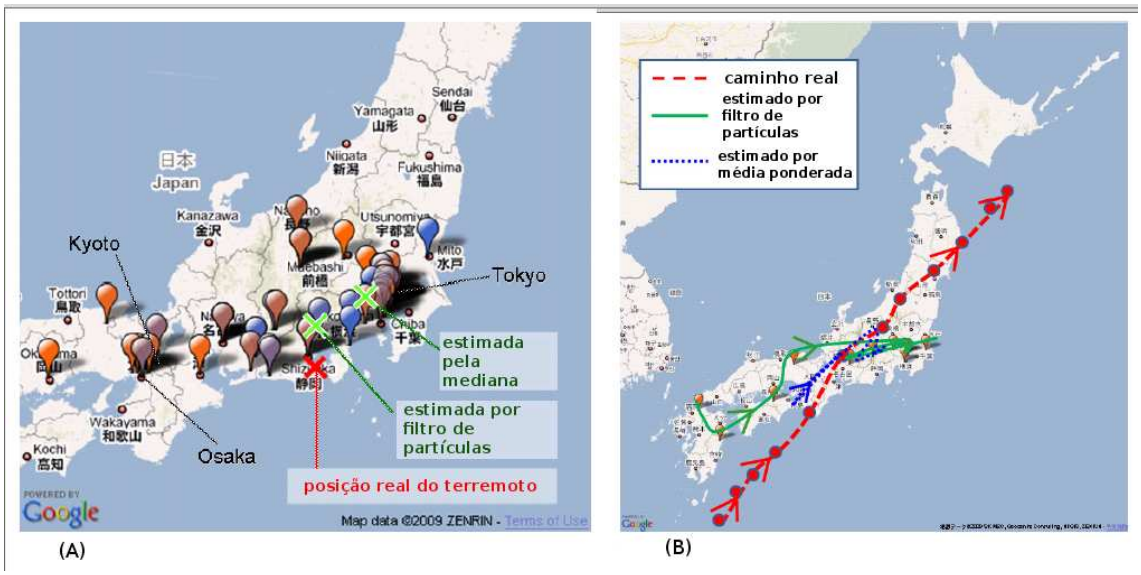

Figura 3.1: Precisão na identificação de eventos.

(A) Detecção de terremoto. (B) Detecção de tufão

Fonte: adaptada de (SAKAKI; OKAZAKI; MATSUO, 2010).

Num mesmo direcionamento, o estudo de Viewg et al. (2010) analisou as postagens feitas em um período que ocorreram dois desastres naturais para tentar identificar informações relevantes que possam servir em casos de emergência. Devido a característica ubíqua, de tempo real e multiplataforma, os microblogs podem ser considerados como meios de comunicação em situações de emergência. Também são vistos como o lugar para coletar informações durante eventos de crise e determinar com mais detalhes o que está acontecendo. Contudo, neste artigo, a identificação da geolocalização é feita por termos encontrados ao longo da mensagem, que se referem à estados/províncias, cidades, ruas, estradas e lugares. O objetivo do artigo foi entender como a informação é afetada por este tipo de evento e demonstrar que um sistema de Consciência Situacional (Situational Awareness, SA) pode ajudar pessoas que trabalham com resposta a emergências.

3.3 Identificação de Eventos por Volume de Pessoas

Num trabalho mais recente, Lee, Wakamiya e Sumiya (2011) apresentaram um sistema para descoberta de regiões com atividade social fora do padrão, utilizando informações do Twitter que tenham marcação geográfica. Este trabalho foi apresentado em outros quatro artigos que serão descritos individualmente a seguir.

No primeiro estudo (FUJISAKA; LEE; SUMIYA, 2010a) foram introduzidos os principais conceitos o que seria a proposta final da metodologia de análise para obtenção das regiões com atividades fora do esperado. O inicio do fluxo de informações começa com a coleta dos tweets através da Search API do Twitter e nesta parte foi necessário criar um algoritmo de divisão e conquista para resolver a restrição de 1500 resultados por consulta impostos na API quando se requisita os últimos tweets de uma determinada região geográfica. Caso a consulta retorna 1500 tweets, então se divide a região pesquisada em quatro áreas para se obter mais dados, até um limite que o raio da região seja maior que 2 km.

Cada mensagem tem uma coordenada geográfica associada. Assim, estes pontos no espaço, são agrupados através do algoritmo K-means formando um local a ser

analisado. Para cada local, analisam-se os comportamentos de agregação – quantas pessoas que estavam em outros locais estão agora no local *x* – e dispersão – quantas pessoas que estavam no local *x* agora estão em outros locais. Este primeiro artigo considera que um pico nos dados de agregação representa um evento social regional.

No segundo trabalho apresentado (FUJISAKA; LEE; SUMIYA, 2010b) os autores perceberam que analisando o conteúdo das mensagens que geraram o pico de agregação, seria possível entender quais as emoções e pensamentos do público, assim como quais eventos ocasionarem este acumulo de pessoas. Além disso, evoluíram as medidas de agregação/dispersão para o conceito de *Crowd Activity* (atividade da multidão), que leva em consideração o quanto um usuário se deslocou para chegar até o local atual, potencializando eventos nos quais as pessoas vieram de longe para prestigiar. No terceiro artigo (FUJISAKA; LEE; SUMIYA, 2010c) os locais formados pelo agrupamento de pontos é chamado de Region-of-Interest (RoI, Região de Interesse). Estes três trabalhos mencionados foram validados com 1 (um) exemplo de evento identificado pelo método, utilizando mensagens capturadas durante uma semana, na região do Japão e entorno.

O quarto estudo evolui consideravelmente os conceitos já bem descritos nos três primeiros trabalhos (LEE; SUMIYA, 2010). Três medidas são utilizadas para identificar um possível evento: número de tweets observados (#Tweets), número de pessoas (#Crowd) e movimentação da multidão (#MovCrowd). Três tipos de possíveis situações são associadas à combinação entre estas medidas:

- Festivais: aumento de #Tweets, #Crowd e #MovCrowd
- Festivais locais: aumento de #Tweets e #MovCrowd
- Feriados: aumentos de #Crowd e #MovCrowd

Assim um evento anormal é representado pela seguinte expressão:

(#Tweets OR #Crowd) AND #MovCrowd

Para detectar a ocorrência anormal dessas medidas é calculada a Regularidade Geográfica de cada uma, que corresponde ao comportamento normalmente observado em cada região. Depois, utiliza-se a técnica de Boxplot, que fornece também os pontos anormais (outlier). O método foi validado com 15 eventos, obtendo 60% de Recall e 1,8% de Precision Rate. Recall é a quantidade de eventos identificados que estavam em uma listagem pré-definida de eventos relevantes, enquanto Precision Rate indica quantos, de todos os eventos identificados, eram relevantes. No último trabalho do grupo (LEE; WAKAMIYA; SUMIYA, 2011), todos os conceitos são reapresentados, aumentando a quantidade de eventos na validação para 50 e modificando o nome das medidas para:

- #Tweets → Degree of Crowd Activity based on Tweets (DCA_T)
- #Crowd → Degree of Crowd Activity based on Crowd (DCA_C)
- #MovCrowd → Degree of Crowd Activity based on Moving Crowd (DCA_M)

Como trabalhos futuros, os autores sugerem a utilização de outras medidas, outros algoritmos de agrupamento, análise dos textos postados para entender as razões dos eventos e identificar fenômenos regionais adicionais, baseado na sua associação com os dados geográficos e temporais. Cabe ressaltar que a medida DCA_M tem complexidade exponencial e somado aos algoritmos de agrupamento podem inviabilizar a aplicação deste método em um ambiente de tempo real.

Para entender a relevância deste tipo de estudo na modelagem do comportamento das pessoas, em outro trabalho do mesmo grupo (WAKAMIYA; LEE; SUMIYA, 2011), as medidas descritas anteriormente foram utilizadas para caracterizar áreas urbanas, no Japão, baseado no padrão comportamental de seus habitantes ao longo do dia. Com base nestes padrões foram identificados quatro tipos de áreas urbanas:

- Cidades quarto: onde a população ativa se concentra antes do entardecer e diminui de atividade à noite;
- Cidades escritório: a população de concentra durante a tarde e se dispersa ao entardecer;
- Cidades de vida noturna: a atividade é calma até o entardecer e a partir daí a população começa a se concentrar; e
- Cidades multifuncionais: onde a população é ativa durante todo o dia.

O mesmo método de detecção de outliers, Boxplot, foi utilizado no trabalho de Ferrari, Mamei e Colonna (2012) para identificação de eventos. Eles utilizaram dados da principal empresa de telecomunicação da Itália para verificar se é possível identificar eventos na cidade analisando o uso das redes de celular. Apesar da identificação de eventos como partidas de futebol e promoções especiais em shoppings, eles puderam concluir que há um grande desafio em aberto para identificação de eventos que não atraiam um grande volume de pessoas ou eventos que atraiam muitas pessoas que não utilizam celulares. Estes mesmos vieses devem ser considerados na utilização dos dados do Twitter para identificação de eventos.

3.4 Resumo do Capítulo

Neste capítulo foram apresentados os principais trabalhos na área de identificação de eventos. Os principais pontos a serem ressaltados são:

- A definição de eventos por Allan, Papka e Lavrenko (1998), utilizada neste trabalho.
- O pioneirismo de Sakaki, Okazaki e Matsuo (2010) na utilização de dados georreferenciados do Twitter para identificação de eventos naturais através das mensagens dessa rede social;
- As medidas de Lee, Wakamiya e Sumiya (2011) para identificação de eventos sociais e os passos utilizados para concretizar esta identificação, no qual este trabalho se baseia amplamente.

4 MÉTODO PROPOSTO

O método de identificação de eventos utilizando dados de redes sociais desenvolvido neste trabalho é baseado nas definições do trabalho de Lee e Sumiya (2010). Foi o primeiro trabalho a utilizar o comportamento dos usuários de redes sociais para identificação de eventos, o qual pode ser separado em quatro etapas (ilustradas na Figura 4.1):

- Obter dados georreferenciados;
- Identificar regiões de interesse ou divisões políticas (bairros, cidades, estados, etc.);
- Medir o comportamento geográfico dos usuários dessas regiões;
- Detectar atividades incomuns ou medições fora do padrão.

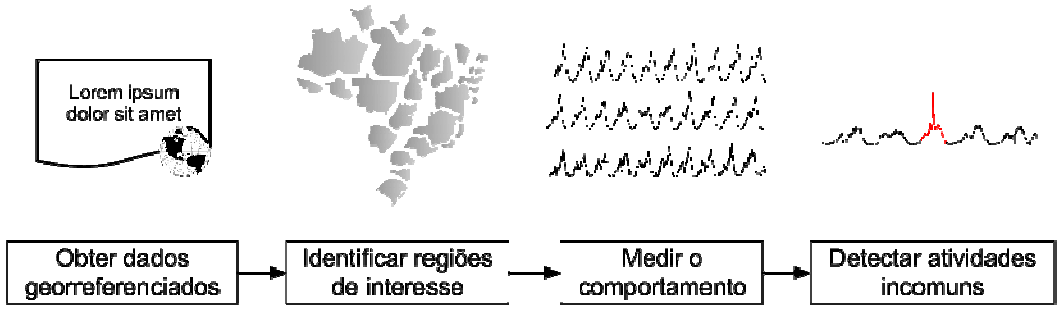

Figura 4.1: Modelo genérico para identificação de eventos

A obtenção de dados geográficos pode utilizar diversas mídias de informação como celular, GPS, Wi-Fi, etc. Em pequena escala pode-se considerar também a utilização de RFID (ZHAO; ZHANG; YE, 2006) e bluetooth (HAY; HARLE, 2009) para localização espacial. Porém o foco deste trabalho se restringe a informações coletadas em redes sociais, principalmente pelo fácil acesso a um grande volume de dados que estas proporcionam.

Na identificação das regiões de interesse devem-se determinar quais as delimitações espaciais que a área observada apresenta. Essas delimitações podem mudar dependendo do interesse da pesquisa. Em uma escala pequena, podem ser consideradas regiões de interesse as lojas de um shopping ou os leitos de um hospital. Já em áreas maiores, essas regiões podem ser bairros, cidades, estados, províncias ou países.

Uma vez delimitada a área de observação dos dados e subdividida em regiões de interesse, é necessário fazer medições de variáveis que possam representar o comportamento social dentro de cada região. Uma variável importante é a quantidade de pessoas presentes em cada região em um instante de tempo e a mudança deste dado ao

longo do tempo. Se possível, é desejável identificar quem são estas pessoas para futuramente observar suas movimentações ao longo da área observada entre as regiões de interesse.

Finalmente, utilizam-se os dados medidos ao longo do tempo para identificar padrões de ocupação das regiões de interesse, movimentação entre as regiões e outras informações dependentes das variáveis disponíveis. Para detectar as atividades incomuns utilizam-se algoritmos ou técnicas estatísticas. A característica dos dados é determinante para a escolha da técnica certa. Por exemplo, a utilização de Boxplot (LEE; SUMIYA, 2010) para detecção de fugas do padrão em uma variável com dependência temporal não é recomendada já que Boxplot é uma ferramenta estatística univariada (BEN-GAL, 2005).

Ao longo do desenvolvimento deste trabalho de pesquisa, várias tentativas para implementação deste método foram executadas para chegar à presente proposta. O volume de dados gerado pelo Twitter é muito grande e por isso uma abordagem com baixo tempo e custo de processamento precisa ser implementada para que o sistema torne-se escalável sem impactar seu desempenho. No estado atual do sistema é possível observar cinco fases principais para a identificação de eventos utilizando as mensagens geolocalizadas do Twitter, ilustrados na Figura 4.2:

- Coleta dos dados: um coletor captura os dados do Twitter utilizando o serviço Streaming API;
- Extração de medidas: duas séries temporais são criadas uma a partir do número de mensagens e outra com o número de usuários identificados em um instante de tempo;
- Criação do modelo: uma rede neural é utilizada para criar o modelo dos dados e a partir deste modelo detectar outliers destes dados;
- Identificação dos outliers: consiste em detectar os instantes de tempo que foram considerados como outliers em ambas as séries temporais ao mesmo tempo;
- Descrição dos eventos: através das mensagens contidas no instante de tempo, detectado como outliers, é possível avaliar e entender o evento que originou este outlier.

