

UNIVERSIDADE DE CAXIAS DO SUL  
CENTRO DE COMPUTAÇÃO E TECNOLOGIA DA INFORMAÇÃO  
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ÉDIPO DEON TERRA

**Ferramenta para Extração de Dados  
do Twitter para Mineração de Dados**

Helena Grazziotin Ribeiro  
Orientador

Caxias do Sul, Dezembro de 2015

# **Ferramenta para Extração de Dados do Twitter para Mineração de Dados**

por

Édipo Deon Terra

Projeto de Diplomação submetido ao curso de Bacharelado em Sistemas de Informação do Centro de Computação e Tecnologia da Informação da Universidade de Caxias do Sul, como requisito obrigatório para graduação.

## **Trabalho de Conclusão de Curso**

Orientador: Helena Grazziotin Ribeiro

Banca examinadora:

Daniel Luís Notari  
CCTI/UCS

Sheila de Ávila e Silva  
CCTI/UCS

Projeto de Diplomação apresentado em  
8 de Dezembro de 2015

Daniel Luís Notari  
Coordenador

## **AGRADECIMENTOS**

Inicialmente, agradeço imensamente a minha orientadora que orientou, incentivou e colaborou muito com o desenvolvimento do trabalho. Colaborou com artigos recentes, onde a cada encontro para orientação, aparecia com conteúdos e sugestões que foram bem aproveitadas no trabalho. E o mais importante, me direcionou para onde e como abordar determinados assuntos.

Agradeço também, de forma não menos importante, a compreensão, tolerância e incentivo de minha família em todos os momentos do andamento deste trabalho. Desde os momentos mais simples até os momentos de crise e desânimo. O incentivo foi mais que decisivo na finalização deste trabalho.

O incentivo direto e indireto que foi prestado pelos colegas e amigos não pode ser esquecido, e merece o seu respectivo agradecimento. Que varia desde a apresentação de determinadas ferramentas que facilitaram o desenvolvimento do trabalho a conversas informais de andamento dos outros trabalhos, que indiretamente acabaram incentivando no desenvolvimento deste.

Queria dizer a todos os envolvidos, muito obrigado!

# SUMÁRIO

<b>LISTA DE ACRÔNIMOS</b>	4
<b>LISTA DE FIGURAS</b>	5
<b>LISTA DE TABELAS</b>	6
<b>RESUMO</b>	7
<b>1 INTRODUÇÃO</b>	8
1.1 Problema	9
1.2 Questão de Pesquisa	10
1.3 Objetivos	10
1.3.1 Objetivo Geral	10
1.3.2 Objetivos Específicos	10
1.4 Estrutura do Texto	10
<b>2 REFERENCIAL TEÓRICO</b>	12
2.1 <i>Big Data</i>	13
2.1.1 Volume	13
2.1.2 Velocidade	13
2.1.3 Variedade	13
2.1.4 Veracidade	14
2.1.5 Valor	14
2.2 Descoberta de Conhecimento em Banco de Dados	14
2.2.1 Aplicações da Descoberta de Conhecimento	15
2.2.2 Seleção de Dados	15
2.2.3 Pré-Processamento	16
2.2.4 Codificação de Dados	17
2.2.5 Mineração de Dados	17
2.2.6 Atividades Preditivas	17
2.2.7 Atividades Descritivas	18
2.2.8 Exibição dos Resultados	19
2.3 Redes Sociais	19
2.3.1 <i>Twitter</i>	20

<b>2.4</b>	<b>Extração de Dados</b>	22
<b>2.5</b>	<b>Trabalhos Relacionados</b>	23
2.5.1	Observatório do Trânsito	23
2.5.2	Descobrimento de eventos locais do Twitter	24
2.5.3	<i>Tweetmining</i>	24
2.5.4	Comparativo de Semelhança	25
<b>3</b>	<b>PROPOSTA DE SOLUÇÃO</b>	26
<b>3.1</b>	<b>Hipóteses</b>	26
3.1.1	Engarrafamentos	27
3.1.2	Acidentes	28
3.1.3	Atropelamentos	29
<b>3.2</b>	<b>Rede Social Escolhida</b>	29
3.2.1	<i>APIs (Application Programming Interfaces)</i> do Twitter	30
3.2.2	Bibliotecas do Twitter	31
<b>3.3</b>	<b>Escolha do Algoritmo de Mineração</b>	32
3.3.1	Algoritmo Apriori	33
3.3.2	Algoritmo Tertius	33
<b>3.4</b>	<b>Ferramentas de Mineração de Dados</b>	33
<b>3.5</b>	<b>Organização das Informações</b>	33
<b>3.6</b>	<b>Extração de Dados</b>	34
<b>3.7</b>	<b>Proposta de Protótipo de Pós-processamento</b>	35
<b>4</b>	<b>IMPLEMENTAÇÃO</b>	36
<b>4.1</b>	<b>Banco de Dados Mongo DB</b>	36
4.1.1	Instalação	37
4.1.2	Modelo de Banco de Dados	38
<b>4.2</b>	<b><i>API Twitter</i></b>	39
4.2.1	Aplicação do <i>Twitter</i>	40
4.2.2	Parâmetros para Busca	40
<b>4.3</b>	<b>Desenvolvimento do Protótipo de Extrator de Dados</b>	41
4.3.1	Tecnologias aplicadas	41
4.3.2	Metodologias aplicadas	42
4.3.3	Extrator	43
4.3.4	Termos de Busca	45
4.3.5	Coordenadas	46
4.3.6	Exportação de Dados	47

<b>5</b>	<b>ANÁLISE E AVALIAÇÃO DAS INFORMAÇÕES . . . . .</b>	<b>49</b>
5.1	Parâmetros Cadastrados . . . . .	49
5.2	Coleta dos Dados . . . . .	50
5.3	Utilização do Software de Mineração . . . . .	51
5.4	Resultados Obtidos . . . . .	53
5.4.1	Apriori . . . . .	53
5.4.2	Tertius . . . . .	54
5.4.3	Apresentação de Resultados . . . . .	54
<b>6</b>	<b>CONCLUSÃO . . . . .</b>	<b>58</b>
6.1	Trabalhos Futuros . . . . .	59
	<b>REFERÊNCIAS . . . . .</b>	<b>60</b>

# LISTA DE ACRÔNIMOS

<b>API</b>	<i>Application Programming Interface</i>
<b>ARFF</b>	<i>Attribute Relationship File Format</i>
<b>CSV</b>	<i>Comma Separated Values</i>
<b>HTTP</b>	<i>Hypertext Transfer Protocol</i>
<b>JSON</b>	<i>Java Script Object Notation</i>
<b>KDD</b>	<i>Knowledge Discovery Database</i>
<b>PHP</b>	<i>PHP: Hipertext Processor</i>
<b>SVM</b>	<i>Support Vector Machine</i>
<b>XML</b>	<i>Extensible Markup Language</i>
<b>GSP</b>	<i>Generalized Sequential Patterns</i>
<b>CEP</b>	Código de Endereçamento Postal
<b>CRM</b>	<i>Customer Relationship Management</i>
<b>SQL</b>	<i>Strutured Query Language</i>
<b>MVC</b>	<i>Model-View-Controller</i>
<b>CSS</b>	<i>Cascading Style Sheet</i>

## LISTA DE FIGURAS

Figura 2.1: Etapas da Descoberta de Conhecimento em Banco de Dados . . .	15
Figura 2.2: Exemplo de mensagem do <i>Twitter</i> . . . . .	21
Figura 2.3: Análise de comprovação do método de análise das localizações das mensagens do Twitter . . . . .	24
Figura 3.1: Exemplo de informação de Trânsito Lento no Twitter . . . . .	27
Figura 3.2: Exemplo de informação de Acidente de Trânsito no Twitter . . .	28
Figura 3.3: Exemplo de informação de Atropelamento no <i>Twitter</i> . . . . .	29
Figura 4.1: Modelo de banco de dados . . . . .	38
Figura 4.2: Estrutura da Aplicação . . . . .	42
Figura 4.3: Estrutura da Aplicação - Aguardando a próxima extração . . . .	43
Figura 4.4: Fluxo de coleta de dados do <i>Twitter</i> . . . . .	45
Figura 4.5: Protótipo de tela dos termos de Busca . . . . .	46
Figura 4.6: Protótipo de tela dos dois pares de coordenadas . . . . .	47
Figura 4.7: Exemplo do layout do arquivo gerado pelo Extrator . . . . .	48
Figura 5.1: Tela do Weka na etapa de pré-processamento . . . . .	52
Figura 5.2: Gráfico de relação entre Componentes do Trânsito e as ocorrências das regras de associação levantadas . . . . .	55
Figura 5.3: Gráfico de relação entre os períodos e as ocorrências das regras de associação levantadas . . . . .	56
Figura 5.4: Gráfico de relação entre os dias da semana e as ocorrências das regras de associação levantadas . . . . .	56



## **LISTA DE TABELAS**

Tabela 2.1: Utilização do KDD conforme área . . . . .	15
Tabela 2.2: Comparativo de Trabalhos Relacionados . . . . .	25

## RESUMO

As redes sociais possuem uma imensa quantidade de informações, dentre estas informações, muitas estão relacionadas com o trânsito, distribuídas entre as principais redes sociais como *Facebook* e *Twitter*. Este trabalho teve como objetivo explorar estas informações relacionadas com o trânsito através da aplicação das etapas do processo de Descoberta de Conhecimento em Banco de Dados. Como resultado desta pesquisa, tem-se alguns perfis do trânsito traçados para auxiliar na identificação dos problemas com o trânsito nas localidades da Serra Gaúcha e região metropolitana do Rio Grande do Sul.