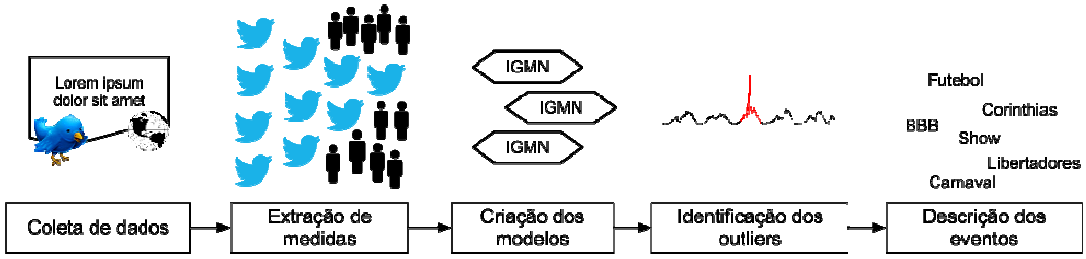

Figura 4.2: Método utilizado por este trabalho para identificação de eventos

Nas subseções a seguir cada uma destas etapas será explicada na sua respectiva ordem até a seção 4.4.

4.1 Captura dos Dados

Esta seção tem por objetivo mostrar detalhadamente o processo de captura dos dados do Twitter. Para isso, uma breve descrição da API do Twitter será apresentada na

subseção 4.1.1, seguindo da explicação de como as mensagens obtidas são marcadas geograficamente pelo serviço, na subseção 4.1.2. A subseção 4.1.3 descreve a implementação do coletor. A subseção 4.1.4 detalha como as mensagens que tem somente coordenada de latitude e longitude terão um nome geográfico atribuído a elas.

4.1.1 Twitter API

O Twitter possui, além de seus serviços para o usuário final, diversas API que permitem aos desenvolvedores de software criar suas próprias aplicações para utilização do serviço de microblog ou integrar funcionalidade do Twitter nas suas aplicações. A API do Twitter foi uma das principais alavancas do serviço, onde, nos primeiros anos de seu funcionamento, 91% dos tweets eram gerados através da API[8].

Esse ecossistema de desenvolvimento em torno do serviço é um dos motivos pelo qual o microblog se tornou popular. Ele é tão relevante que algumas funcionalidades incorporadas na raiz do serviço se originaram de aplicações de terceiros, como a utilização de @ na frente do nome do usuário para referenciá-lo em alguma mensagem, a sigla RT para simbolizar os tweets que estão copiando em seu texto uma mensagem já enviada para rede, as hashtags, palavras que contém o símbolo # na frente para representar o assunto principal da mensagem, entre outras. Um grande diferencial desta API é que, por filosofia da sua equipe, ela pode manipular quase todos os elementos e funcionalidades do microblog.

Para entender um pouco mais sobre os dados manipulados pela API do Twitter, é importante explicar quais são as entidades pertencentes a este ecossistema.

- Tweets: também conhecidos como Status Update, são a unidade básica de todas as coisas do Twitter. Usuários criam Tweets e eles podem ser embarcados em um site, respondidos, adicionados como favoritos, encaminhados (retweetados) e removidos;
- Usuários: podem ser pessoas ou softwares. Eles podem enviar mensagens (tweetar), seguir as mensagens de outros usuários (follow), criar listas de usuários, ter uma linha do tempo (timeline, ou seja, coleção de tweets de um usuário, ordenadas por data), ser mencionados em mensagens e/ou monitorados por uma aplicação;
- Entidades: são metadados e informações adicionais de contexto sobre conteúdos postados no Twitter;
- Lugares: são locais específicos, nomeados, os quais têm coordenadas geográficas correspondentes. Eles podem ser anexados a tweets por um place_id (identificador de local) específico quando a mensagem for enviada. Postagens associadas a lugares não necessariamente foram enviadas daquele local, mas potencialmente podem ser sobre aquele local. Lugares podem ser criados ou pesquisados. As mensagens podem ser localizadas através do place_id ao qual foi associado a elas.

Essas entidades são os principais recursos manipulados ou obtidos através da API. Os campos que descrevem estes elementos podem sofrer alteração ao longo do tempo, não têm garantia de ordem e não estão presentes em todos os casos. Assim, toda aplicação que fizer consultas deve tratar essas variações nos dados obtidos (no Apêndice A é possível visualizar mais informações sobre a API do Twitter).

[8] http://blog.programmableweb.com/2007/09/10/twitter-api-traffic-is-10x-twitters-site/.

São dois os principais grupos de serviços: REST API e Streaming API. O primeiro provê interfaces simples para a maioria das funcionalidades do Twitter, enquanto o segundo é uma poderosa família de API em tempo real para tweets e outros eventos do ecossistema. O presente trabalho utiliza para coleta de dados o recurso *statuses/filter* da Streaming API, que neste trabalho é acessado através do seguinte URI e parâmetros de POST abaixo:

URI: https://stream.twitter.com/1/statuses/filter.json
POST Data: locations=-179.99,-89.99,179.99,89.99

Figura 4.3: Dados utilizados para acessar a Streaming API do Twitter

O Post Data especifica um *Bounding Box*, que é o termo utilizado na documentação da API para descrever regiões geográficas filtradas pelo parâmetro *locations*, cada uma é definida por 4 números. Os dois primeiros representam o ponto sudoeste e os outros dois o ponto nordeste da delimitação. Várias *box* podem ser definidas no parâmetro *locations* para filtrar as mensagens com marcação geográfica.

O recurso acima foi disponibilizado na API em Agosto de 2009, porém a opção de enviar dados geolocalizados como marcação das mensagens foi incluída no serviço somente em Novembro de 2009. Naquela época não era possível filtrar o streaming por localização. A utilização do parâmetro *locations* foi habilitada somente em Janeiro de 2010 e em Agosto de 2012 a API possibilitou a utilização de um Bounding *Box* de 360º de longitude por 180º de latitude, permitindo uma cobertura global de todos tweets com marcação geográfica. Há uma restrição na Streaming API que limita o cliente a acessar somente 1% do total de mensagens enviadas para o serviço.

4.1.2 Mensagens com Marcação Geográfica

As mensagens com marcação geográfica podem ser enviadas para o serviço do Twitter desde Novembro de 2009 e a maneira como o usuário insere essa marcação geográfica depende da implementação do aplicativo cliente que o mesmo utiliza. Primeiramente, o usuário necessita habilitar a funcionalidade de tweetar com informação de localidade, já que essa função é desabilitada por padrão quando um usuário cria sua conta no serviço. Tal característica é também utilizada para fornecer conteúdo com maior semântica ao usuário, como exibir ao usuário tendências e mensagens específicas para sua atual localização.

Estando o serviço de localização habilitado, o usuário pode escolher qual a localização em que ele se encontra, ao enviar uma mensagem. Para isto é exibida uma lista dos locais mais próximos de onde o usuário está, sendo que essa lista fica limitada à base de lugares do Twitter. A localização do usuário pode ser definida por diversas maneiras:

- base de dados que associam IP à localização geográfica;
- coordenadas geográficas obtidas por triangulação da rede Wi-Fi em que o usuário está conectado (Google Gears e Skyhook Wireless);
- triangulação das antenas de celular em torno do usuário;
- GPS e/ou A-GPS;
- uma combinação dos recursos anteriores.

Ao utilizar o serviço do Twitter em dispositivos móveis, os aplicativos clientes permitem ao usuário escolher a granularidade da informação geográfica a ser enviada para o Twitter. Isso pode variar desde o nome do lugar (como bairro e cidade), que é a

configuração padrão, até a localização exata fornecida através de coordenadas geográficas de latitude e longitude, se esta informação estiver disponível pelo hardware do usuário. Quando a mensagem foi enviada com essa informação mais precisa, de latitude e longitude, ela é exibida com um *pin* ao lado do nome do local, nas interfaces de visualização. O usuário tem autonomia para remover as marcações geográficas de suas mensagens a qualquer momento do uso do serviço, através da interface Web do Twitter.

4.1.3 Coletor

O coletor é o primeiro componente da arquitetura, sendo responsável por:

- consumir todas as informações enviadas pela API;
- converter a informação não estruturada em estruturada;
- resolver problemas de codificação de caracteres;
- identificar duplicidade nos dados recebidos;
- adicionar informação de localidade nas mensagens (vide seção 4.1.4);
- registrar o andamento da coleta, taxa de mensagens por minuto;
- inserir as informações no banco de dados.

Na implementação inicial do coletor, apenas um componente de coleta era necessário para consumir os dados de todo o mundo, 360º de longitude por 180º de latitude. Contudo, com o passar do tempo, a utilização do Twitter assim como das mensagens geolocalizadas foram aumentando, passando a ser necessário dividir a tarefa em dois coletores. Um coletor fica então responsável por obter os dados vindos das Américas, maior região produtora de dados do Twitter, e o outro do resto do globo, Europa, África, Ásia e Oceania, fazendo com que a perda de dados, ocasionada pela limitação de 1%, seja minimizada. Através das mensagens de controle, enviadas pela API, é possível observar uma perda média de 2,31%, ocorrida principalmente no período entre 21h e 01 UTC no coletor das Américas. As perdas no coletor Euro Ásia são mínimas e se concentram no período entre 09h e 12h UTC.

Parte da implementação deste coletor foi realizada através da biblioteca Phirehose[9], responsável por implementar a conexão e o consumo do fluxo de dados. Dessa forma, este trabalho utiliza, para implementação do coletor, a linguagem PHP como uma aplicação de linha de comando, e não como interface de um servidor web.

Os campos armazenados pelo sistema de coleta são:

- created_at: data/hora UTC em que o tweet foi criado;
- id_str: representação textual do identificador único do tweet;
- text: o texto da mensagem (ou atualização de status);
- geo['coordinates'][0]: valor correspondente à latitude de onde o usuário estava no momento do envio da mensagem;
- geo['coordinates'][1]: idem ao item anterior, referente à longitude;
- user['id_str']: representação textual do identificador único do usuário;

[9] https://github.com/fennb/phirehose/.

- place['country']: nome do país que contém este lugar;
- place['place_type']: tipo de localização representada por este lugar, pode ter os valores admin, city, country, neighborhood e poi (ponto de interesse);
- place['full_name']: representação legível e completa do nome do lugar;
- user: bloco JSON que representa todas as informações do perfil do usuário, neste caso armazenadas de forma não estruturada (para uso futuro).

4.1.4 Enriquecimento de Nomes Geográficos

As marcações geográficas das mensagens podem ter dois tipos, não mutualmente excludentes: a descrição do lugar e a identificação numérica das coordenadas geográficas, latitude e longitude. Através de uma amostra de dados obtidos no período de Outubro e Novembro de 2010 pelo coletor, identificou-se que os usuários do Twitter produzem em torno de 4,1 milhões de mensagens, com marcação geográfica, por dia, onde 42,25% contêm coordenadas geográficas e 93,49% contêm o nome do lugar. Os 6,51% restante das mensagens, que só têm a informação de latitude e longitude, podem ser enriquecidas com o nome do lugar através da consulta dessas coordenadas em um banco de dados geográficos.

A Tabela 4.1 mostra em que proporção se encontra os tipos de lugares em mensagens com latitude e longitude. A descrição do lugar é obtida pelo campo place['full_name'] e o tipo de lugar se encontra no campo place['place_type']. Essa última informação especifica a precisão dos dados geográficos em relação ao local em que os tweets são criados. Cabe lembrar que, se um tweets possuir dados de latitude e longitude, é possível aumentar a granularidade da informação através de um banco de dados geográfico.

Tabela 4.1: Proporção de tipos de lugares observados

Tipo de Lugar	*Proporção*	*Conteúdo do campo full_place*
Cidade	73,43 %	<cidade>, <estado>
Sem descrição de lugar	11,25 %	
Unidade Administrativa	7,18 %	<unidade administrativa>, <país>
País	6,34 %	<país>
Ponto de Interesse	1,67 %	<ponto de interesse>, <cidade>
Bairro	0,14 %	<bairro>, <cidade>

Na Internet há diversos bancos de dados geográficos disponíveis gratuitamente com diversos níveis de granularidades de dados. Alguns deles possuem informação das divisões políticas de continentes, países, estados, províncias, microrregiões, unidades administrativas, cidades, bairros, até mesmo da malha viária de estradas, rodovias e ruas, ou banco de pontos de interesse que contém locais relevantes da cidade como estabelecimentos comerciais, praças, bancos, postos de gasolina, etc.

Dessa forma é possível enriquecer o nome do lugar de onde a mensagem foi enviada em diversos níveis. Neste trabalho foi utilizado um banco de dados geográfico que contém a divisão políticas dos países do mundo, com o propósito de atribuir o nome do país às mensagens que não têm essa informação, os 11,25% da segunda linha da Tabela 4.1. O referido banco[10] possui 211 registros, e para cada registro há informações de

[10] http://geocommons.com/overlays/85161/.

área, população (no ano de 2005), região e o nome do pais. Foi necessário ainda mapear o nome dos países deste banco para o mesmo utilizado pelo Twitter, a fim de aprimorar o desempenho das consultas por nome de país, como *Brazil* para *Brasil*, *Netherlands* para *The Netherlands*, entre outros.

Utilizando este processo de enriquecimento sobre as informações de países é possível observar um aumento de 11,97% nos dados de nomes geográficos, passando de 88,75% para 99,37% a quantidade de mensagens com nomes geográficos. Dos 11,25% das mensagens que não têm nenhuma informação referente ao nome do lugar, 94,44% foram identificadas com o nome do país. Ao analisar as mensagens restantes, que não têm nome do lugar mesmo após o processo, percebeu-se que muitas destas foram enviadas de regiões costeiras na qual há erro no banco de países, ou em áreas internacionais e, por este motivo, não tendo correspondência com banco de dados geográficos.

4.2 Medidas extraídas dos dados

Para geração das medidas, que serão utilizadas para criação das séries temporais, é necessária a escolha de alguns parâmetros:

- a região geográfica de onde essas medidas serão extraídas;
- o período no qual essas medidas serão aferidas;
- as variáveis utilizadas para geração dessas medidas;

Primeiro deve-se escolher uma região geográfica, Espanha, Paraná, Miami, e depois a granularidade temporal desejada, 1 dia, 1 hora, 15 minutos. A partir dessas informações, podem-se iniciar os agrupamentos das informações e a realização das medições. Utilizando os parâmetros anteriores é possível observar dois conjuntos de informações que representam o comportamento dos usuários, assim como observado por Lee e Sumiya (2010): a quantidade de mensagens e a quantidade de usuários. A primeira é mais simples de contabilizar, já que é uma contagem de quantas mensagens foram enviadas em um determinado período no tempo para a região escolhida. Na segunda, por outro lado, esta contagem é realizada sobre os usuários distintos que correspondem aos produtores daquelas mensagens, na mesma granularidade temporal e geográfica da primeira.

Há outras variáveis mais complexas que podem ser extraídas dos dados, porém, por terem um custo computacional muito alto estão fora do escopo deste trabalho. Um exemplo são as variáveis identificadas no trabalho de Lee, Wakamiya e Sumiya (2011), agregação e dispersão, ou seja, quantas pessoas estavam em outra região geográfica e agora estão na região escolhida para análise e quantas estavam na região escolhida e agora estão em outra região, respectivamente. Outro exemplo seria calcular um índice de estrangeiros na região escolhida a partir da média das distâncias dos indivíduos ali presentes em relação às suas localidades historicamente observadas, sugerida como trabalho futuro.

Para cada medida extraída uma série temporal é criada. Para cada região duas séries temporais são criadas a partir das medidas, uma que representa a variação da quantidade de mensagens ao longo do tempo e outra que representa a quantidade de usuários. Conforme os parágrafos anteriores, essas séries foram escolhidas por serem mais simples e mesmo assim representarem o comportamento dos usuários. Assim, tem-se:

- TS_m: Série temporal formada pela quantidade de mensagens. Cada valor representando a quantidade de mensagens enviadas para o Twitter, originadas da região geográfica especificada (*r*), durante o instante de tempo correspondente ao slot selecionado;
- TS_u: Série temporal formada pela quantidade de usuários. Cada valor representa a quantidade de usuários distintos que enviaram mensagem ao serviço, originadas da região geográfica especificada (*r*), durante o instante de tempo correspondente ao slot especificado.

A Figura 4.4 mostra exemplos de séries temporais formada pela quantidade de mensagens (TS_m) em três cidades brasileira. Através da análise visual dessas séries é possível observar um padrão de comportamento com sazonalidade diária, sendo que os menores valores da série ocorrem durante a madrugada (das 4h às 6h), havendo um aumento ao longo do dia e pico no fim do dia (entre as 20h e 2h). Este padrão se repete para as três cidades e nos mesmos instantes de tempo, já que estão sob um mesmo fuso horário. É possível perceber que o Rio de Janeiro possui maior volume de mensagens que Porto Alegre, e essa maior que Belo Horizonte. Um baixo valor de mensagens observado entre os dias 30 e 31 ocorreu pela falha do coletor por um problema de conectividade.

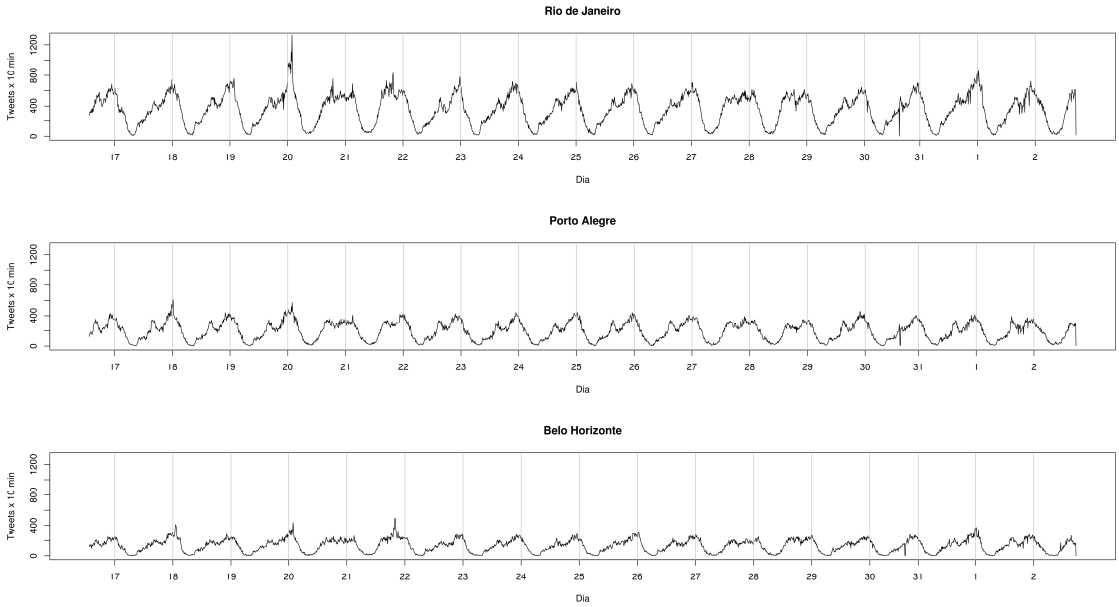

Figura 4.4: Séries Temporais formadas pela quantidade de mensagens nas cidades do Rio de Janeiro, Porto Alegre e Belo Horizonte

Dados de 16/10/2012 a 02/11/2012 em uma escala de 0 a 1200 mensagens para um slot de tempo com tamanho de 10 minutos.

4.3 Análise das Séries Temporais

Nesta seção serão expostos os métodos de análise de séries temporais, utilizadas e avaliadas neste trabalho para resolver o problema de detecção de outliers. Também se explica como ocorre a detecção dos outliers, etapa fundamental para o processo de identificação dos eventos.

4.3.1 Métodos

Como visto na seção 2.4, para fazer a modelagem de séries temporais existem diversos métodos, porém, neste trabalho três deles foram analisados a fim de verificar qual melhor se adapta ao problema de detecção de outliers, sendo eles: ARIMA, STL e a IGMN.

Para comparação dos três métodos de modelagem foi utilizado o tempo de processamento e a raiz quadrada do erro quadrático médio (RMSE, *root-mean-squared error*), uma medida frequentemente usada para avaliação de modelos de previsão, que mede a diferença entre os valores previstos pelo modelo e os valores realmente observados. Assim, o RMSE é aplicado ao ruído restante da subtração dos dados originais pelo modelo previsto.

O algoritmo STL possui um tempo de execução menor em comparação aos outros métodos, porém um erro médio quadrático mais elevado, criando um modelo pouco ajustado aos dados. Mais detalhes serão apresentados no capítulo de Experimentos (Capítulo 5). A IGMN quando comparado com o ARIMA tem um erro médio quadrático equivalente sem a necessidade de um conhecimento anterior dos componentes da série temporal tampouco da correlação dos dados imposta pelos parâmetros do ARIMA, isso facilita o processo de integração de novos locais a serem analisados pelo método. O processo incremental é outra vantagem em relação ao ARIMA, que precisa de um longo período da série temporal para fazer a modelagem. Desta forma é possível estender o método para uma análise em tempo real do fluxo do Twitter.

4.3.2 Extração de Outliers

Neste trabalho, a extração de outliers é peça fundamental para a identificação de eventos, já que para este estudo um evento é uma observação destoante do resto da série temporal, formada pelas medidas extraídas dos tweets, vindos de uma região do globo.

Cada uma das séries temporais, sobre a quantidade de mensagens e a quantidade de usuários, é processada pela IGMN a fim de modelar esses dados. Como entrada da rede utiliza-se uma série temporal de cada vez. Como saída deste processamento tem-se:

- O modelo gerado a partir da rede;
- Lista de pontos detectados como outliers superiores, aqueles que ficam muito acima do previsto pelo modelo. Ponto azuis na Figura 4.5; e
- Lista de pontos detectados como outliers inferiores, aqueles que ficam muito abaixo do previsto pelo modelo. Pontos vermelhos na Figura 4.5.

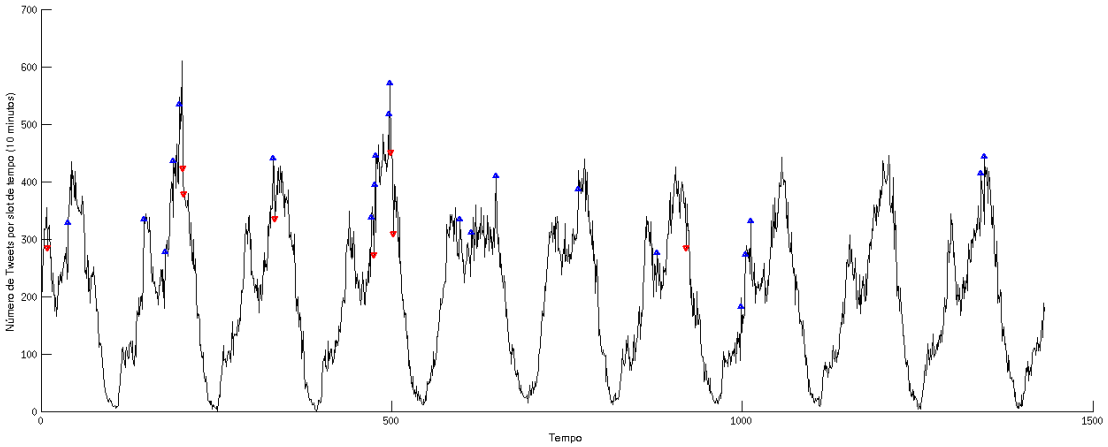

Figura 4.5: Exemplo de outliers superiores e inferiores detectados pela IGMN em uma série temporal de quantidade de mensagens

A lista dos outliers inferiores não é utilizada para este trabalho, pois indica baixa atividade em mensagens e usuário. Este trabalho se interessa nas altas atividades desses dados, assim, têm-se as duas listas de outliers superiores, uma de cada série temporal. Os eventos candidatos são os slots de tempo detectados como outliers em ambas as séries, significando que, em um slot de tempo, houve um número de pessoas tweetando muito acima do normal, assim como muitas mensagens foram enviadas. Está é uma importante decisão na lógica da detecção, embasada no fato de que ao considerar somente a quantidade de mensagens, o envio de um grande número de mensagens por um mesmo usuário seria considerado um evento, por exemplo.

A Figura 4.6 mostra exemplos de eventos identificados por essa abordagem, identificados ao longo de uma série temporal de mensagens (TS_m). Outras perturbações podem ser observadas na figura, porém será destacada somente uma de cada gráfico, para mostrar os diferentes padrões de comportamentos observados nos eventos escolhidos. Através da análise visual da figura, podem-se identificar os seguintes padrões de comportamento dos eventos (mas não únicos).

- Explosão em Oslo: grande perturbação na série temporal, com longa duração;
- Jogo de futebol em Munique: grande perturbação, com curta duração;
- Apuração do carnaval em São Paulo: pequena perturbação na série, com curta duração.

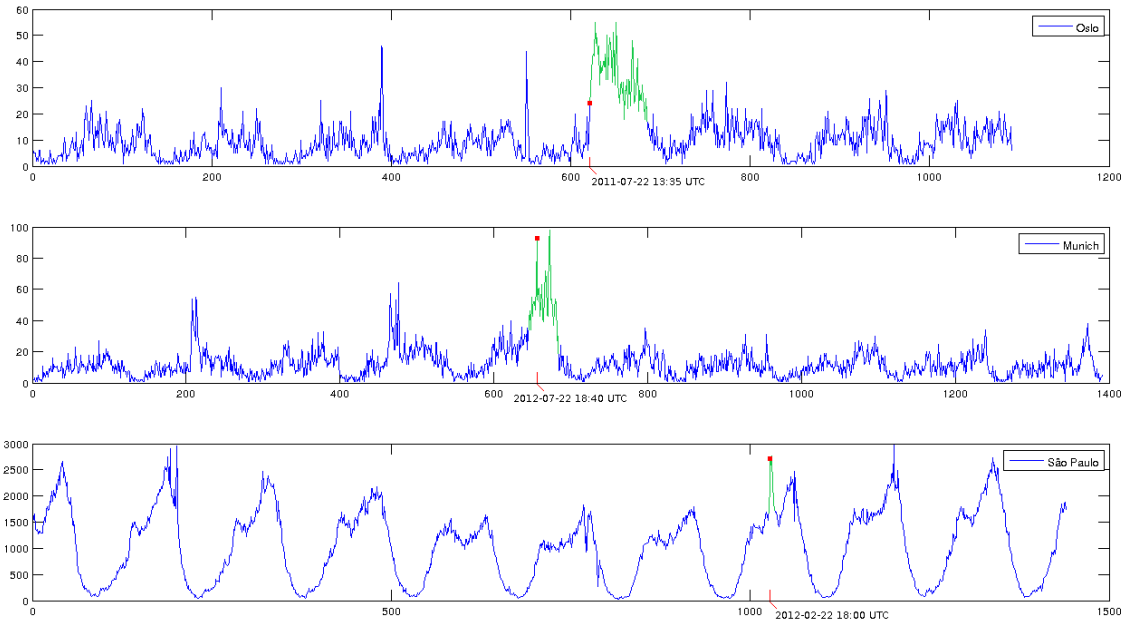

Figura 4.6: Eventos identificados em Oslo, Munique e São Paulo

Os eventos estão destacados em verde.

4.4 Descrição dos Eventos

A partir dos instantes de tempos detectados como outliers, tem-se o dia e a hora dos slots de tempo como eventos candidatos. No entanto, ainda é necessário descrever tais eventos utilizando como subsídio os tweets enviados para o serviço no slot de tempo em questão. Para aumentar a quantidade de texto e permitir uma melhor descrição do evento, é possível ampliar o período de tempo a ser analisado para além do slot de tempo, já que mensagens referentes ao evento podem ter sido enviadas antes e/ou depois do slot de tempo detectado.