**Palavras-chave:** Mineração de Dados, Trânsito.

# 1 INTRODUÇÃO

A cada ano a tecnologia está mais presente no cotidiano das pessoas, através de computadores, *smartphones*, televisores, meios de transporte, entre outros. Assim como todos estes equipamentos citados, a *internet* é uma destas tecnologias que estão recebendo bastante destaque devido sua disponibilidade nos mais diversos tipos de aparelhos. Além de computadores e *smartphones* ela está disponível em televisores, equipamentos de segurança e até automóveis. Fatos estes que aumentaram significativamente o número de acessos à *internet*.

Com o número maior de usuários que estão utilizando a *internet*, a transmissão e criação de dados está mais elevado, como consequência a quantidade de informações disponíveis na mesma aumentou consideravelmente. Parte destas informações são derivadas dos grandes portais de notícias, *blogs* e redes sociais, onde este último teve um crescimento considerável nestes últimos anos, devido a popularidade do *Facebook* e *Twitter*. Este crescimento acabou influenciando na estruturação de alguns programas de televisão, onde as mensagens servem como forma de complemento do assunto que está sendo abordado no programa.

As redes sociais como *Twitter*, *Facebook* e *LinkedIn* permitem que os usuários deixem de ser apenas consumidores de informações e se tornem produtores de informações, contribuindo com as informações. Isto reflete numa maior quantidade de informações inseridas a cada hora, onde estas podem ser dos mais diversos assuntos, sejam sobre atualizações do cotidiano pessoal, notícias de acontecimentos, anúncios de publicidade e tantos outros tipos de mensagens.

No meio de tanta informação, uma quantidade considerável está relacionada com o trânsito, sabendo que muitas mensagens delas não tratam exatamente de ocorrências ou situações do trânsito. Existem também os tipos de mensagens onde a atualização da situação é mais precisa e completa, que normalmente é mantida por alguns órgãos especializados no trânsito que disponibilizam um serviço de publicação informações, como é o caso da EPTC de Porto Alegre, alguns perfis da Polícia Rodoviária Federal e órgãos de observatório de trânsito.

Alguns trabalhos foram realizados possuindo o objetivo de explorar as informações contidas nas redes sociais, podendo explorar os dados do trânsito ou não, como exemplo os trabalhos de (SILVA; BENEVUTO; ALMEIDA, 2010) e de (SILVEIRA, 2010). Nos trabalhos que não estão relacionados com o trânsito, mas que podem se destacar pela pesquisa que fazem em relação à descoberta de eventos através das redes

sociais, em especial eventos mencionados através do *Twitter*. Já em trabalhos que efetuam a análise das informações do trânsito, existe uma grande preocupação em determinar qual o local onde ocorreu determinado evento. Em grande parte destas pesquisas são utilizados os conceitos de Descoberta de Conhecimento em Banco de Dados, que visa efetuar a coleta, limpeza e análise dos dados.

A Descoberta de Conhecimento em Banco de Dados é uma área bastante estudada e explorada em diversos ambientes, e com o surgimento e grande disseminação do *Big Data*, esta busca pelos processos de busca em banco de dados ou até mesmo redes sociais se intensificou muito nos últimos anos. Por exemplo, o *KDD (Knowledge Discovery Database)* está presente no estudo de diversos trabalhos acadêmicos, como no Observatório da Web (INWEB, 2014). O mesmo comportamento se repetiu na Universidade de Caxias do Sul, onde nos últimos anos podem ser citados os trabalhos da Marina Bellenzier e o trabalho do Rodrigo Francischelli (FRANCISCHELLI, 2013). Ambos os trabalhos tiveram como base de estudo o *KDD* e aplicaram algum tipo de mineração sobre os dados disponíveis.

## 1.1 Problema

A cada ano a quantidade de veículos presente nas ruas está cada vez maior e como consequência disso aparecem, ou se agravam os problemas no trânsito. Alguns dos exemplos que podem ser facilmente identificados são o aumento dos congestionamentos, ocorrência de mais atropelamentos e aumento de acidentes em geral. Os acidentes influenciam muito na lentidão do trânsito, contribuindo para o surgimento de engarrafamentos.

Tentando amenizar estes problemas do trânsito, foram desenvolvidos alguns aplicativos para dispositivos móveis que visam alertar os motoristas de situações de perigo e indica quais as rotas de desvio destes pontos. Neste tipo de aplicativo se destaca o *Waze*, que além de funcionar como um aplicativo de mapas para o trânsito, também funciona como uma rede social onde as pessoas publicam as atualizações do trânsito de forma fácil e rápida, entre os usuários do aplicativo. Além disso este aplicativo verifica a velocidade média de circulação dos veículos nas vias para avisar os usuários de possíveis congestionamentos.

As redes sociais mais populares possuem informações similares e que podem tratar da praticamente da mesma situação que a disponibilizada pelo *Waze*, porém de uma forma um pouco diferenciada e restrita. Estas redes sociais pecam pela praticidade, pois a inserção ainda deve ser feita manualmente e o consumo destas informações ainda é muito demorado, pois exige a leitura e interpretação das mensagens em sua forma integral. A prática do consumo de informações das redes sociais utilizando esta forma aumenta consideravelmente o perigo de ocasionar algum aci-

dente. Entretanto estas redes sociais servem muito bem para uma base de análise do comportamento dos componentes do trânsito e para análises estatísticas, já que estas contribuem muito na riqueza de detalhes, coisa que não pode ser percebidas nas informações do *Waze*.

Toda a informação disponível pode ser extraída das redes sociais, através de aplicações específicas para esta atividade ou através do uso de *APIs* (*Application Programming Interfaces*), muitas vezes disponibilizadas pelas próprias redes sociais. Aliás, as redes sociais em si incentivam muito a criação de aplicações para utilizar de suas informações, de forma que muitas delas possuem áreas exclusivas para os desenvolvedores aprenderem como funciona a rede social, e auxilia na sua manipulação contribuem para a popularidade da mesma. Em conjunto com as *APIs*, pode ser efetuada a aplicação de técnicas de mineração de dados, definidas e estruturadas pelo procedimento de *KDD*, onde é possível encontrar algumas relações das ocorrências e componentes do trânsito.

## 1.2 Questão de Pesquisa

Com base nas informações disponíveis no *Twitter*, é possível explorar estas informações de forma que seja possível adquirir algum tipo de conhecimento através de mineração de dados?

## 1.3 Objetivos

### 1.3.1 Objetivo Geral

Desenvolver um protótipo de extrator de dados da rede social *Twitter*. Explorar um estudo de caso a fim de avaliar o funcionamento do protótipo de extrator.

### 1.3.2 Objetivos Específicos

Como forma de auxiliar o atingimento do objetivo geral, foram definidos os seguintes objetivos específicos:

- Coletar e processar os dados do *Twitter* sobre o trânsito da região da Serra Gaúcha e Região Metropolitana;
- Utilizar uma ferramenta de mineração de Dados para analisar os dados extraídos;
- Apresentar os resultados obtidos da análise das informações do trânsito.

## 1.4 Estrutura do Texto

Este trabalho de conclusão está organizado em cinco capítulos, os quais visam dar toda a base inicial necessária para um bom entendimento dos conceitos e técnicas utilizadas neste trabalho. Além disso apresenta também algumas técnicas e experimentos já foram efetuadas em outros trabalhos e por último explica todo o processo que foi adotado para definir uma solução para o problema proposto, assim como o processo de desenvolvimento do extrator e análise dos dados.

No capítulo um (1) está definida a contextualização e a descrição detalhada do problema, a definição dos objetivos e a organização do presente trabalho.

No capítulo dois (2) estão descritos os conceitos necessários para compreensão do trabalho, assim como algumas das técnicas que foram aplicadas neste trabalho e os algoritmos de mineração necessários para efetuação da análise. Dentre os trabalhos relacionados, destacam-se a Descoberta de Conhecimento em Banco de Dados, os conceitos sobre as redes sociais e os conceitos sobre *Big Data*. Ainda no segundo capítulo estão os trabalhos relacionados onde já foram efetuadas algumas atividades semelhantes ao que foi proposto neste trabalho. Neste capítulo também foi efetuado uma análise de semelhança com este trabalho.

No capítulo três (3) apresenta-se a proposta de solução, detalhando cada uma das etapas necessárias para atingir os objetivos delineados e descritos na introdução.

No capítulo quatro (4) descreve-se todo o processo e modelo de desenvolvimento do extrator de dados, que foi necessário para a análise dos dados, assim como o processo necessário para instalar os componentes para o funcionamento do extrator de dados.