Como primeiro passo obtém-se todos os textos enviados para o Twitter que pertençam a este intervalo de tempo e tenham sido enviados a partir da região analisada. Para a extração das palavras-chave mais frequentes é necessário antes remover caracteres especiais como “.”, “:”, “¬”, “/”, “#”, “%”, “&”, “@”, etc., palavras irrelevantes (ou stopwords) como artigos, pronomes, conjunções, preposições, interjeições, letras avulsas e palavras reservadas do escopo Internet/Twitter como rt (retweet), 4sq (foursquare) lol, http, etc. Eventualmente essa lista pode ser atualizada através da análise das palavras-chave extraídas em outros eventos, também irrelevantes para descrição dos eventos.

Através da lista de palavras-chave é possível calcular a frequência absoluta e frequência relativa de cada palavra para interpretar o quão relevante ela é dentro de todos os textos utilizados. Essa relevância é importante para identificar se a palavra realmente descreve o evento, pois mesmo quando ranqueadas por ordem de frequência absoluta as palavras do topo podem ter baixa frequência relativa, indicando uma dispersão nos textos dos diversos tweets.

Quando os textos convergem resultando em palavras-chave com alta frequência absoluta e relativa, há um indicativo de que os textos abordam um mesmo assunto. Para identificar se o assunto está relacionado a um evento, utilizam-se as palavras mais frequentes em ferramentas de busca de notícias na internet, utilizando como critério de filtragem adicional a data do evento. Quando o resultado é positivo, tem-se a

confirmação de que o evento identificado pela abordagem proposta é um evento do mundo real.

Somente esta confirmação não é suficiente para garantir eficácia (taxa de precisão) total da metodologia para este evento, já que é necessário ainda validar o quanto a filtragem geográfica dos tweets resultou em uma identificação de um evento local, evento ocorrido na região geográfica selecionada. Esta confirmação é possível caso o conjunto de notícias ou fontes de informação contenha dados referentes à localização do evento identificado. Na sessão de experimentos, essas verificações serão apresentadas com dados e análises estatísticas.

4.5 Resumo do Capítulo

Este capítulo apresentou os elementos utilizados na identificação de eventos. Os principais pontos a serem destacados no processo de detecção e identificação, em comparação aos métodos já existentes, são:

- Fazer uso da Streaming API ao invés da Search API, obtendo desta forma somente dados com marcação geográfica;
- Utilização de dados sem latitude e longitude, mas com nome da localizade;
- Uso de modelagem de séries temporais para detecção de outliers;
- Detalhamento de como ocorre a descrição de um evento.

5 AVALIAÇÃO

Este capítulo avalia o método proposto no capítulo anterior. Para tanto, são realizadas diversas baterias de experimentos, cada qual com um objetivo específico e descrita em uma subseção diferente. Para compreensão dos experimentos realizados e descritos a seguir, é necessário considerar as explanações realizadas no capítulo anterior sobre como os dados são arranjados. Em todos os experimentos, duas séries temporais, TS_m (mensagens) e TS_u (usuários distintos), são utilizadas para a identificação de eventos, sendo formadas pelas medidas calculadas para uma determinada região geográfica *r*, em slots de tempo *s* de tamanho *t* que são computados com uma periodicidade *p*.

5.1 Algoritmos de Séries Temporais

A primeira bateria de experimentos teve por objetivo avaliar os métodos de modelagem de séries temporais para que um deles fosse escolhido, dentro da arquitetura de identificação de eventos, na etapa de detecção de outliers. O algoritmo clássico para modelagem de séries temporais ARIMA foi o primeiro a ser colocado na lista. Em seguida, um mais simples para comparação de eficiência (tempo de execução), o STL, assim como um mais adaptativo, a IGMN.

O foco é identificar um algoritmo que tenha principalmente eficácia (taxa de precisão) na detecção de outliers e que seja adaptativo na modelagem (representação) de dados com diferentes características (de regiões e periodicidade). Um ponto importante a ser destacado é que a cada período de tempo *p* um novo slot *s* é adicionado ao conjunto de dados, sendo necessário realizar novamente a detecção de outliers sobre as séries TS_m e TS_u para verificar se um novo evento ocorreu no mundo real.

Como descrito na seção 2.4, uma das dificuldades da utilização do ARIMA é a escolha dos parâmetros *p*, *d* e *q*. Análises visuais e testes estatísticos devem ser realizados para avaliar qual valor de parâmetros leva a um modelo que minimize o erro quadrático médio. Supondo que uma vez identificados os parâmetros corretos esses possam ser utilizados para a série nos períodos seguintes de análise, a escolha dos parâmetros tem que ser feita novamente para as séries de outras regiões geográficas *r* escolhidas, por possuírem características distintas (Tabela 5.5). Isso dificulta a automatização dessa etapa, o que será demonstrado nos experimentos a seguir. De toda forma, o ARIMA é utilizado dentro dos experimentos para comparação de eficácia (detecção de outliers) com os demais algoritmos.

O método STL tem uma alta eficiência (tempo de execução), entretanto o erro quadrático médio é alto quando comparado com o ARIMA, ou seja, ele cria um modelo não muito aproximado aos dados reais. Apesar da baixa eficácia, o STL pode ser utilizado para analisar as componentes sazonais e de tendência das séries temporais obtidas com os dados do Twitter. Na Figura 5.1, por exemplo, é possível visualizar a

sazonalidade diária, na segunda célula com valores mais altos no meio e no final do dia que se repetem todos os dias, e a ausência de tendência nos dados, na terceira célula não há comportamento linear da tendência.

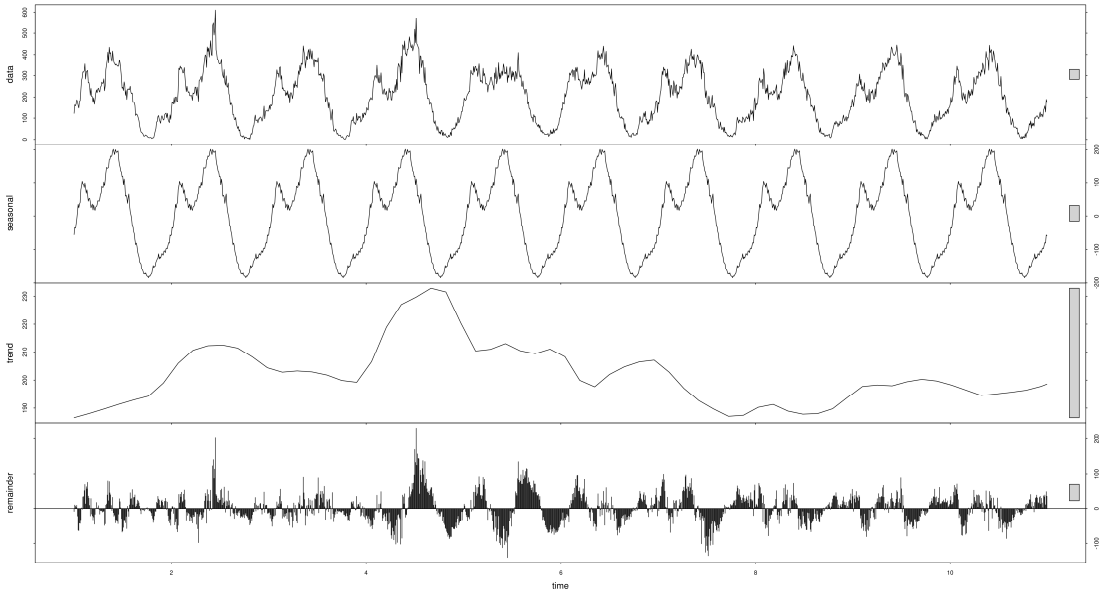

Figura 5.1: Método STL aplicado a série temporal de quantidade de mensagens

A rede neural IGMN apresentou uma eficácia (RMSE) equivalente ao ARIMA e uma eficiência (tempo) pior que o ARIMA quando utilizada para treinar as séries temporais. Porém, uma das vantagens dessas redes é a possibilidade de incluir slots novos na modelagem da série sem a necessidade de recalculá-la novamente. Dessa forma, a utilização em um ambiente online se torna propícia, pois diminui o tempo de execução do algoritmo para o processamento de novos slots. Outro ponto importante a favor da IGMN é que o calculo de outliers é realizado junto com a criação do modelo, levando em conta a similaridade dos dados em relação à sua vizinhança local.

Abaixo são apresentados os resultados da comparação entre os três algoritmos de modelagem de séries temporais quando aplicados a TS_m de cinco regiões geográficas (*r*), variando o tamanho do slot de tempo (*t*). Para computação do modelo ARIMA foi utilizada a função *auto.arima*[11] do software R, que procura o melhor modelo ARIMA para os dados variando os parâmetros *p*, *d* e *q*. Duas computações dessa função são realizadas, uma procurando pelo melhor modelo variando os parâmetros *p* e *q* de 0 a 5, nomeado como aa5, e outra variando os mesmos parâmetros de 0 a 10, com o nome de aa10. O primeiro experimento (A) utiliza um slot com o tamanho de 10 minutos e aplica os algoritmos em diferentes regiões geográficas.

A

Tabela 5.1 mostra a descrição dos dados utilizados nas comparações do experimento A, as cidades listadas foram escolhidas por terem diferentes características no volume de tweets e por serem de regiões reográficas diferentes. Na

Tabela 5.2 são comparados os valores da raiz quadrada do erro quadrático médio (RMSE) dos algoritmos, assim como o tempo de execução na criação do modelo. Não é parte do objetivo desses experimentos fazer uma comparação minuciosa da eficiência

[11] http://cran.r-project.org/web/packages/forecast/forecast.pdf

(tempo de execução) entre os algoritmos, o tempo de execução é utilizado apenas para uma comparação do comportamento dos algoritmos em diferentes situações.

Tabela 5.1: Descrição dos dados para o experimento A, slot de 10 minutos.

Cidade	*Quantidade de slots*	*Total de Tweets*	*Média de Tweets por slot de tempo (10 minutos)*
São Paulo	5471	5471882	1000,16
São Paulo (2)	2464	2561950	1039,75
Rio de Janeiro	2464	858045	348,23
Porto Alegre	2462	488816	198,54
Belo Horizonte	2446	323916	132,43
Nova York	2430	284401	117,04

Tabela 5.2: RMSE e tempo de execução dos diferentes algoritmos no experimento A, slot de 10 minutos. Menores RMSE em negrito.

	RMSE				*Tempo (s)*			
Cidade	*aa5*	*aa10*	*STL*	*IGMN*	*aa5*	*aa10*	*STL*	*IGMN*
São Paulo	90,453	**90,060**	179,409	90,255	5,218	243,975	0,011	26,841
São Paulo (2)	**102,074**	103,232	195,078	103,644	2,489	122,580	0,037	10,428
Rio de Janeiro	**41,717**	42,378	79,622	42,872	2,092	120,093	0,025	11,296
Porto Alegre	**27,340**	27,779	51,182	34,988	2,428	107,512	0,056	11,752
Belo Horizonte	22,518	**21,828**	50,146	26,738	1,774	109,534	0,033	9,763
Nova York	18,042	**17,681**	26,293	17,813	2,632	83,264	0,036	14,192
Média	**50,357**	50,493	96,955	52,718	2,772	131,160	0,033	14,016

No experimento B as mesmas análises foram realizadas, mas utilizando 20 minutos para o tamanho do slot de tempo, com o objetivo de verificar o comportamento dos algoritmos variando o parâmetro *t*. A descrição dos dados e o resultado são mostrados na Tabela 5.3 e Tabela 5.4, respectivamente.

Tabela 5.3: Descrição dos dados para o experimento B, slot de 20 minutos

Cidade	*Quantidade de slots*	*Total Tweets*	*Média de Tweets por slot de tempo (20 minutos)*
São Paulo	2736	5471882	1999,96
São Paulo (2)	1232	2560500	2078,33
Rio de Janeiro	1233	858045	695,90
Porto Alegre	1232	488816	396,77
Belo Horizonte	1225	323916	264,42

Nova York	1216	284401	233,88

Tabela 5.4: RMSE e tempo de execução dos diferentes algoritmos no experimento B, slot de 20 minutos. Menores RMSE em negrito

	RMSE				*Tempo (s)*			
Cidade	*aa5*	*aa10*	*STL*	*IGMN*	*aa5*	*aa10*	*STL*	*IGMN*
São Paulo	196,606	**193,258**	356,878	196,828	2,678	121,096	0,050	11,058
São Paulo (2)	224,470	**219,425**	387,309	226,968	1,743	88,277	0,015	4,765
Rio de Janeiro	84,486	**83,352**	155,074	142,010	1,915	67,552	0,017	5,575
Porto Alegre	53,827	53,618	96,231	**52,038**	2,033	83,763	0,015	7,051
Belo Horizonte	42,780	**42,063**	94,889	42,803	2,076	72,756	0,018	7,026
Nova York	33,603	32,450	48,822	**32,362**	23,462	60,494	0,013	6,427
Média	105,962	104,028	189,867	115,501	5,651	82,323	0,021	6,984

O objetivo desses experimentos era confirmar que a utilização da IGMN traz resultados tão bons quanto à do ARIMA. A utilização da IGMN é motivada principalmente pelo seu potencial em criar um modelo de forma incremental, pela forma de calculo dos outliers e por se adaptar melhor a dinamicidade dos dados com diferentes características. Outra confirmação dos experimentos é a de que os modelos ARIMA gerados para cada uma das séries temporais têm parâmetros *p*, *d* e *q* diferentes entre si, necessitando calculá-los sempre que novos dados forem modelados, apresentados na Tabela 5.5.

Tabela 5.5: Melhor combinação dos parâmetros *p, d* e *q* segundo a função auto.arima

	Slot de 10 minutos		*Slot de 20 minutos*	
Cidade	*aa5*	*aa10*	*aa5*	*aa10*
São Paulo	(4,0,1)	(4,0,9)	(4,0,1)	(9,0,1)
São Paulo (2)	(4,0,1)	(4,0,10)	(2,0,1)	(10,0,3)
Rio de Janeiro	(2,0,2)	(3,0,10)	(3,0,1)	(4,0,6)
Porto Alegre	(4,0,1)	(3,0,8)	(3,0,1)	(8,0,5)
Belo Horizonte	(1,0,1)	(8,0,4)	(4,0,1)	(9,0,3)
Nova York	(1,1,3)	(10,1,5)	(2,1,3)	(6,1,3)

5.2 Comparação de variáveis

Nesta seção será apresentada a influência de algumas variáveis na detecção de outliers, sendo elas: tamanho do slot de tempo, região geográfica (e consequentemente a diferença no volume de dados produzidos entre estas) e forma do cálculo de outliers.