No capítulo cinco (5) estão apresentados os processos de análise, a análise através de cada algoritmo e os resultados da análise encontrados no término da mineração de dados.

## 2 REFERENCIAL TEÓRICO

Trabalhar com grandes bases de dados e com grandes conjuntos de informações é uma das atividades que tem sido muito estudada atualmente, principalmente nas pesquisas relacionadas com tecnologia da informação, nas quais estão voltadas para o desenvolvimento de novas formas de manipular as informações que estão contidas nestas bases de dados. Além de tentar descobrir novas formas de descobertas, estão sendo atualizadas e aprimoradas as técnicas mais antigas, para que funcionem melhor com as tecnologias mais atuais. Desta forma será efetuado um aproveitamento maior da capacidade de processamento de dados que temos atualmente.

Assim como existem muitos dados em bancos empresariais, a *internet* também possui esta característica, estando repleta de informações muito úteis. Estes dados provém de portais de notícias, aplicativos móveis, redes sociais, *blogs*, vídeos, *streaming* e tantas outras fontes. As redes sociais detém de uma grande quantidade de informações, devido à forma de interação que este tipo de aplicação permite, permitindo que a fonte produtora de dados seja descentralizada e múltipla. Com esta descentralização, os assuntos tratados nas redes sociais variam entre muitos, desde os mais cotidianos a até assuntos mais específicos como as atualizações da situação de trânsito.

Este trabalho possui o principal foco de desenvolver um protótipo de extrator de dados e analisar as informações contidas no *Twitter* como forma de estudo de caso para validar o funcionamento do protótipo implementado. Das informações relacionadas ao trânsito, elas são derivadas da região da Serra Gaúcha e região metropolitana, desta forma torna-se possível traçar alguns perfis do trânsito e identificar os padrões e desvio padrões da situação avaliada.

Para que o objetivo traçado possa ser cumprido, é importante apresentar os conceitos, métodos e técnicas que serão essenciais para o desenvolvimento da pesquisa. Dentre estes conceitos estão os relacionados à descoberta de conhecimento em banco de dados, que está apresentado neste capítulo na sessão 2.2, juntamente à ele estão as definições de Mineração de Dados, pré-processamento, pós-processamento, cujas informações estão nesta mesma sessão desenvolvida para a Descoberta de Conhecimento.

Os conceitos que definem o que é uma Rede Social e como ela se organiza é tão importante quanto como saber como trabalhar com as informações dela. Por este motivo, estão inseridas nas sessões seguintes as definições e conceitos de uma

rede social, assim como a estruturação e organização do *Twitter*, pois é fundamental entender como são organizadas as informações desta rede social para que uma estratégia de coleta, organização e mineração dos dados ser determinada.

## 2.1 *Big Data*

Conforme foi apresentado por (ALECRIM, 2015), o *Big Data* surgiu como uma forma de complementar os processos da mineração de dados e Descoberta de Conhecimento, e com ele vieram uma série de ferramentas que auxiliam nas consultas, buscas e montagem da base de dados. Desta forma os dados podem ser encontrados e devidamente aproveitados de forma hábil. Os procedimentos básicos do *Big Data* não são completamente novos, onde grande parte deles são derivados dos processos do *KDD*, *Data Mining* e *CRMs* (*Customer Relationship Managements*) de utilizados para análise de dados históricos.

A tecnologia está cada dia mais avançada e integrada com o cotidiano, e com isto os dados são gerados por qualquer tipo de equipamento, porém estas informações ficam armazenadas de forma centralizada e controladas. Com toda esta quantidade de dados, não é possível efetuar a análise de forma manual, como era efetuado antigamente.

No *Big Data* existem cinco características que são desejadas para que uma aplicação de análise de dados seja enquadrada pelas normas definidas por esta tecnologia. Estas características são comumente conhecidas como os cinco Vs. Sendo que estas cinco características determinadas não necessitam estar devidamente representadas, pois dependendo das regras de negócio necessárias, algumas delas possam estar omitidos ou não implicar tanta importância para a exigência destas regras de negócio, podendo conseguir resultados significativos desta forma.

### 2.1.1 Volume

É necessária ou desejável um grande volume de dados para que o *Big Data* possa ser devidamente executados.

### 2.1.2 Velocidade

Na atualidade, o curto tempo de acesso às informações é essencial. E ter um acesso veloz às informações analisadas é essencial, e é exatamente para isto que esta característica provê.

### 2.1.3 Variedade

O grande volume de dados possui como consequência também a grande quantidade de dados. Muitos destes dados estão devidamente organizados, devido estarem



armazenados e organizados em bancos de dados estruturais como *SQL (Strutured Query Language)*, *Oracle*, *PostgreSQL* e *Oracle*, assim como dados não organizados como imagens, vídeos, áudios e documentos. Por isto é necessário saber como tratar estes dados como um todo. Um dado pode ser considerado inútil caso ele não esteja vinculado com algum outro conjunto de dados.

#### 2.1.4 Veracidade

Os dados devem ser verídicos, pois de nada adianta os dados estarem em grande volume, em uma imensa variedade e disponíveis em alta velocidade se estes dados não sejam confiáveis. É necessário que existam alguns processos que garantam que a maior consistência de dados possível.

#### 2.1.5 Valor

Uma classificação que define se os dados analisados e gerados por todos os processos anteriores são válidos e úteis para a análise em si.

Além do grande volume de dados e dos mais variados tipos de dados, as soluções de *Big Data* também precisam trabalhar com distribuição de processamento e elasticidade, ou seja, suportar o grande crescimento do volume dos dados, que crescem em pouco tempo. Por conta disto, os bancos relacionais não são os mais apropriados para compor este tipo de solução, já que os mesmos não são muito flexíveis.

## 2.2 Descoberta de Conhecimento em Banco de Dados

A descoberta de conhecimento em Bancos de Dados ou *KDD* consiste em muito mais do que apenas a mineração de dados. Conforme definido por Fayyad (1996), a **Descoberta de Conhecimento em Banco de Dados é um procedimento não trivial, automatizado e que visa buscar em grande bases de dados, novos conhecimentos e novos padrões de relacionamento de dados, que possuem algum tipo de produtividade e informação, caso os resultados sejam apresentados para um especialista da área.**

**O *KDD* é composto por seis etapas: seleção de dados, limpeza de dados, enriquecimento, transformação ou codificação dos dados, mineração de dados e apresentação dos resultados, conforme foi apontado Navathe (2011).** Podem existir algumas variações destes conceitos, apontadas por cada autor, por exemplo Fayyad (1996) determina que a divisão seja entre 5, onde as etapas de limpeza e enriquecimento dos dados sejam efetuados na mesma fase de pré-processamento, unificação devida a semelhança entre estas duas etapas.

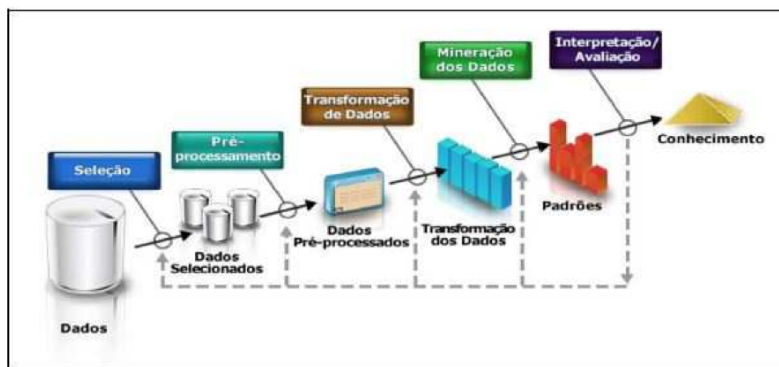


Figura 2.1: Etapas da Descoberta de Conhecimento em Banco de Dados

Como pode ser visto na figura 2.1, as etapas da Descoberta de Conhecimento estão sendo apresentadas conforme a definição de Fayyad. Existe um fluxo de atividades que se relacionam diretamente entre si e que deve ser obedecido para executar o processo de *KDD*, demonstrando que cada uma das etapas têm a sua devida importância e que uma só deve ter início ao término da etapa anterior.

### 2.2.1 Aplicações da Descoberta de Conhecimento

A descoberta de conhecimento é amplamente utilizada em área de marketing e de análise de informações de supermercado, porém a descoberta de conhecimento não é restrita à regras de negócio, de tal forma que ela tem sido utilizada nas mais distintas áreas. A tabela 2.1 apresenta algumas das diversas áreas em que a Descoberta de Conhecimento pode ser empregada, assim como uma descrição sucinta das utilidades em que se aplica.

Tabela 2.1: Utilização do KDD conforme área

Área	Utilização
Marketing	é utilizado para encontrar o perfil do consumidor e padrões de compras, tendo o objetivo de sugerir produtos de forma personalizada para os compradores.
Finanças	utilizado para efetuar a detecção de fraudes, que muitas vezes acabam passando por despercebido, devido terem um desvio muito sutil do padrão normal. Além de identificar clientes leais, padrões de uso do cartão de crédito e encontrar relações entre os mais diversos indicadores financeiros.
Medicina	utilizado para prever visitas, identificar terapias, avaliação clínica conforme os sintomas relacionados, além de identificar novos padrões de doenças.
Governo	Em instituições governamentais são utilizadas para melhorar as coletas de taxas, assim como na área financeira, utilizada para detectar fraudes.
Ciência	as técnicas de mineração de dados auxiliam os cientistas nas pesquisas, para encontrar padrões moleculares, dados genéticos, mudanças climáticas e gerando os relatórios e contribuindo para as conclusões valiosas em um tempo relativamente menor.
Qualidade	No Controle de Processos e Qualidade, traça os padrões de condições físicas e auxilia no planejamento estratégico relacionada a linha de produção.

### 2.2.2 Seleção de Dados

A Seleção de Dados é uma das principais etapas dos principais processos da Descoberta de Conhecimento em Banco de Dados, pois esta etapa é a responsável em

efetuar a seleção e filtragem dos dados que deverão ser avaliados pelas etapas seguintes do *KDD*. Os dados coletados nesta etapa refletem diretamente na qualidade do resultado final da análise da mineração de dados, de forma que estes dados são a principal fonte de informações da análise.

Nesta etapa são efetuadas as análises de levantamentos das variáveis e grupo de variáveis necessárias para efetuar a seleção e extração dos dados da base selecionada. Para a execução desta etapa normalmente são escritas aplicações que ficam responsabilizadas por efetuar a extração das bases de dados, das quais estas podem ser originadas de diferentes fontes de dados. Esta aplicação também fica responsabilizada por efetuar as filtrações necessárias, para não trazer para a análise dados que não devem ser analisados e por conta disso, é muito importante também definir quais são os tipos de informações e quais são os filtros que a aplicação deve ter implementada para que a qualidade dos dados seja mantida.

### 2.2.3 Pré-Processamento

Os dados levantados pela etapa anterior podem possuir alguns caracteres indesejados, alguns ruídos ou possuir informações incompletas. Este tipo de situação normalmente ocorre quando os dados são selecionados de bases heterogêneas, quando são originárias de bancos de dados que não possuem um devido tratamento no gerenciamento de dados ou quando provém da *internet*. Para conseguir resolver estes problemas com os dados, é necessário efetuar a limpeza dos dados para que não existam interferências durante a execução do algoritmo de mineração de dados, onde são removidos qualquer tipo de caracter indesejado e descartadas as mensagens que possuem informações incompletas ou algum outro tipo de ruído que não possa ser tratado.

O uso de *Data Warehouse* pode facilitar a tarefa de limpeza dos dados. Normalmente eles possuem uma organização e gerenciamento mais bem definido, que acabam mantendo os dados do banco de dados em um estado limpo. O uso dele contribui e facilita o processo de limpeza dos dados, para o qual será necessário menos tempo e esforço.

Nesta segunda parte, são efetuadas buscas secundárias para complementar estas informações selecionadas previamente. É bom lembrar que esta complementação é de informações que são incompletas por padrão, como por exemplo informações selecionadas dos cadastros de clientes. Neste cadastro existem apenas o logradouro, o número do endereço e o CEP (Código de Endereçamento Postal), mas o objetivo é buscar as vendas efetuadas por cidade/estado. Neste caso será executada a tarefa de enriquecimento para complementar estas informações faltantes para o problema proposto, carregando as informações da cidade e estado através do CEP.

### 2.2.4 Codificação de Dados

A codificação de dados tem como objetivo adequar os dados em uma estrutura e formatação necessários conforme é exigido pelo procedimento do algoritmo de mineração de dados. Além de adequar estes dados, cada algoritmo exige uma determinada estruturação para que o algoritmo consiga executar as análises corretamente, sem interferências ou erros durante a execução. Nesta etapa ocorre a conversão dos tipos de dados, para que o algoritmo de mineração de dados consiga efetuar a análise, assim como a filtragem de treinamento para o algoritmo de mineração de dados.

### 2.2.5 Mineração de Dados

A mineração de dados, segundo a definição de Fayyad(1996) envolve métodos e aplicações iterativas e interações de mineração de dados em particular. Ou seja, a mineração de dados é um processo automático ou semi-automático que visa explorar e analisar grandes bases de dados. Baseando-se nestas análises, poderão ser encontradas novos padrões e regras úteis e compreensíveis para o analista.

Este tipo de análise que é efetuado de forma automática pela mineração de dados não seria possível de ser efetuadas por humanos em procedimentos convencionais devido ao grande volume de dados que teria que ser processado, de mesma forma que a imensidão de relações que a análise exige também não seria humanamente possível. Para resolver este tipo de problema, foram desenvolvidos algoritmos que trabalham de forma automática e semi-autônomos, que conseguem retornar o mesmo resultado que o efetuado por um ser humano.

A interação e intervenção do analista ainda é exigida em alguns momentos para conseguir interpretar os resultados que a execução do algoritmo retornou, que é através da interpretação do analista que os padrões serão determinados como úteis ou não. Os algoritmos por si só efetuam o trabalho de encontrar as principais relações e possivelmente válidos, como foi descrito por (NAVEGA, 2002).

Como o processo de mineração de dados não possui um procedimento padronizado para resolver qualquer problema, existem diversos algoritmos que podem ser usados para cada tipo de problema proposto. Estes algoritmos são classificados em dois grandes grupos que variam de acordo com o tipo de conhecimento que se deseja extrair, conforme é apontado por Fayyad (1996). Estes grupos são Predição ou Atividades Preditivas e Descrição ou Atividades Descritivas, e dentro de cada um deles existem algoritmos para cada tipo de análise em específico.

### 2.2.6 Atividades Preditivas

As atividades preditivas ou simplesmente Predição, possuem como característica a análise de dados que visa encontrar formas de prever futuras situações ou aconte-

cimentos através da análise de dados históricos. Este tipo de atividade também é conhecida como aprendizagem de máquina.

Existem dois grupos de algoritmos dentro das Atividades que são conhecidos como Classificação e Regressão.

#### 2.2.6.1 Classificação

Na classificação, existe uma forma de aprendizagem, onde é mapeado um conjunto de registros em um conjunto de variáveis pré-definidas, estas que são denominadas Classes. Este tipo de função se enquadra em prever qual o tipo de classe que determinado registro se enquadra.

Muitos algoritmos se enquadram na classificação, mas os que mais se destacam são as Redes Neurais, *Back Propagation*, Classificadores *Bayesianos* e Algoritmos Genéticos.

#### 2.2.6.2 Regressão

A regressão consiste numa forma de aprendizagem semelhante à classificação, porém suas variáveis de análises devem ser valores numéricos contínuos presentes no banco de dados com valores reais. Neste tipo de algoritmo se utilizam os métodos de estatística e as Redes Neurais.

### 2.2.7 Atividades Descritivas

Segundo (FAYYAD; PIATETSKY-SHAPIO; PADHRAIC, 1996), as atividades descritivas ou Descrição possuem como característica fazer a busca de padrões utilizados como proposta para a forma de entendimento humano. Normalmente a descrição trabalha com grupos de informações, para facilitar a forma de localização dos padrões. Dentro da categoria de atividades descritivas se enquadram os grupos de algoritmos de Agrupamento, Regras de Associação e Numeração.

#### 2.2.7.1 Agrupamento

Conforme explicado por (GALVÃO; MARIN, 2009), a tarefa de Agrupamento ou *Clusterização* é utilizada para separar os registros de uma base de dados em subconjuntos. Este agrupamento é efetuado com as características semelhantes entre os registros, de forma que sejam distinguidos dos elementos de outros subconjuntos. Diferentemente da classificação, onde as variáveis são predefinidas, no agrupamento é necessário identificar automaticamente os grupos de dados que serão relacionados. Dentre os algoritmos disponíveis, os mais utilizados são os *K-Means*, *KModes*, *K-Prototypes*, *K-Medoids* e *Kohonen*.

### 2.2.7.2 Associação

Este tipo de tarefa consiste em identificar e descrever associações entre variáveis no mesmo item ou entre itens diferentes que ocorram simultaneamente de forma frequente no banco de dados, como apresenta (GALVÃO; MARIN, 2009). Esta pesquisa de associações entre variáveis e itens também é efetuada entre itens dentro de um espaço temporal. Dentre os algoritmos mais utilizados se destacam os algoritmos *Apriori* e *GSP* (*Generalized Sequential Patterns*).

### Numeração

Como é apresentado por (GALVÃO; MARIN, 2009), a numeração procura identificar e indicar características comuns entre um conjunto de dados. Esta tarefa é aplicada nos agrupamentos obtidos na tarefa de *clusterização*, sendo a Lógica Indutiva e Algoritmos Genéticos alguns tipos de algoritmos que são utilizados.