Um dos parâmetros que merecem uma investigação maior é o tamanho do slot de tempo. Como será apresentado nesta seção, ele influencia a volatilidade da série temporal assim como a quantidade de outliers detectados. O primeiro experimento

realizado mostra o comportamento da detecção de outliers em regiões geográficas com diferentes volumes de tweets, variando o tamanho do slot de tempo.

A Figura 5.2 mostra um gráfico apresentando regiões geográficas com diferentes volumes de dados gerados. O número médio de tweets, considerando um slot de tempo de 1 minuto, é exibido ao lado do nome de cada região. O gráfico é desenhado em escala logarítmica para facilitar a comparação dos dados, que são muito altos no início e muito baixos no fim, quando se varia os valores do tamanho do slot de tempo. O slot de tempo é variado no seu tamanho conforme os valores apresentados na figura. Foi utilizado a IGMN para detecção de outliers É possível observar que nos slots com tamanho de 5 e 10 minutos há uma proximidade na quantidade de outliers detectados nas diversas regiões, indicando tamanhos de slot que possam ser usado para detecção de outliers independente do volume de dados da região.

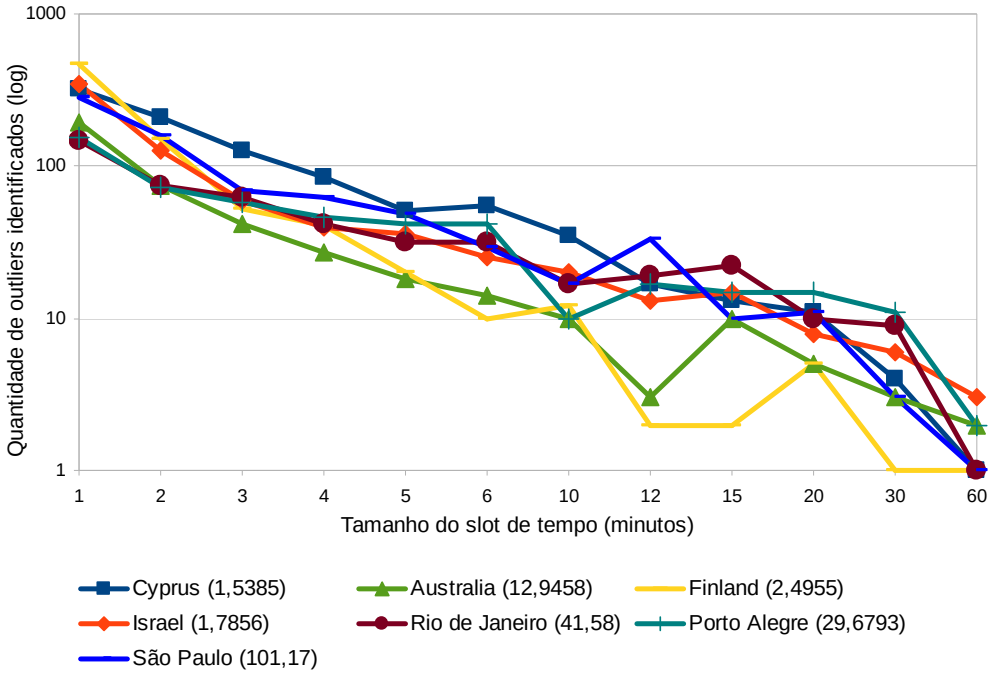

Figura 5.2: Comparação entre quantidade de outliers detectados, variando o tamanho do slot de tempo em diferentes regiões geográficas

A Figura 5.3 mostra os outliers variando o tamanho dos slots entre 1, 5 e 10 minutos. A dificuldade está em decidir em um slot de tempo pequeno, que gera uma série temporal mais detalhada, dados com uma volatilidade maior e consequentemente mais outliers serão detectados; ou um slot de tempo maior, que gera uma série temporal mais suavizada e menos outliers detectados. As implicações desta variação na eficácia de identificação de eventos são apresentadas na seção 5.3.

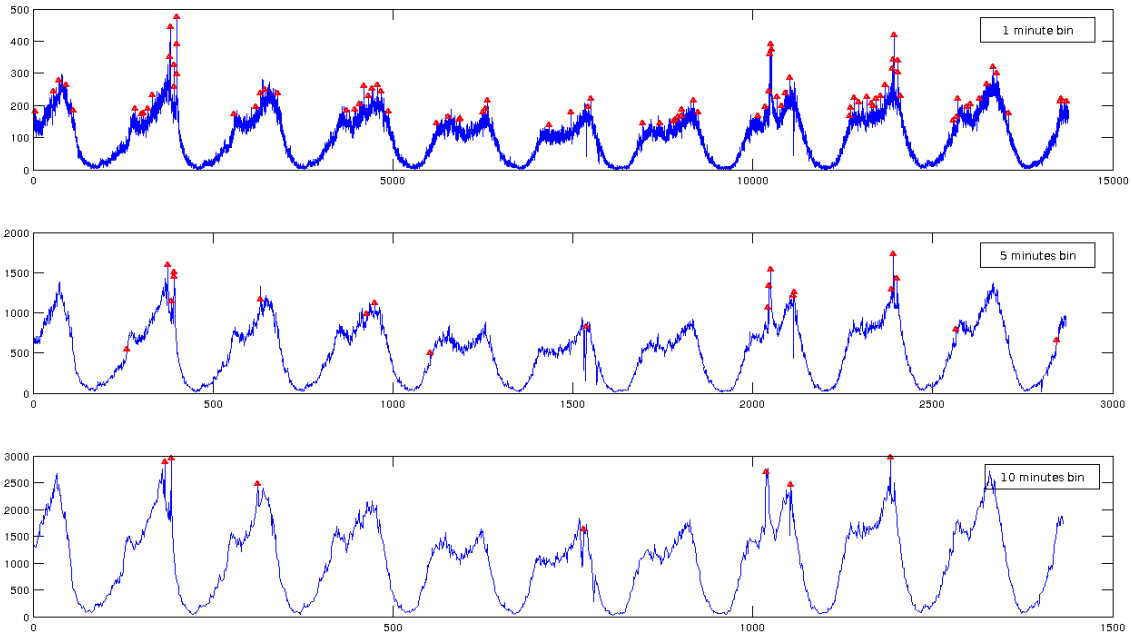

Figura 5.3: Detecção de outliers variando o tamanho do slot de tempo

A comparação entre os métodos de detecção de outliers é exibida na Figura 5.4. Os eixos, escalas e legendas tem a mesma característica da Figura 5.2. O objetivo desta figura é mostrar que o método Boxplot, para detecção de outliers, resulta em todos os casos um maior número de outliers detectados em relação à IGMN, confirmando a Hipótese 2 de que a utilização de séries temporais resulta em melhora na eficácia de detecção de outliers. Isso ocorre pois o Boxplot considera que todos os dados têm a mesma distribuição, já a IGMN utiliza a similaridade local da região na série temporal em que os valores se encontram para detectar outliers. Neste experimento o Boxplot foi aplicado assim como no trabalho de Lee, Wakamiya e Sumiya (2011), ou seja, diretamente sobre os valores da série, agrupados de acordo com a sua ocorrência no dia, e não sobre o resíduo da modelagem. (Mais informações dos dados utilizados para criar a Figura 5.2 e Figura 5.4 podem ser encontradas no Apêndice C)

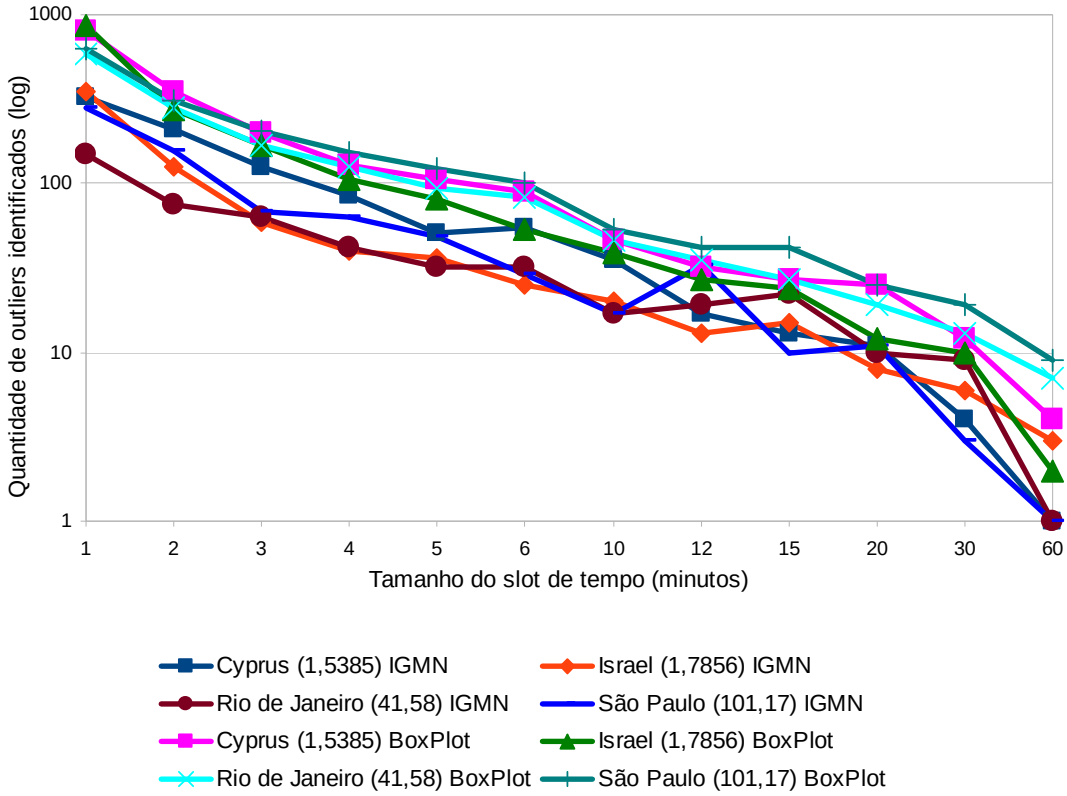

Figura 5.4: Comparação entre métodos de detecção de outliers, variando o tamanho do slot de tempo em diferentes regiões geográficas

5.3 Experimentos com IGMN

Foi decido a utilização da IGMN, pois essa possui melhor adaptabilidade na modelagem dos dados, mesmo tendo um RMSE maior que o ARIMA. Escolhendo a IGMN como o algoritmo para modelar as séries temporais dos dados coletados, não será mais utilizado o RMSE para avaliar o melhor modelo gerado. O objeto de interesse a partir deste ponto é a identificação de quais parâmetros do modelo melhor identificam os eventos e para isso será medida a precisão de acerto desses modelos.

Para avaliar a precisão do método com diferentes parâmetros de dados e do algoritmo, foi escolhida a cidade de São Paulo (Brasil) como região geográfica (divisão política). Essa escolha foi feita por esta cidade ser a que mais produz tweets com informação geográfica no mundo. O período de 10 a 24 de Fevereiro de 2012 foi escolhido para estes testes por possuir eventos conhecidos no mundo real, como o carnaval e campeonatos de futebol.

A qualidade dos outliers detectados é medida através da avaliação dos mesmos contra eventos ocorridos, únicos, duplicados e perdidos. Eventos ocorridos são eventos que aconteceram no mundo real e que foram avaliados utilizando as palavras mais frequentes das mensagens pertencentes ao slot de tempo *s* detectado como outlier, posteriormente comparado com as notícias no jornal Folha de São Paulo, usando como filtro a data e hora do slot de tempo *s*. Dependendo dos parâmetros escolhidos, um evento é identificado mais de uma vez. Desta forma denota-se que o evento é único e suas múltiplas identificações são consideradas duplicadas. Eventos identificados com um determinado conjunto de valores de parâmetros e não com outros conjuntos, são denotados como perdidos.

A primeira avaliação se preocupa em verificar quanto o tamanho (*t*) do slot influencia na identificação de eventos. A Tabela 5.6 exibe o acerto do método quando esses outliers são comparados com eventos do mundo real e os gráficos destes dados são apresentados na Figura 5.3. À medida que o tamanho do slot aumenta, há uma suavização da série fazendo com que as perturbações se diluam com a vizinhança, desta forma detectando outliers mais distintos do restante dos dados. Quando a precisão aumenta, diminui a identificação de eventos duplicados, porém alguns eventos que não ocasionaram muita perturbação na série não são identificados.

Tabela 5.6: Taxa de precisão em diferentes tamanhos de slot de tempo

Tamanho do Slot	*Total de Outliers*	*Eventos ocorridos identificados*	*Eventos únicos*	*Detecção duplicada*	*Eventos perdidos*	*Taxa de precisão*
1 minuto	90	22	6	16	0	24,44%
5 minutos	20	12	4	8	2	60,00%
10 minutos	7	5	3	2	3	71,43%

A IGMN ajusta seus modelos para os dados apresentados utilizando técnicas de agrupamento. A similaridade entre as entradas é medida pela probabilidade de cada entrada pertencer aos agrupamentos já existentes na rede. Para isso, o desvio padrão é utilizado indicando quando um novo agrupamento deve ser criado caso a entrada seja muito diferente dos agrupamentos existentes. Esse mesmo parâmetro é utilizado para detectar se uma determinada entrada é um outlier, baseado na sua similaridade local.

O desvio padrão foi avaliado com um slot de tempo com o tamanho de 1 minuto e diferentes valores de desvio padrão. O valor padrão da rede para esse parâmetro é de 3. Foram testados também os valores de 4 e 5 para este parâmetro com o objetivo de verificar os diferentes resultados. O número de outliers diminui à medida que o valor de desvio padrão aumenta (Figura 5.5), mas a mudança na taxa de precisão não evolui como no experimento anterior (Tabela 5.7). A variação no desvio padrão também se mostrou fundamental na eficácia do método de Ferrari, Mamei e Colonna (2012) para minimizar a identificação de falsos positivos e falsos negativos.