### 2.2.8 Exibição dos Resultados

Após a mineração de dados, é efetuada a exibição dos resultados, que por alguns autores é enquadrada no grupo de etapas de pós-processamento. Nesta etapa são efetuadas as seleções e ordenações das descobertas interessantes, efetuado o mapeamento de apresentação dos resultados obtidos e gerando relatório dos resultados. Existem muitas formas de apresentar estes dados, sendo através de gráficos, relatórios, tabelas ou qualquer outra forma de apresentação dos resultados.

## 2.3 Redes Sociais

A vida das pessoas é cercada por uma série de relações pessoais que são moldadas durante toda vida, iniciando no ambiente familiar, seguindo na escola e comunidade e por fim no ambiente de trabalho. As relações pessoais seguem se desenvolvendo e fortalecendo a esfera social de forma que faz com que a mesma fique organizada em uma estruturação de rede, conforme aponta (TOMÁEL; ALCARÁ; CHIARA, 2005). Por causa desta organização de rede, o termo "Redes Sociais" foi utilizado na Antropologia social por volta da década de 40.

Para avaliar estas relações e o grau de separação entre as pessoas, Milgram realizou um teste enviando uma determinada quantidade de cartas à várias pessoas selecionadas aleatoriamente, onde tinha-se o objetivo de redirecioná-las a um alvo específico. Caso estas pessoas não conhecessem o alvo, deveriam enviar as cartas para alguém que julgassem estar mais próximo a ele. Com este experimento, Milgram conseguiu avaliar que das cartas que chegaram ao seu alvo, a grande maioria das cartas haviam passado por uma pequena quantidade de pessoas, sugerindo assim que as pessoas estariam a poucos graus de separação das outras, consistindo assim

o que é chamado de fenômeno do mundo pequeno.

De outra forma, Barabási demonstrou que as redes sociais são formadas de uma forma dinâmica, seguindo um padrão de estruturação, apontando que redes ricas ficam mais ricas”, ou seja, quanto maior for o número de conexões que uma ator possuir, maiores serão as chances deste mesmo ator adquirir novas conexões. De modo que as redes sociais não são constituídas de relações igualitárias, ou seja, tendo a possibilidade de ter em média o mesmo número de conexões.

Avaliando as duas teorias, pode-se identificar que existem relações igualitárias para determinados grupos de pessoas, porém pode ter uma grande variação a medida que vai se mudando de grupos. Em outras palavras, pessoas de mesma classe social e que pertencem a um mesmo grupo tendem a possuir um número de conexões semelhantes, mas que se comparado à pessoas de um outro grupo social, a tendência do número de conexões é ser completamente diferentes.

As redes sociais online como conhecemos hoje é basicamente uma representação das redes sociais em um ambiente informatizado. Como a Internet possui as características de encurtar distâncias e estar presente em qualquer lugar a qualquer momento, as redes sociais *online* herdaram estas características, encurtando as distâncias entre as pessoas e elevando a proporção das conexões entre as pessoas nas redes sociais para níveis inimagináveis. Com a característica de encurtamento de distâncias, pessoas que estão a milhares de quilômetros e que mal se falavam por conta disso, através das redes sociais estas pessoas conseguem se comunicar facilmente. Além de todas estas características, as redes sociais possuem uma série de locais que podem ser utilizadas, pois como a *internet* é a sua base de funcionamento, elas estão disponíveis na rede de computadores, celulares, *tablets* e qualquer outro tipo de equipamento que tenha acesso às redes sociais.

Em uma rede social online, os usuários publicam seus perfis com as suas informações pessoais, preferências e características pessoais. Com todas estas definições, os usuários acabam se relacionando através da similaridade entre si, justamente por este comportamento, as redes sociais ainda prestam um serviço de sugestões de amizades com base nas semelhanças entre os perfis e dos perfis das conexões de um segundo nível, ou seja, as conexões de uma conexão de um usuário.

### 2.3.1 *Twitter*

O *Twitter* (TWITTER, 2014) atualmente é uma das principais redes sociais, na qual seus usuários compartilham e trocam mensagens entre si. As mensagens que alimentam a rede social é limitada em 160 caracteres, dos quais 20 deles são reservados para o usuário originário e os outros 140 caracteres são destinados à mensagem de texto. Estas mensagens, dentro da rede social, recebe o nome de *tweet* e nelas podem existir mensagens de texto puro, imagens, *links* para endereços

externos e vídeos curtos. Os vídeos e imagens postados são apresentados na linha de tempo da página principal do usuário logado, conforme pode ser visto na imagem 2.2, onde pode-se verificar a estrutura que um *tweet* possui, apresentando os textos, os links externos e imagens.



Figura 2.2: Exemplo de mensagem do *Twitter*

O *Twitter* possui o sistema de seguidor, onde se tem como objetivo um usuário acompanhar as publicações dos outros usuários mesmo que esses usuários não acompanhem as publicações deste usuário. Estruturação esta que não exige que os dois lados de uma conexão precisem ser efetuados para que a troca de informações entre elas possa existir. Ele possui também o sistema de listas, onde pode-se criar uma lista de pessoas que se deseja acompanhar, sem ter a necessidade de criar uma conexão de "seguidor" com este usuário. Além disso, existe as palavras-chave, que são chamadas de *hashtags*, onde as mensagens com estas *hashtags* possuem a facilidade de serem encontradas através de buscas. As *hashtags* mais utilizadas formam o *hanking* que é chamado de *Trend Topics*.

Na página inicial do usuário são exibidas as mensagens em tempo real, assim como as mensagens mais próximas do horário de acesso, seguindo uma classificação decrescente do horário de postagem. Dentre estas mensagens estão as publicações do próprio usuário e as as publicações dos usuários que são seguidos pelo mesmo. Estas mensagens podem ser repassadas para os próprios seguidores, através de *retweets*, assim como podem ser marcadas com estrela, podendo ser visualizada facilmente mais tarde. A rede social também tem disponível uma página de busca, que facilita localizar usuários e/ou assuntos específicos, que utilizam de palavras-chave e *hashtags* para agilizar a localização das informações.



Existem algumas outras páginas, onde é efetuado o gerenciamento dos seguidores e dos seguidos, bloqueio de um usuários indesejado, além das configurações do usuário, do tipo de apresentação da rede social, cores, etc. O perfil público do Twitter pode ser visualizado por qualquer um, estando ou não cadastrado na rede. Portanto, estas atualizações podem ser acompanhadas via RSS <sup>1</sup> ou pelo próprio site do Twitter.

## 2.4 Extração de Dados

O processo de Extração de Dados exige muito mais do que simplesmente fazer a coleta de informações. A extração de dados inicia desde o entendimento e definição das regras de negócio, que acabam evoluindo até a etapa de pré-processamento. Esta etapa compreende as atividades de extração de dados, limpeza e estruturação das informações.

Conforme apresentado por (FRANCISCHELLI, 2013), inicialmente é necessário entender as regras do problemas, para definir o que se deve ser analisado para poder retornar os resultados mais coerentes. Ou seja, é necessário definir os tipos de informações necessários para a análise. Os objetivos bem delineados acabam refletindo nos resultados que são obtidos ao final da mineração de dados.

Com os objetivos definidos, é necessário definir as hipóteses e questões que pretende-se responder. Estas definições influenciam na decisão nos termos de buscas e palavras chave para levantar as informações para a extração de dados em si.

Conforme descrito por (SOUSA, 2012), técnicas de extração que se baseiam em modelos linguísticos estão descritas em um determinado idioma. Existe o problema de que em uma mensagem curta, como um *tweet*, técnicas de identificação de partes de discursos de documentos-alvo não podem ser aplicados, pois podem não apresentar esta estrutura. Porém, devido à estrutura do *Twitter*, pode-se resolver alguns outros problemas que podem ser apresentados, como interlocutor e receptor da mensagem, locais dos atores das mensagens, quais os tópicos abordados, entre outros.

Com todas as definições já efetuadas, pode-se extrair os dados do Twitter. Como ele possui uma limitação de 750 mensagens por hora, como foi apresentado por (SILVA et al., 2010), a definição das palavras-chave serve de estratégia para melhor direcionar as informações coletadas para a análise.

Para auxiliar nesta tarefa, existem as *APIs* e bibliotecas das linguagens de desenvolvimento. Elas já possuem o tratamento necessário para conseguir acessar as informações das redes sociais. Algumas destas ferramentas já integram as atividades de limpeza dos dados, trazendo as informações já organizadas e tratadas.

---

<sup>1</sup>RSS é um formato que permite distribuir o conteúdo de um site de forma padronizada para leitores de notícias

## 2.5 Trabalhos Relacionados

Dentro da área abordada pelo *KDD*, existem diversos trabalhos e pesquisas efetuadas. Algumas delas estão relacionadas com as redes sociais, assim como outras estão relacionadas com o tratamento de informações do trânsito.

Aqui estão relacionados alguns trabalhos que acabam abordando algumas das etapas que estão sendo utilizadas para o desenvolvimento deste trabalho. Alguns dos processos definidos nestes trabalhos serviram de referência para a adoção de algumas estratégias para determinadas etapas.