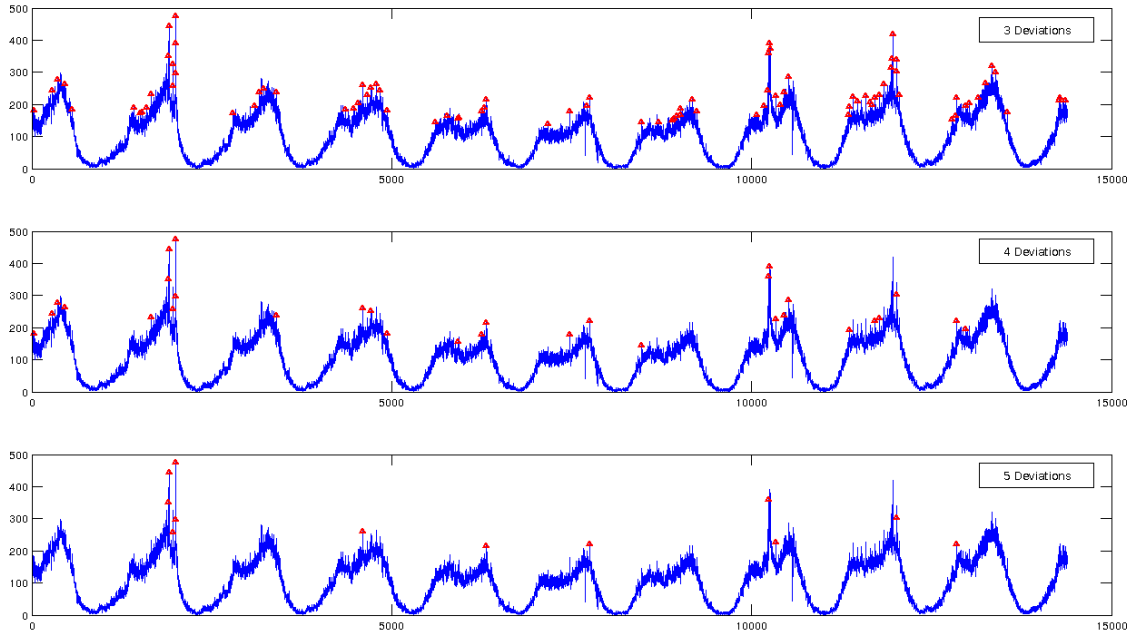

Figura 5.5: Detecção de outliers variando o parâmetro de desvio padrão

Tabela 5.7: Taxa de precisão em diferentes desvios padrão

Desvio padrão	*Total de Outliers*	*Eventos ocorridos identificados*	*Eventos únicos*	*Detecção duplicada*	*Eventos perdidos*	*Taxa de precisão*
3	90	22	6	16	0	24,44%
4	31	11	5	6	1	35,48%
5	12	8	3	5	3	66,67%

A primeira verificação é que o slot com tamanho de 1 minuto torna a série temporal grosseira e sensível a pequenas perturbações. Nessas condições o ajuste do parâmetro de desvio padrão é incapaz de chegar a resultados como o do experimento anterior. Por outro lado, somente aumentar o tamanho do slot de tempo irá causar atraso em uma abordagem que utilize esta metodologia num ambiente de tempo real, assim como a perda de eventos de pequena perturbação.

Na tarefa de validação dos outliers com eventos do mundo real, a utilização das palavras-chave mais frequentes permite entender quais tipos de tópicos mobilizam os usuários do Twitter a postar mais mensagens que o comum, assim como incentivar outros usuários não tão ativos. É importante entender que aspectos culturais podem influenciar a utilização dos serviços de redes social e os experimentos apresentados consideram somente o comportamento social da cidade de São Paulo. Todos os eventos ocorridos e identificados pelo método proposto tiveram grande cobertura televisiva, alguns com abrangência local e outros com abrangência global (Tabela 5.8). É fundamental entender que o processo de identificar os melhores valores para os parâmetros apresentados depende fortemente da lista dos eventos que realmente ocorreram no mundo real assim como da definição de evento, como verificado também por Ferrari, Mamei e Colonna (2012).

Os eventos identificados e não ocorridos no mundo real, falsos positivos, são, em 81% dos casos, momentos do dia em que o número que representa a hora coincide com

o que representa o minuto, por exemplo: 19:19, 22:22, 13:13. Outros casos são eventos virtuais como: promoção de uma rádio para quem mandar mensagens com o nome da música que gosta e a hashtag (#) com o nome da rádio, ou anúncio de uma nova música de uma banda ou músico popular (Justin Bieber, Rebeldes, One Direction).

Tabela 5.8: Alguns eventos identificados pela abordagem proposta

Descrição do evento	*Termos mais frequentes*	*Abrangência*
Partida de futebol da Copa Libertadores, na Venezuela	Corinthians, jogo, libertadores, gol, timão	Global
Programa de TV nacional, *reality show*	Yuri, fael, bbb, lider, ganhar	Global
Partida de futebol do campeonato estadual, fora da cidade	Corinthians, willian, douglas, gol, jogo	Global
Tumultos na apuração dos votos do carnaval	Gaviões, carnaval, nota, fogo, apuração, escola	Local
Duas partidas de futebol do campeonato estadual, fora da cidade	Gol, jogo, bragantino, time, corinthians	Global
Partida de futebol do campeonato estadual, na cidade	Ganhar, vergonha, deus, palmeiras	Local

5.4 Resumo do Capítulo

Neste capítulo de experimentos foi possível avaliar o comportamento das diversas variáveis envolvidas na identificação de eventos: tamanho do slot de tempo, perfil da região geográfica quando ao volume de mensagens, quantidade de desvios padrão utilizado e o método para detecção de outliers. Confirmou-se a Hipótese 1 de que é possível identificar eventos utilizando dados do Twitter. Também foi apresentada a eficácia do método proposto em identificar eventos, assim como a dificuldade em identificar eventos que realmente ocorreram na região monitorada. Isso sugere que novas investigações sejam realizadas a fim de especializar a identificação de eventos com relevância somente local.

6 CONCLUSÕES

Este trabalho apresentou um novo método para a descoberta de eventos baseado em localização, utilizando o fluxo de dados do Twitter e realizado através de análise de séries temporais. Esta abordagem pode levar a eventos representativos sem a necessidade prévia da escolha de palavras-chave, nem da utilização de algoritmos de agrupamento para seleção de regiões geográficas. A proposta provê os primeiros passos em uma série de métodos para melhorar a identificação de eventos com relevância local.

O trabalho confirmou as hipóteses apresentadas de:

- Detectar eventos do mundo real (fora da Internet) utilizando os dados do Twitter;
- Utilizar de dados que possuem somente o nome das localidades para aumentar a eficácia (taxa de precisão) na identificação de eventos;
- Utilizar análise de séries temporais para aumentar a eficácia da identificação de eventos se comparada com trabalhos que utilizem outros métodos;
- Usar a IGMN para permitir o processamento das informações de modo incremental.

Uma das limitações encontradas por este trabalho foi a de identificar com maior relevância eventos que aconteceram de fato na região geográfica analisada. Isso ocorre, pois em eventos com cobertura televisiva o público que interage com o Twitter não está necessariamente presente no evento ou na região do evento. Entretanto esta limitação pode não ser um problema, pois depende da necessidade de cada usuário na utilização de um sistema de identificação de eventos.

A definição de evento para realização dos experimentos, assim como a escolha do método para realizar a validação desses eventos são também elementos limitantes deste trabalho. A definição de o que é um evento pode mudar de acordo com a interpretação de cada pesquisador. Assim como a lista de eventos utilizada para validação pode influenciar fortemente o resultado da taxa de precisão.

Outro grande problema na identificação de eventos pelo Twitter é a característica desta rede. A maioria das mensagens trocadas pelos usuários trata sobre conversas ou assuntos do cotidiano. Isso atrapalha bastante a modelagem das séries temporais em casos como, por exemplo, no meio dia, quando as pessoas saem para almoçar, há um grande volume de mensagens sobre onde vão comer, informando a hora, demonstrando felicidade por estarem saindo do trabalho ou da escola, etc. Por este motivo uma nova etapa no método pode ser incluída com a remoção de diversos tipos de mensagens, baseado em palavras-chave, para deixar a detecção mais eficaz.

Embora este trabalho se utilize de dados de um microblog, o método pode ser aplicado a outras redes sociais ou sistemas que possuam dados semelhantes, dados

geográficos e grande volume de informação. Um exemplo de plataforma que poderia ser utilizada para obtenção deste tipo de dados são os Smartphone com sistema operacional Android. Estes dados já são utilizados, por exemplo, para popular os mapas que informam o tráfico nas ruas de diversas cidades do mundo (Google Maps).

Muitas questões precisam ser estudadas para evoluir o tratamento deste tipo de informação, pois os dados geográficos em redes sociais são recentes. Outra questão é a pouca quantidade de trabalhos com outros métodos de detecção e identificação de eventos, o que dificulta a evolução da descoberta de conhecimento neste tipo de dado. Percebe-se que este tipo de técnica agrega muito valor às redes sociais e, por esse motivo, muitos avanços estão escondidos dentro das grandes empresas. À medida que estas utilizações vão se tornando visíveis e populares, o meio acadêmico se apropria e propõe melhorias e novos paradigmas.

6.1 Trabalhos Futuros

Através deste trabalho foi possível perceber que os dados de redes sociais podem conter informações relevantes para a modelagem do comportamento das pessoas. A investigação dos dados geográficos nestas redes pôde despertar melhorias no método atual ou novas perspectivas de utilização destes dados como:

- Avaliar como a identificação de eventos se comporta nos períodos de vale na série temporal, em qual período do dia a identificação é mais freqüente e como os parâmetros de slot de tempo e desvio padrão influenciam na identificação de eventos em diferentes cidades.
- Após identificar um evento, agrupar as mensagens do slot de tempo para identificar qual ponto da região geográfica selecionada possui maior densidade de usuários, e investigar se aquele é o local do evento identificado;
- Para melhorar a identificação de eventos com relevância somente local, remover da lista de eventos detectados os que tenham sido detectados em um escopo mais abrangente. Por exemplo, um evento que tenha sido identificado no Brasil e em São Paulo ao mesmo tempo e possua palavras-chave semelhantes pode indicar um evento global;
- Criar uma nova medida que indique o quão próximo os usuários de uma região estão ao longo do dia para observar se quando esta medida tiver um aumento significativo pode indicar eventos com outras características;
- Criar outra medida que calcule a taxa de estrangeiros que ocupam uma determinada região. Para criação desta taxa antes teria que ser calculada a região mais frequente de cada usuário, sua região de moradia, e depois para cada região geográfica calcular o quão distante estão os usuários daquela região em relação a sua região de moradia. Uma alta taxa de estrangeiros em uma região pode indicar eventos com interesse global, como as Olimpíadas, shows internacionais, etc.

REFERÊNCIAS

ADRIAANS, P.; ZANTINGE, D. **Data mining**. [S.l.]: Addison-Wesley-Longman, 1997. I-XI, 1-158p.

ALLAN, J. **Topic Detection and Tracking**: event-based information organization. [S.l.]: Kluwer Academic Publishers, 2002. (Kluwer International Series on Information Retrieval).

ALLAN, J. et al. Topic Detection and Tracking Pilot Study Final Report. In: IN PROCEEDINGS OF THE DARPA BROADCAST NEWS TRANSCRIPTION AND UNDERSTANDING WORKSHOP. **Anais. . .** [S.l.: s.n.], 1998. p.194–218.

ALLAN, J.; PAPKA, R.; LAVRENKO, V. On-line new event detection and tracking. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21. **Proceedings. . .** [S.l.: s.n.], 1998. p.37–45.

BECKER, H.; NAAMAN, M.; GRAVANO, L. Beyond Trending Topics: real-world event identification on twitter. In: ICWSM. **Anais. . .** The AAAI Press, 2011.

BEN-GAL, I. Outlier detection. **Data Mining and Knowledge Discovery Handbook**, [S.l.], p.131–146, 2005.

BERMINGHAM, A.; SMEATON, A. F. Crowdsourced real-world sensing: sentiment analysis and the real-time web. In: AICS 2010 - SENTIMENT ANALYSIS WORKSHOP AT ARTIFICIAL INTELLIGENCE AND COGNITIVE SCIENCE. **Anais. . .** [S.l.: s.n.],2010.

BOX, G.; JENKINS, G.; REINSEL, G. **Time series analysis**: forecasting and control. [S.l.]: Prentice Hall, 1994. (Forecasting and Control Series).

BROCKWELL, P.; DAVIS, R. **Time series**: theory and methods. [S.l.]: Springer-Verlag,1987. (Springer series in statistics).

CLEVELAND, R. B. et al. STL: a seasonal-trend decomposition procedure based on loess (with discussion). **Journal of Official Statistics**, [S.l.], v.6, p.3–73, 1990.

COWPERTWAIT, P.; METCALFE, A. **Introductory Time Series with R**. [S.l.]: Springer-Verlag New York, 2009. (Use R).

CURBERA, F. et al. Unraveling the Web services web: an introduction to soap, wsdl, and uddi. **Internet Computing, IEEE**, [S.l.], v.6, n.2, p.86 –93, march-april 2002.

DARLINGTON, R. **Regression and Linear Models**. [S.l.]: McGraw-Hill, 1990. (McGraw-Hill series in psychology).

ELDER IV, J. F.; PREGIBON, D. A statistical perspective on knowledge discovery in databases. In: FAYYAD, U. M. et al. (Ed.). **Advances in knowledge discovery and data mining**. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. p.83–113.

ENGEL, P.; HEINEN, M. Incremental Learning of Multivariate Gaussian Mixture Models. In: ROCHA COSTA, A. da; VICARI, R.; TONIDANDEL, F. (Ed.). **Advances in Artificial Intelligence – SBIA 2010**. [S.l.]: Springer Berlin / Heidelberg, 2011. p.82–91. (Lecture Notes in Computer Science, v.6404). 10.1007/978-3-642-16138-4$_9$.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: an overview. In: **Advances in Knowledge Discovery and Data Mining**. [S.l.]: American Association for Artificial Intelligence, 1996. p.1–34.

FERRARI, L.; MAMEI, M.; COLONNA, M. People get together on special events: discovering happenings in the city via cell network analysis. **Pervasive Computing and Communications Workshops, IEEE International Conference on**, Los Alamitos, CA, USA, v.0, p.223–228, 2012.