### 2.5.1 Observatório do Trânsito

O trabalho relacionado ao Observatório do Trânsito, efetuado por (RIBEIRO JR. et al., 2012), remete uma grande preocupação em localizar as regiões que estão inseridas dentro das mensagens do *Twitter*. Em questão de similaridade, é o trabalho que possui uma maior quantidade de características semelhantes com o trabalho que está sendo desenvolvido.

Neste trabalho existe um grande foco no processo de seleção e enriquecimento dos dados, etapas que estão citados no Referencial Teórico. Nestas etapas, são selecionadas as mensagens que receberão o processo de mineração e análise. Em seguida está o processo de enriquecimento das informações que inicialmente não estão incompletas, mas que exigem um complemento. É **efetuada a complementação da localização citada na mensagem com o objetivo de encontrar o ponto exato de onde a informação está se relacionando.**

Para efetuar a complementação das informações coletadas, **são efetuadas novas coletas e análises de outras mensagens com o objetivo de encontrar a rua perpendicular mais próxima da localização citada.** Neste processo é efetuada uma busca de similaridade de mensagens que tratam de um mesmo assunto. Este processo é definido como localização por casamento exato.

Além da localização, foram utilizados perfis específicos de Minas Gerais, que atualizam constantemente as informações do trânsito local. Além disso, são apresentados alguns quadros de percentual de acurácia da execução do método proposto, conforme está apresentado na imagem 2.3, logo abaixo.

Table IV.

	Acerto Total (%)	Acerto Parcial (%)
Nada	61	61
Somente um bairro	60	78
Somente logradouro	74	76
Um logradouro + um bairro	90	90
Pelo menos um bairro	48	72
Pelo menos um logradouro	68	81

Figura 2.3: Análise de comprovação do método de análise das localizações das mensagens do Twitter

### 2.5.2 Descobrimento de eventos locais do Twitter

Santos (SANTOS, 2013) apresentou um trabalho com o objetivo de descobrir os eventos locais através de análise de informações do Twitter. Nele estão abordados os processos de georreferenciamento para localizar a origem das mensagens e os locais referenciados. O georreferenciamento é efetuado a partir localização da origem da mensagem, quando este não está disponível é analisado o local citado na mensagem. Também está sendo abordada a análise de quantidade de pessoas que estão neste determinado local, assim como é efetuada a análise de comportamento destas pessoas.

O método definido de georreferenciamento é necessário para conseguir determinar a localização das pessoas conforme está descrito nas mensagens do Twitter. A determinação da localização é importante para efetuar a filtragem de mensagens que estão relacionada a aquela localização, para assim conseguir analisar o comportamento das pessoas que estão naquela determinada localização.

Ao fim de determinar a localização, é necessário acompanhar o comportamento das pessoas em um determinado espaço de tempo. Esta análise, juntamente com a localização encontrada permite determinar a ocorrência de algum evento naquele determinado local e por um determinado período de tempo.

### 2.5.3 *Tweetmining*

O trabalho de *Tweetmining*, de Giulia Luan Santos de Sousa(SOUSA, 2012) possui o foco em fazer a análise de opinião que está disponível no *Twitter*. Para efetuar esta análise é necessário, além de todo o processo de descoberta de conhecimento em banco de dados, um intenso processo de mineração de textos para tornar estes dados informações relevantes.

A mineração e texto utilizada é o método de mineração *SVM (Support Vector Machine)*, onde serve como medida de relevância para os termos constantes no texto. Após é gerado o arquivo *SVMLib* para fins de análise através desta biblioteca.

Dentro da mineração de textos, são apresentadas algumas teorias e referências,

exemplificando os problemas que podem aparecer na mineração de textos e a forma de resolvê-los. Além da mineração de textos, a aplicação possui uma estrutura bem interessante que define desde a extração de informações da rede social até a análise dos dados selecionados.

#### 2.5.4 Comparativo de Semelhança

Relacionando os três trabalhos aqui citados, foi efetuado um comparativo para avaliação dos trabalhos que podem contribuir mais com o presente trabalho.

Tabela 2.2: Comparativo de Trabalhos Relacionados

Características	Trabalhos		
	Observatório do Trânsito	Descobrimento de Eventos	Tweetmining
<i>KDD</i>	X	X	X
Mineração de Textos		X	X
Informações do Trânsito	X		
Georreferenciamento	X	X	
Extração de Dados da Web	X	X	X
Twitter	X	X	X

Conforme demonstrada na tabela 2.2, os trabalhos do Observatório do Trânsito e o trabalho de Descobrimento de Eventos possuem o maior número de características de interesse do trabalho aqui realizado, possuindo ambos os trabalhos 5 características de interesse.

### 3 PROPOSTA DE SOLUÇÃO

Inicialmente para que um problema possa ser solucionado, é interessante entender qual é a sua origem, quais são suas causas e com estas informações montar o perfil do problema que está sendo abordado, para que seja possível determinar a estratégia necessária para solucioná-lo. No caso do presente trabalho, o principal problema são ocorrências do trânsito e portanto o objetivo principal é traçar os diversos perfis que o trânsito pode apresentar. Desta forma consegue-se delimitar quais os dias que possuem uma maior incidência destas ocorrências, quais os tipos de veículos que se envolvem em mais acidentes e outras coisas.

O trânsito é um ambiente que existem diversas variáveis e como consequência disso existem muitos tipos de ocorrências que podem acontecer em paralelo, e por isso é necessário determinar quais destas ocorrências precisam ser abordadas numa primeira instância. Para isto foram determinadas algumas hipóteses que servem como base para determinar qual será o tipo de informação coletada das redes sociais e quais serão as premissas necessárias para a análise.

Com as hipóteses definidas, é necessário definir qual será a origem dos dados. Para que a disponibilidade dos dados esteja garantida durante o processo de extração, foi utilizado o *Twitter* como base de dados. Esta rede social foi selecionada para servir como origem dos dados da análise, pois possui uma imensa quantidade de informações, onde parte delas está relacionada com o trânsito, permitindo uma maior facilidade na extração destes dados. Além de definir a origem dos dados também é necessário saber como e onde estes dados serão salvos, por isto existe a necessidade de definir qual o tipo do banco de dados, qual a estrutura que este banco vai adotar para que as informações fiquem armazenadas e organizadas, facilitando o momento de pesquisa e mineração dos dados.

Para facilitar a análise dos dados que foram coletados, limpadados e organizados, foram utilizados os processos definidos pela Descoberta de Conhecimento em Bancos de Dados. Com isso, foram definidos os algoritmos de mineração, além das ferramentas utilizadas para efetuar a análise dos dados.

#### 3.1 Hipóteses

Como o trânsito possui uma grande quantidade de variáveis que constantemente interagem entre si e geralmente ocasionam diversas ocorrências, é necessário deter-

minar quais destas ocorrências são as mais graves e que ocorrem com uma maior frequência. Dentre as ocorrências, foram escolhidas as mais comuns e que possuem uma maior ocorrência, como são os casos de engarrafamentos, lentidão, acidentes e atropelamentos. A ocorrência destes problema é de grande disponibilidade nas redes sociais, fato que contribui para a facilidade na extração destes dados.

Para entender qual o impacto que estes problemas têm sobre o trânsito e para conseguir delimitar o perfil do trânsito, foram determinadas algumas hipóteses a serem provadas.

- Quais os períodos de maior ocorrência destes problemas do trânsito? Como horários e dias.
- Quais os dias da semana que possuem a maior parte destas ocorrências?
- Quais os tipos de veículos que se envolvem em mais acidentes.

### 3.1.1 Engarrafamentos

A lentidão do trânsito e os engarrafamentos ocorrem por diversos motivos, onde um dos principais motivos é a situação que ocasiona este tipo de problema é a grande quantidade de veículos que estão disputando o espaço das ruas, porém elas não foram projetadas para receber tamanha densidade de veículos, o que acaba gerando os gargalos e consequentemente as grandes filas no trânsito. Os engarrafamentos podem ser gerados a partir de outras ocorrências do trânsito como acidentes, atropelamentos, entre outras causas.



Figura 3.1: Exemplo de informação de Trânsito Lento no Twitter

Na imagem 3.1 pode ser observado um exemplo de mensagem no *Twitter* demonstrando um problema de lentidão. Também pode ser visualizado que o problema de lentidão ocorreu devido a um acidente, portando as causas de lentidão e engarrafamentos podem ser originados de qualquer outro problema. Se houver qualquer ocorrência que acabe criando um gargalo nas vias do trânsito, acabará gerando lentidão e congestionamentos.

A base de origem é uma rede social, onde a estrutura dos dados não está organizada formalmente, de forma com que as mensagens não sejam apresentadas em um determinado padrão. Para resolver este problema, foram utilizados termos de busca sinônimos com os termos principais para conseguir referenciar a um mesmo assunto. Também foram definidos algumas das mais prováveis causas de engarrafamentos,

desde que não invadam os outros problemas abordados.