FIELDING, R. T. **Architectural Styles and the Design of Network-based Software Architectures**. 2000. Tese (Doutorado em Ciência da Computação) — UNIVERSITY OF CALIFORNIA, IRVINE.

FUJISAKA, T.; LEE, R.; SUMIYA, K. Monitoring Geo-social Activities through Microblogging Sites. In: YOSHIKAWA, M. et al. (Ed.). **Database Systems for Advanced Applications**. [S.l.]: Springer Berlin / Heidelberg, 2010. p.374–384. (Lecture Notes in Computer Science, v.6193). 10.1007/978-3-642-14589-6$_3$8.

FUJISAKA, T.; LEE, R.; SUMIYA, K. Detection of Unusually Crowded Places through Micro-Blogging Sites. In: ADVANCED INFORMATION NETWORKING AND APPLICATIONS WORKSHOPS (WAINA), 2010 IEEE 24TH INTERNATIONAL CONFERENCE ON. **Anais. . .** [S.l.: s.n.], 2010. p.467 –472.

FUJISAKA, T.; LEE, R.; SUMIYA, K. Exploring urban characteristics using movement history of mass mobile microbloggers. In: ELEVENTH WORKSHOP ON MOBILE COMPUTING SYSTEMS & APPLICATIONS, New York, NY, USA. **Proceedings. . .** ACM, 2010. p.13–18. (HotMobile '10).

GARTON, L.; HAYTHORNTHWAITE, C.; WELLMAN, B. Studying Online Social Networks. **Journal of Computer-Mediated Communication**, [S.l.], v.3, n.1, p.0–0, 1997.

GÜTING, R. H. An introduction to spatial database systems. **The VLDB Journal**, [S.l.], v.3, p.357–399, 1994.

HAND, D. J. Data Mining: statistics and more? **The American Statistician**, [S.l.], v.52, n.2, p.112–118, 1998.

HAY, S.; HARLE, R. Bluetooth Tracking without Discoverability. In: CHOUDHURY, T. et al. (Ed.). **Location and Context Awareness**. [S.l.]: Springer Berlin / Heidelberg, 2009. p.120–137. (Lecture Notes in Computer Science, v.5561). 10.1007/978-3-642-01721-6$_8$.

HUBERMAN, B.; ROMERO, D.; WU, F. Social networks that matter: twitter under the microscope. **Available at SSRN 1313405**, [S.l.], 2008.

JAVA, A. et al. Why we twitter: understanding microblogging usage and communities. In: WEBKDD AND 1ST SNA-KDD 2007 WORKSHOP ON WEB MINING AND SOCIAL NETWORK ANALYSIS, 9., New York, NY, USA. **Proceedings. . .** ACM, 2007. p.56–65. (WebKDD/SNA-KDD '07).

KOSALA, R.; BLOCKEEL, H. Web mining research: a survey. **SIGKDD Explor. Newsl**., New York, NY, USA, v.2, n.1, p.1–15, June 2000.

KWAK, H. et al. What is Twitter, a social network or a news media? In: WORLD WIDE WEB, 19., New York, NY, USA. **Proceedings. . .** ACM, 2010. p.591–600. (WWW ’10).

LANAGAN, J.; SMEATON, A. F. Using Twitter to Detect and Tag Important Events in Sports Media. In: ICWSM. **Anais. . .** The AAAI Press, 2011.

LEE, R.; SUMIYA, K. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: ACM SIGSPATIAL INTERNATIONAL WORKSHOP ON LOCATION BASED SOCIAL NETWORKS, 2., New York, NY, USA. **Proceedings. . .** ACM, 2010. p.1–10. (LBSN ’10).

LEE, R.; WAKAMIYA, S.; SUMIYA, K. Discovery of unusual regional social activities using geo-tagged microblogs. **World Wide Web**, [S.l.], v.14, p.321–349, 2011.10.1007/s11280-011-0120-x.

MILLER, H. J.; HAN, J. **Geographic Data Mining and Knowledge Discovery, Second Edition**. [S.l.]: Taylor & Francis Groups, 2009. (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series).

MISLOVE, A. et al. Measurement and analysis of online social networks. In: ACM SIGCOMM CONFERENCE ON INTERNET MEASUREMENT, 7., New York, NY, USA. **Proceedings. . .** ACM, 2007. p.29–42. (IMC ’07).

NEWMAN, M. E. J. The Structure and Function of Complex Networks. **SIAM Review**, [S.l.], v.45, n.2, p.167, 2003.

ROUSSEEUW, P. J.; LEROY, A. M. **Robust Regression and Outlier Detection**. [S.l.]: John Wiley & Sons, 1987. (Wiley series in probability and mathematical statistics: Applied probability and statistics).

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In: WORLD WIDE WEB, 19., New York, NY, USA. **Proceedings. . .** ACM, 2010. p.851–860. (WWW ’10).

SHEKHAR, S.; CHAWLA, S. Spatial databases: a tour. **Upper Saddle River, New Jersey**, [S.l.], v.7458, 2003.

SHUMWAY, R. H.; STOFFER, D. S. **Time Series Analysis and Its Applications**: with r examples. [S.l.]: Springer, 2010. (Springer Texts in Statistics).

TULACH, J. **Practical API Design**: confessions of a java framework architect. [S.l.]: Apress, 2012. (Apressus Series).

VIEWEG, S. et al. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, New York, NY, USA. **Proceedings. . .** ACM, 2010. p.1079–1088. (CHI ’10).

WAKAMIYA, S.; LEE, R.; SUMIYA, K. Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In: ACM SIGSPATIAL INTERNATIONAL WORKSHOP ON LOCATION-BASED SOCIAL NETWORKS, 3., New York, NY, USA. **Proceedings. . .** ACM, 2011. p.77–84. (LBSN ’11).

WELLMAN, B. et al. Computer Networks as Social Networks: collaborative work, telework, and virtual community. **Annual Review of Sociology**, [S.l.], v.22, p.pp. 213–238,1996.

YE, M.; YIN, P.; LEE, W.-C. Location recommendation for location-based social networks. In: SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, 18., New York, NY, USA. **Proceedings. . .** ACM, 2010. p.458–461. (GIS ’10).

ZHAO, J.; ZHANG, Y.; YE, M. Research on the Received Signal Strength Indication Location Algorithm for RFID System. In: COMMUNICATIONS AND INFORMATION TECHNOLOGIES, 2006. ISCIT ’06. INTERNATIONAL SYMPOSIUM ON. **Anais. . .** [S.l.: s.n.], 2006. p.881 –885.

APÊNDICE A - DETALHES DA API DO TWITTER

A resposta da API pode estar em dois formatos XML e JSON, sendo que o segundo consome menor banda, por ser mais compacto, e dependem de cada função da API suportá-los ou não. São dois os principais grupos de serviços: REST API e Streaming API. O primeiro provê interfaces simples para a maioria das funcionalidades do Twitter, enquanto o segundo é uma poderosa família de API em tempo real para tweets e outros eventos do ecossistema.

A Streaming API possui um número menor de recursos porém com um volume de informação que pode ser muito grande, dependendo do filtro de pesquisa utilizado na consulta. São eles:

- Public streams: Três tipos de fluxo de todos os tweets públicos que são submetidos ao Twitter.
    - Sample: retorna uma amostragem aleatória das mensagens públicas, aproximadamente 1% do total de postagens públicas;
    - Filter: condiciona os resultados à tweets que pertençam aos critérios de busca, que podem ser por ids de usuários, palavras ou região geográfica (utilizando coordenadas geográficas), também restrita à um máximo de 1% de todos os tweets públicos;
    - Firehose: retorna todos os tweets públicos. Poucas aplicações têm permissão para este nível de acesso.
- User streams: Fluxo de todas as informações relacionadas às visões de um determinado usuário;
- Site streams: Conjunto de user streams de diversos usuários, para serem utilizados por um serviço que tenha vários usuários conectados ao mesmo tempo.

Após o início da conexão com o recurso Public da Streaming API, quatro tipos de informação podem ser recebidas através do fluxo de dados, obrigatoriamente no formato JSON:

- Limit: quantidade de tweets que, desde o início da conexão foram omitidos por terem ultrapassado a cota de 1% da taxa total de tweets. No exemplo abaixo 1234 tweets não foram enviados, pela Twitter API, para processo que recebeu esta mensagem;

```
{
   "limit":{
      "track":1234
    }
```

```
}
```

- Delete: uma lista contendo os identificadores de mensagens que foram deletadas pelos usuários, as quais devem ser deletadas também pelo consumidor dos dados;

```
{
  "delete":{
    "status":{
    "id":1234,
    "id_str":"1234",
    "user_id":3,
    "user_id_str":"3"
  }
}
```

- Scrub_geo: lista de identificadores de usuários que configuraram suas contas para remover informações geográficas de suas mensagens, as quais devem também ser removidas pelo sistema consumidor destes dados.

```
{
  "scrub_geo":{
    "user_id":14090452,
    "user_id_str":"14090452",
    "up_to_status_id":23260136625,
    "up_to_status_id_str":"23260136625"
  }
}
```

- Tweet: uma mensagem enviada pelo usuário;

```
{
  "created_at":"Mon Jul 02 22:30:01 +0000 2012",
  "id":219920891264503808,
  "id_str":"219920891264503808",
  "text":"@lubargas eu vou sim kkk",
  "source":"<a href=\"http:\/\/www.twitter.com\" rel=\"nofollow\">Twitter for Windows Phone<\/a>",
  "truncated":false,
  "in_reply_to_status_id":219916417041055745,
  "in_reply_to_status_id_str":"219916417041055745",
  "in_reply_to_user_id":99414243,
  "in_reply_to_user_id_str":"99414243",
  "in_reply_to_screen_name":"lubargas",
  "user":{
    "id":63591725,
    "id_str":"63591725",
   "name":"Gabriela Marçal",
   "screen_name":"gabimarcal_",
   "location":"Belo Horizonte, Brasil",
   "description":"'Onde estão teus olhos agora que eu tô bem na foto?'",
   "url":"http:\/\/varias-variaveis.tumblr.com",
   "protected":false,
   "followers_count":226,
```

```
        "friends_count":193,
        "listed_count":8,
        "created_at":"Fri Aug 07 00:54:40 +0000 2009",
        "favourites_count":62,
        "utc_offset":-10800,
        "time_zone":"Brasilia",
        "geo_enabled":true,
        "verified":false,
        "statuses_count":20508,
        "lang":"pt",
        "contributors_enabled":false,
        "is_translator":false,
        "profile_background_color":"F04FFF",
        "profile_background_image_url":"http:\/\/a0.twimg.com\/profile_background_
images\/575210979\/hyhgb243pf9ywh4i4mlm.jpeg",
        "profile_background_image_url_https":"https:\/\/si0.twimg.com\/profile_
background_images\/575210979\/hyhgb243pf9ywh4i4mlm.jpeg",
        "profile_background_tile":true,
        "profile_image_url":"http:\/\/a0.twimg.com\/profile_images\
/2368457324\/d8p078jgsy6z233ht6rd_normal.jpeg",
        "profile_image_url_https":"https:\/\/si0.twimg.com\/profile_images\
/2368457324\/d8p078jgsy6z233ht6rd_normal.jpeg",
        "profile_link_color":"D16DD1",
        "profile_sidebar_border_color":"AB84BA",
        "profile_sidebar_fill_color":"FFEBFA",
        "profile_text_color":"000000",
        "profile_use_background_image":true,
        "show_all_inline_media":false,
        "default_profile":false,
        "default_profile_image":false,
        "following":null,
        "follow_request_sent":null,
        "notifications":null
      },
      "geo":{
      "type":"Point",
      "coordinates":[-19.92967987,-43.94337463]
    },
    "coordinates":{
    "type":"Point",
    "coordinates":[-43.94337463,-19.92967987]
    },
    "place":{
    "id":"d9d978b087a92583",
    "url":"http:\/\/api.twitter.com\/1\/geo\/id\/d9d978b087a92583.json",
    "place_type":"city",
    "name":"Belo Horizonte",
    "full_name":"Belo Horizonte, Minas Gerais",
    "country_code":"BR",
    "country":"Brasil",
```

```
"bounding_box":{
"type":"Polygon",
"coordinates":[
[
[-44.062789,-20.059816],
[-43.856856,-20.059816],
[-43.856856,-19.777568],
[-44.062789,-19.777568]
]
]
},
"attributes":{}
},
"contributors":null,
"retweet_count":0,
"entities":{
"hashtags":[],
"urls":[],
"user_mentions":[{
"screen_name":"lubargas",
"name":"Luísa Bargas",
"id":99414243,
"id_str":"99414243",
"indices":[0,9]
}
]
},
"favorited":false,
"retweeted":false
}
```

APÊNDICE B - SISTEMA ONLINE DE IDENTIFICAÇÃO DE EVENTOS

As evoluções obtidas ao longo deste trabalho, para se chegar à proposta de método definida no Capítulo 4, foram alcançadas a partir da criação de um sistema online para identificação de eventos, utilizando como fonte os dados do serviço Twitter. À medida que este sistema apresentava os eventos identificados, melhorias eram elaboradas para aperfeiçoar a identificação em eficiência (tempo de execução) e eficácia (identificar menos falsos positivos). Um conjunto de outras ferramentas foi desenvolvido para apoiar o diagnóstico do sistema e definir a eficácia do mesmo.

Esse sistema online de identificação de eventos (candidatos), que ficou operando no período de 30 de Abril a 10 de Agosto de 2012. A Tabela B.1 mostra a quantidade de eventos detectados em cada uma das cidades utilizadas como regiões de interesse. As primeiras cidades a terem os eventos identificados foram Porto Alegre, Rio de Janeiro, São Paulo, Brasília, Curitiba, Belo Horizonte, Santos e Recife as demais foram incluídas ao longo do tempo. O sistema teve seu funcionamento interrompido quando, por um erro administrativo, as máquinas foram desligadas e removidas do laboratório em seguida formatadas, perdendo assim dois bilhões de dados coletados no período de mais de um ano e meio.