Dos termos relacionados ao engarrafamento, foram determinadas as seguintes palavras chave:

- Engarrafamentos
- Lentidão
- Trânsito Lento
- Transito Parado
- Bloqueios
- Obras

### 3.1.2 Acidentes

Como existe uma grande quantidade de veículos que estão dividindo o espaço das ruas, é esperado que existam conflitos entre eles, de forma que acidentes ocorram com uma certa frequência. Eles podem ser originados de vários fatores como a lentidão, falta de atenção dos motoristas ou até mesmo um outro acidente, ou até mesmo de todos estes fatores juntos. Como os acidentes possuem muitas variações, para efetuar as devidas buscas na rede social, foi utilizado da forma mais genérica possível para conseguir abranger a maior quantidade de informações foram definidas as palavras chave para manter esta característica de generalização.



Figura 3.2: Exemplo de informação de Acidente de Trânsito no Twitter

Como palavras-chave para este problema, foram definidos inicialmente poucos termos para que seja coletada a maior quantidade de informação possível. Foram definidos os seguintes termos de busca:

- Acidente
- Acidente de Trânsito

Além dos termos genéricos, também foram definidos os seguintes termos sinônimos que remetem ao mesmo assunto, tratando do mesmo tipo de informação. São eles:

- Batida
- Pechada
- Engavetamento
- Capotamento

### 3.1.3 Atropelamentos

Os atropelamentos se enquadram como acidentes de trânsito, porém devido às suas características específicas, foi definido que este tipo de ocorrência seria tratado de forma separada. De mesma forma que os acidentes, são ocorrências que possuem uma frequência considerável.

Os atropelamentos se enquadram na categoria de acidentes de trânsito, porém devido à suas características específicas, serão tratados separadamente. Os atropelamentos assim como os acidentes são problemas que acabam ocorrendo frequentemente no trânsito e normalmente envolvem algum veículo e um pedestre. Também existe uma determinada classificação dos atropelamentos que segue da mesma forma que para os acidentes, variando de lesões leves, lesões graves e atropelamentos com vítimas fatais.

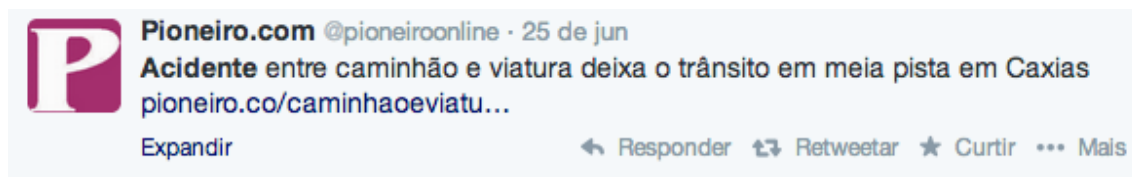


Figura 3.3: Exemplo de informação de Atropelamento no *Twitter*

A imagem 3.3 serve como exemplo de uma mensagem postada no *Twitter* referenciando uma ocorrência de atropelamento. Algumas mensagens relacionadas com o trânsito possuem um certo padrão mostrando a ocorrência e a descrição do local aproximado do local que ocorreu tal situação, porém elas não são muito frequentes entre os usuários normais.

## 3.2 Rede Social Escolhida

Existem diversas redes sociais na *internet* e sabe-se que elas movimentam diariamente milhões de mensagens, independente do assunto. Dentre elas existem alguns tipos de redes sociais onde tem as priorizam o uso de texto puro, como é o caso do *Twitter*; as que priorizam o compartilhamento de imagens como é o caso do *Instagram*; e redes sociais que possuem o foco nas informações, onde podem trabalhar com texto, imagens, vídeos, entre outros como é o caso do *Facebook*. Existem também algumas redes sociais que fogem um pouco da relação social e tratam mais sobre a plataforma de mapas e compartilha as informações e condições do trânsito entre os contatos mais próximos, como é o caso do *Waze*.

O *Waze* é um tipo de rede social que permite atualizar as informações do trânsito de forma fácil e rápida. No aplicativo móvel existem opções pré-determinadas de ocorrências do trânsito, facilitando assim o seu uso no trânsito, mesmo que não



seja recomendado utilizar aparelhos móveis enquanto se esteja dirigindo. Estas atualizações são efetuadas do local de onde se está enviando a informação. Além de monitorar as atualizações compartilhadas pelos usuários, é efetuado o monitoramento da velocidade de deslocamento dos usuários nas ruas rodovia, e utilizando isto para alertar sobre possível lentidão ou engarrafamento no trânsito.

Devido a algumas características apresentadas pelo *Twitter*, esta foi a rede social escolhida para base de dados para análise das informações do trânsito: o fato dela não ser uma rede social direcionada exclusivamente para o trânsito, a disponibilidade de *APIs* para extração de informações, faz com que o processo de análise das suas informações relacionadas ao trânsito acabe se tornando interessante.

As informações provenientes de perfis pessoais e de perfis específicos responsáveis em atualizar as informações do trânsito via rede social. Este último tipo de perfil, é muito encontrado nas capitais e nas principais cidades do Brasil. Neles são disponibilizadas as informações do trânsito, obras, acidentes e estatísticas em geral.

O *Twitter* possui uma área exclusiva para desenvolvedores, desta forma disponibiliza alguns serviços e ferramentas para auxiliar no desenvolvimento de aplicações. Para tal, basta cadastrar uma aplicação na rede social, que a mesma já permite que sejam definidas as permissões de acesso e as chaves necessárias para acessar as informações da rede social. Estas regras não são tão restritas quanto os termos de uso do *Facebook*, mas ainda assim garantem uma boa segurança e privacidade para os usuários. Em questão dos dados para os aplicativos que se integram com o *Twitter*, este possui uma série de *APIs* que permitem acessar estas informações.

Juntamente com a grande quantidade de informações, a estrutura e a organização do *Twitter* acabaram influenciando na escolha desta como a rede social a ser utilizada para servir como fonte de dados. A estrutura possui um certo tipo de padronização das informações, o que acaba facilitando o processo de limpeza e identificação das informações presentes nas mensagens coletadas. Assim como existe toda uma estrutura facilitada no processo de adquirir informações.

### 3.2.1 *APIs* do *Twitter*

Dentro da área de desenvolvedor do *Twitter*, existe muita documentação e as principais *APIs* que permitem acessar as informações da rede social. Dentre estas *APIs*, se destacam o *Twitter API Streaming* e o *Twitter Rest API*.

Conforme apresentado pelo próprio *Twitter* (DEVELOPERS, 2014), o *Twitter API Streaming* funciona capturando as mensagens em tempo real, ou seja, assim que as mensagens vão sendo postadas no *Twitter*, as mesmas são capturadas e disponibilizadas por esta API. Esta API requer que uma conexão *HTTP* (*Hypertext Transfer Protocol*) se mantenha sempre aberta durante o processo para que a captura funcione corretamente. As informações e as mensagens que são capturadas por esta

*API* estão relacionadas a usuários em específico.

Existe também a documentação da *API Twitter Rest API*, que apresenta uma maior disponibilidade de recursos, das quais é possível capturar as informações de um determinado usuário da *timeline* de um usuário, das mensagens que foram retuitadas por este usuário, mensagens coletadas a partir de buscas e muitas outras ferramentas de seleção interessantes. Além de capturar informações do Twitter, ela permite adicionar informações no Twitter como criar novas mensagens, *retuitar* alguma mensagem em específico, criar mensagens com imagens, etc.

Diferentemente da *API Twitter API Streaming*, esta *API* não faz a coleta das informações em tempo real mas sim através de requisições em massa. Para efetuar uma requisição de informações da rede social, é aberta uma conexão *HTTP* que ao término da coleta, esta requisição *HTTP* é encerrada.

Para coleta de informações, ambas as *APIs* se destacam pois permitem coletar as informações de distintas formas. O *Twitter REST API* consegue processar uma maior quantidade de dados de uma só vez e não necessita que uma conexão *HTTP* fique continuamente ativa, porém existe a limitação onde não se pode coletar mais que 700 mensagens por hora. O *Twitter API Stream*, assim como o *REST API*, permite efetuar a filtragem por palavras-chave, por localização, etc. Sabe-se que existe a característica de que esta *API* necessita que uma conexão fique sempre ativa para receber cada uma das mensagens no exato momento em que elas são postadas.

### 3.2.2 Bibliotecas do Twitter

Assim como as muitas *APIs* que o *Twitter* disponibiliza, existe uma série de bibliotecas para serem utilizadas nas mais diversas linguagens de desenvolvimento, que varia entre *Java*, *Python*, *Ruby*, *PHP*, *Perl* e outras linguagens. Estas bibliotecas permitem conectar a linguagem de programação com as *APIs*, e consequentemente efetuar a coleta das informações da mesma.

Primeiramente é essencial escolher a linguagem de programação que deve ser utilizado no desenvolvimento, e para tal é necessário avaliar uma série de variáveis. Dentre as que mais se destacam a complexidade de desenvolvimento exigida pela linguagem, a quantidade de documentação relacionada com a linguagem e principalmente o conforto que o desenvolvedor possui ao programar.

Seguindo a partir das características apontadas no parágrafo anterior, a linguagem que melhor se enquadra para desenvolvimento da aplicação é o *PHP* (*PHP: Hypertext Processor*). Como a proposta do protótipo de extrator é simples e não deve possuir uma estrutura complexa, será seguido o modelo *MVC* (*Model-View-Controller*), com algumas telas disponíveis com as opções para cadastro e exportação dos dados do banco.