Tabela B.1: Quantidade de eventos identificados em cada cidade e a data de inclusão da cidade no sistema de detecção e identificação

Cidade	*Quantidade de Eventos*	*Data de Inclusão*
Porto Alegre	244	30/04/2012
Rio de Janeiro	181	30/04/2012
São Paulo	308	30/04/2012
Brasília	190	30/04/2012
Curitiba	218	30/04/2012
Belo Horizonte	215	30/04/2012
Santos	188	30/04/2012
Recife	161	30/04/2012
Los Angeles	128	07/05/2012
Madrid	359	07/05/2012
London	254	14/05/2012

Manchester	183	17/05/2012
Barcelona	150	17/05/2012
New York	112	23/05/2012
Moscow	287	23/05/2012
Paris	93	23/05/2012

Uma das principais melhorias na redução do tempo de execução do sistema, que permitiu a inclusão de mais cidades no monitoramento, é descrita a seguir. O sistema executava sua rotina de detecção e identificação a cada 30 minutos, assim para cada cidade eram consultados os dados referentes a 10 dias de coleta a contar do instante atual. Essa consulta demorava muito tempo para ser executada por se tratar de um banco de dados com milhões de registros, desta forma a melhoria realizada foi armazenar os dados referentes à primeira consulta de 10 dias e nas consultas seguintes pesquisar somente as informações dos minutos que se passaram desde então.

O sistema processava as informações de 10 dias de dados a cada 30 minutos, desta forma um mesmo evento era identificado mais de uma vez. Outra melhoria obtida na redução do tempo foi a de não calcular as palavras-chave para eventos já identificados anteriormente. A média de vezes que um evento era identificado foi de 90,03, com um desvio padrão de 142,26.

Um dos dados armazenados pelo sistema era a frequência em que as palavras-chave apareciam no conjunto de tweets de um slot de tempo identificado como evento. As palavras-chave com maior média de frequência de ocorrência no momento em que aparecem, e que ocorreram em pelo menos 10 eventos foram: *corinthians, paulo, libertadores, palmeiras, fogos, globo, campeão, inter*. As palavras que ocorreram em mais eventos foram: *amo, deus, melhor, time, linda, mãe, jogo, dormir, vem, gol*.

À medida que palavras sem sentido ocorriam em muitos eventos, elas eram adicionadas na lista de stopwords. Essa inclusão era frequente em palavras utilizadas no vocabulário da internet, simplificações de stopword (*mesmo* para *msm*) ou palavras que não ajudam a descrever eventos como: *vou, bom, achei, quero*.

B.1 Arquitetura

O sistema funcionava com uma arquitetura de três servidores por questões técnicas e financeiras. Apresentados na Figura B.1 o primeiro servidor, *coletor*, executava dois processos responsáveis por coletar os tweets com a Twitter API e a cada 30 minutos (minuto 0 e minuto 30 de cada hora) enviar para o segundo servidor, *banco de dados*. Uma terceira máquina, *monitor*, era responsável pela identificação dos eventos (candidatos), a cada 30 minutos (minuto 10 e minuto 40 de cada hora) ele consultava o *banco de dados* para obter as medidas mais recentes de cada região monitorada e realizava o processo de identificação dos eventos. As informações dos eventos identificados, data, local e palavras-chave, eram enviadas para o *banco de dados* e lá ficavam armazenadas.

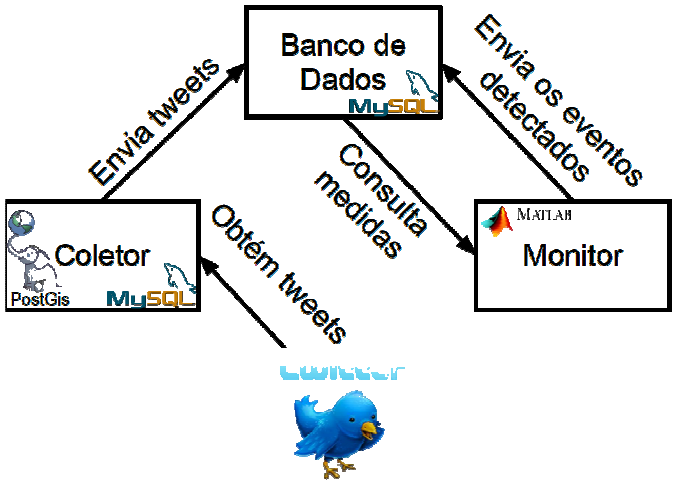

Figura B.1: Arquitetura utilizada no sistema online de detecção e identificação

Abaixo são enumerados os servidores e os motivos técnicos e financeiros para existência de cada um:

- *Coletor*: localizado nos EUA, tinha um custo financeiro de trafego de rede baixo, alta disponibilidade, excelente conectividade com o Twitter e possibilidade de instalar e gerenciar os softwares dentro dele. A contra partida era o alto custo para armazenamento.
- *Banco de dados*: localizado no Brasil, tinha um custo fixo e uma possibilidade quase ilimitada de armazenamento. Não era possível instalar e gerenciar os softwares desta máquina.
- *Monitor*: localizado no Instituto de Informática, não tinha custo de trafego e armazenamento, tinha acesso para instalação e gerenciamento de softwares. Possuía componentes de baixa qualidade que impactavam no processamento das informações e baixa capacidade de armazenamento.

Devido a problemas com a infraestrutura de hardware, rede e intermitências da API se viu necessário a implementação de um processo auxiliar, responsável por verificar se o processo coletor ainda estava ativo ou não, religando-o caso necessário. Um desses problemas de hardware foi a utilização de um servidor no qual as memórias RAM não eram ECC[12] ocasionando erros no recebimento das informações assim como o desligamento do processo coletor. A infraestrutura do coletor foi alterada para um servidor nos EUA, o qual tem mais poder de processamento, maior tolerância a falhas e melhor banda de rede reduzindo em muito os problemas de desligamento do processo coletor. Mesmo assim o processo auxiliar é necessário, pois mudanças na infraestrutura e configurações do Twitter podem ocasionar o desligamento do processo, o que ocorre de modo aleatório e imprevisível. No mês de Junho de 2012 os processos foram reiniciados oito vezes, já em Julho do mesmo ano foi apenas uma vez. O processo auxiliar existe separadamente para cada coletor, é executado a cada 2 minutos e verifica se o processo coletor está vivo e ativo, ou seja, se ele consumiu mensagens nos últimos 2 minutos, caso contrário o processo atual é finalizado e um novo é criado.

O custo das máquinas localizadas em servidores externos ao da instituição de ensino era mantido com recurso do autor.

[12] Memória ECC é um tipo de armazenamento computacional que pode detectar e corrigir os erros mais comuns de corrupção de dados internos.

APÊNDICE C – OUTRAS TABELAS

Este apêndice apresenta os valores utilizados para a criação da Figura 5.2 e da Figura 5.4. As tabelas possuem os valores abaixo variando o tamanho do slot de tempo em dados de diversas regiões, identificadas em cada tabela. O período de coleta destes dados foi entre os dias 14 de Fevereiro e 22 de Março de 2012.

- Média de tweets por slot de tempo;
- Média de usuários por slot de tempo;
- Quantidades outliers detectados pela rede neural IGMN; e
- Quantidade de outliers detectados utilizando Boxplot.

Tabela C.1: Valores para o país Cyprus

Slot de Tempo (em minutos)	*Média de Tweets por slot*	*Média de Usuários por slot*	*Quantidade de Outliers / IGMN*	*Quantidade de Outliers / BoxPlot*
1	1,54	1,36	324	795
2	1,98	1,66	211	350
3	2,44	1,96	127	202
4	2,91	2,26	84	129
5	3,40	2,56	51	105
6	3,88	2,85	55	89
10	5,89	4,01	35	46
12	6,90	4,58	17	32
15	8,42	5,40	13	27
20	10,93	6,68	11	25
30	15,90	9,04	4	12
60	30,79	15,34	1	4

Tabela C.2: Valores para o país Israel

Slot de Tempo (em minutos)	*Média de Tweets por slot*	*Média de Usuários por slot*	*Quantidade de Outliers / IGMN*	*Quantidade de Outliers / BoxPlot*
1	1,7856	1,6769	349	864
2	2,7149	2,3878	125	274
3	3,6939	3,0811	59	169
4	4,7134	3,7721	40	106
5	5,7398	4,4430	36	80

6	6,7556	5,0818	25	54
10	10,7972	7,5200	20	39
12	12,7911	8,6644	13	27
15	15,8184	10,3815	15	24
20	20,7551	13,0684	8	12
30	30,8331	18,2881	6	10
60	61,1718	33,0069	3	2

Tabela C.3: Valores para o país Finlândia

Slot de Tempo (em minutos)	*Média de Tweets por slot*	*Média de Usuários por slot*	*Quantidade de Outliers / IGMN*	*Quantidade de Outliers / BoxPlot*
1	2,4955	2,3749	467	897
2	4,2294	3,8427	151	293
3	6,0062	5,2561	53	125
4	7,7709	6,5929	41	85
5	9,5545	7,8963	20	49
6	11,3327	9,1525	10	46
10	18,4360	13,8802	12	17
12	22,1037	16,0753	2	16
15	27,5774	19,2076	2	9
20	36,7104	24,2588	5	8
30	55,0446	33,7737	1	6
60	109,9635	58,8676	1	1

Tabela C.4: Valores para o país Austrália

Slot de Tempo (em minutos)	*Média de Tweets por slot*	*Média de Usuários por slot*	*Quantidade de Outliers / IGMN*	*Quantidade de Outliers / BoxPlot*
1	12,9458	12,2147	193	411
2	25,4884	22,7870	74	160
3	38,1306	32,6566	42	98
4	50,8214	42,0621	27	70
5	63,5061	51,0691	18	44
6	76,2088	59,8379	14	27
10	127,0002	92,9971	10	27
12	152,3828	108,8315	3	17
15	190,4458	131,8114	10	15
20	253,8552	168,1550	5	12
30	380,7829	237,2074	3	6
60	760,6963	425,6347	2	3

Tabela C.5: Valores para a cidade Brasília

Slot de Tempo (em minutos)	*Média de Tweets por slot*	*Média de Usuários por slot*	*Quantidade de Outliers / IGMN*	*Quantidade de Outliers / BoxPlot*

1	17,1563	15,0221	232	512
2	33,1607	25,7690	64	281
3	49,0211	34,7373	45	187
4	64,8079	42,5491	27	145
5	80,5941	49,6388	17	139
6	96,4346	56,2058	14	115
10	159,5043	79,1048	11	67
12	191,1194	89,2795	7	68
15	238,7074	103,6922	1	52
20	317,7297	125,7480	6	39
30	475,7770	164,8113	1	28
60	949,9247	263,9623	2	9

Tabela C.6: Valores para a cidade Belo Horizonte

Slot de Tempo (em minutos)	*Média de Tweets por slot*	*Média de Usuários por slot*	*Quantidade de Outliers / IGMN*	*Quantidade de Outliers / BoxPlot*
1	17,3824	15,0976	257	503
2	33,6077	25,8674	38	225
3	49,6353	34,8146	21	170
4	65,7257	42,6827	11	139
5	81,7021	49,7820	11	113
6	97,8593	56,3649	4	96
10	161,8270	79,1005	5	56
12	193,8812	89,1729	2	50
15	242,1158	103,3180	5	43
20	322,6673	124,9375	1	33
30	483,8166	163,1737	4	24
60	966,5285	259,1644	1	13

Tabela C.7: Valores para a cidade Porto Alegre

Slot de Tempo (em minutos)	*Média de Tweets por slot*	*Média de Usuários por slot*	*Quantidade de Outliers / IGMN*	*Quantidade de Outliers / BoxPlot*
1	29,6793	25,3007	155	377
2	58,6578	43,4271	73	146
3	87,6826	58,1842	58	92
4	116,7697	70,8246	46	55
5	145,8730	81,9857	42	50
6	175,0944	92,1273	42	35
10	291,6905	126,3339	10	20
12	349,9886	140,8992	17	12
15	437,4107	160,8311	15	7
20	583,0476	190,4701	15	3
30	874,5714	241,3143	11	3
60	1747,1461	362,8368	2	1

Tabela C.8: Valores para a cidade Curitiba

Slot de Tempo (em minutos)	*Média de Tweets por slot*	*Média de Usuários por slot*	*Quantidade de Outliers / IGMN*	*Quantidade de Outliers / BoxPlot*
1	31,9220	26,7740	176	602
2	62,3108	45,2119	108	315
3	92,6788	60,1684	49	232
4	123,0746	72,9110	29	181
5	153,5198	84,2105	17	144
6	184,1041	94,4882	17	133
10	306,1616	129,2327	6	85
12	367,3660	144,2358	6	67
15	459,0237	164,7925	5	63
20	611,8568	195,6328	5	45
30	917,7851	248,9566	1	29
60	1833,4749	377,3858	1	17

Tabela C.9: Valores para a cidade Rio de Janeiro

Slot de Tempo (em minutos)	*Média de Tweets por slot*	*Média de Usuários por slot*	*Quantidade de Outliers / IGMN*	*Quantidade de Outliers / BoxPlot*
1	41,5801	35,6131	149	579
2	82,6075	61,8955	75	284
3	123,7671	83,6429	63	169
4	164,9378	102,4257	42	125
5	206,1409	119,2028	32	93
6	247,4068	134,6122	32	83
10	412,2032	187,4587	17	46
12	494,5872	210,5612	19	35
15	618,1280	242,6828	22	27
20	823,9352	290,9470	10	19
30	1235,9029	376,6029	9	13
60	2468,9840	588,2386	1	7

Tabela C.10: Valores para a cidade São Paulo

Slot de Tempo (em minutos)	*Média de Tweets por slot*	*Média de Usuários por slot*	*Quantidade de Outliers / IGMN*	*Quantidade de Outliers / BoxPlot*
1	101,17216	85,88791	280	624
2	202,15138	149,32373	158	309
3	303,16347	201,17920	69	207
4	404,16402	245,54247	63	154
5	505,11836	284,91347	49	122
6	606,29227	320,89190	29	102
10	1010,14046	443,26282	17	54
12	1212,02996	496,09101	33	42
15	1514,77765	569,44899	10	42

20	2019,12648	679,97981	11	25
30	3028,68971	873,92457	3	19
60	6050,46461	1352,96461	1	9