### 3.3 Escolha do Algoritmo de Mineração

Definir qual o tipo de Algoritmo de Mineração de dados é primordial para o desenvolvimento de um trabalho que implementa e utiliza os conceitos e técnicas de *KDD*. Esta definição implica diretamente nos resultados que são obtidos ao término da análise.

Como o tipo algoritmo implica diretamente no resultado final, escolhê-lo é muito importante além de ser um processo bem complexo. Antes de ser definido o algoritmo é necessário que os objetivos e o tipo de respostas desejadas já estejam bem delineados, caso contrário os resultados possam ser inesperados ou inconsistentes. Isto ocorre porque cada algoritmo de mineração está direcionado para resolver determinado tipo problema.

Para conseguir obter os resultados necessários para a resolução do objetivo proposto, foram delimitadas algumas hipóteses que estão direcionadas em traçar o perfil do trânsito, relacionando as suas informações com as informações relativas a outros ambientes externos. Estes ambientes externos não estão diretamente relacionados com o trânsito, mas que alteram as suas condições de uso. Dentre eles o mais importante é a condição climática do local.

Verificando cuidadosamente cada uma das hipóteses e verificada a proposta de cada um dos tipos de algoritmos de dados, um algoritmo que se enquadra bem neste tipo de problema proposto é o algoritmo de regras de associação. Nele existe a proposta de relacionar as informações, podendo retornar as estatísticas de maiores ocorrências em determinadas situações. Este tipo de algoritmo é amplamente utilizado na mineração de informações que estão relacionadas ao comércio para traçar o perfil das compras, indicando qual a melhor estratégia de dispor os produtos nas prateleiras.

Nas hipóteses definidas existe o desafio de associar estas ocorrências do trânsito que foram coletadas do *Twitter*, com os grupos de informações pertinentes aos ambientes externos que foram determinados. Devido à esta condição proposta pelas hipóteses em traçar estes perfis, e como este tipo de algoritmo trata de melhor forma este tipo de condição, é o tipo de algoritmo que traz uma melhor gama de informações pertinentes aos determinados perfis que tendem a serem traçados. Assim, como comentado no parágrafo anterior, o problema proposto é muito semelhante com a mineração utilizada em supermercados, devido possuir uma grande quantidade de ocorrências possíveis e que estas podem aparecer junto à outras ocorrências.

Dentre os algoritmos de mineração de regras de associação, estão cotados para serem utilizados no presente trabalho os algoritmos *Apriori* e o *Tertius*.

### 3.3.1 Algoritmo Apriori

Conforme apresentado por (YABING, 2013), o algoritmo *Apriori* é um algoritmo de regras de associações que busca os item que tiveram maior frequência durante a mineração. Em seguida, o algoritmo cruza as informações de um registro com as informações das demais iterações da base de dados de forma que este cruzamento vai alimentando o algoritmo e vai estabelecendo as regras de associações entre os itens de cada registro processado. Ao término, o algoritmo gera todas as regras de associação e efetua a classificação das regras de associação com a maior incidência de grau de confiança.

### 3.3.2 Algoritmo Tertius

Segundo (FLACH; MARALDI; RIGUZZI, 2006), *Tertius* é um algoritmo de extração de regras de associação que utiliza a lógica de programação por indução de confirmação, que possui a finalidade de aumentar a confiabilidade da função de avaliação. Desta forma, os registros que serão descartados pelo algoritmo receberão um valor mínimo para análise, enquanto que os demais valores que serão considerados pelo algoritmo recebem um valor maior para a análise.

Para que o *Tertius* consiga efetuar a análise, o algoritmo precisa preencher duas tabelas de contingência para cada cláusula, uma para os valores observados e outra para os valores esperados. Ao explorar estas tabelas, o *Tertius* efetua a busca a fundo com o objetivo de melhorar a confiabilidade da função de avaliação, dando uma nota mínima para as cláusulas descartadas e uma nota mais elevada para as cláusulas restantes. Por final, é efetuado um cruzamento com as cláusulas com melhores pontuações e montadas em regras de associações.

## → 3.4 Ferramentas de Mineração de Dados

Para executar o processo da etapa de mineração de dados, foi selecionado o software *Weka*, pois o mesmo conta com uma série de algoritmos de mineração de dados já implementados. Além disso, este software é multi-plataforma, o que permite que o mesmo consiga ser rodado em praticamente qualquer sistema operacional, assim como este software é amplamente utilizado em ambientes acadêmicos, informação que pode ser visualizada na grande quantidade de trabalhos que foram realizados utilizando-o como ferramenta de mineração e análise dos dados.

## 3.5 Organização das Informações

Inicialmente, os dados são salvos em banco de dados do tipo *NoSQL*, devido este tipo de banco de dados ser mais indicado no processo de armazenamento de grande

bases de dados assim como são mais facilmente adaptáveis para informações não formalmente estruturadas. Eles também são indicados no processo de armazenar as informações que receberão um processo de mineração de dados.

Como as *APIs* do *Twitter* retornam os resultados normalmente nos formatos *JSON* (*Java Script Object Notation*) e *XML* (*Extensible Markup Language*), o banco de dados *NoSQL* consegue armazenar este tipo de informação, e o banco de dados *MongoDB* está devidamente preparado para trabalhar com *JSON*. Os dados provenientes do *Twitter* serão armazenados neste tipo de banco de dados, após receberem o devido tratamento de limpeza dos dados.

No processo de limpeza, será removida toda a acentuação, cedilha e caracteres especiais que possam comprometer a execução do algoritmo de mineração de dados. Mesmo com a execução deste processo, a estrutura dos dados será mantida, para garantir mesmo comportamento que como se os dados estivessem originais.

Logo em seguida, será criado o arquivo de aprendizagem de máquina, contendo o a estrutura e os parâmetros para montar as regras de associação necessárias para a devida execução do algoritmo de mineração de dados e análise das informações. O software *Weka* consegue trabalhar com arquivos do tipo *ARFF* (*Attribute Relationship File Format*), onde ficam organizadas as informações em duas áreas: regras de estruturação e treinamento. Além disso, o software *WEKA* consegue trabalhar com os arquivos do tipo *CSV* (*Comma Separated Values*), onde as colunas representam a estrutura do arquivo e as linhas representam os dados de treinamento do arquivo.

Para que as informações estejam disponíveis no momento em que for utilizado o algoritmo de mineração de dados, será desenvolvido o extrator de dados. Este extrator de dados vai compreender todas estas etapas listadas acima, que engloba desde a seleção dos dados via *API* até o seu armazenamento no banco de dados, com todos os dados devidamente limpos. Ao término da coleta das informações, será gerado um arquivo na estrutura do tipo *CSV*, que é mais simples e fácil de ser montado.

### 3.6 Extração de Dados

O processo de extração é efetuado utilizando dos dois tipos de *APIs* disponibilizados do *Twitter*. Para garantir que os dados consigam ser coletados corretamente, ao iniciar o processo de coleta será utilizado o processo *REST API* onde são coletadas as postagens anteriores ao início do processamento de coleta. Na sequência do processo, a coleta normal é efetuada utilizando a coleta via *Stream*, no caso usando a *Twitter API Stream* que efetua a coleta dos dados conforme os mesmos vão sendo postados na rede social.

Como está determinado na documentação do *Twitter*, a *API REST* está limitada em efetuar no máximo 750 mensagens por hora para uma mesma aplicação, por este motivo o processo de coleta que envolve esta *API* será executado durante a inicialização da aplicação e repetindo a cada 1 hora para superar esta limitação. Além disso foram definidas algumas estratégias para abranger mais mensagens da rede social ao utilizar esta *API*, no caso foi definido um ponto de geo-localização para servir de ponto central da origem das mensagens e baseando-se deste ponto são coletadas as mensagens em um raio de 300 quilômetros. Complementando e finalizando esta parte da lógica da aplicação, foram definidas as palavras-chave para servirem como filtragem e de parâmetros para a extração dos dados.

A *API via Stream* possui um funcionamento diferenciado da outra *API* já mencionada, onde esta não possui a limitação e efetua a coleta das mensagens assim que as mesmas vão sendo postadas na rede social. Para a utilização desta *API* foram definidos dois pares de coordenadas para delimitar a coordenada inicial e a coordenada final da área em que as mensagens serão coletadas. Além disso estão sendo utilizadas os mesmos termos de busca e as mesmas regras para efetuar a filtragem das mensagens.

Como já foi apresentado na definição das hipóteses e na apresentação dos problemas específicos do trânsito que foram abordados, as mensagens provenientes de órgãos de compartilhamento de informações do trânsito possuem um certo padrão, que auxilia na extração de cada uma das condições de análise acima. Para conseguir encontrar cada uma das condições, será efetuada a mineração de textos e busca de padrões e ocorrências de palavras-chave. Para isto será definido um dicionário de palavras-chave, que será gerado a partir dos dados pré-selecionados.

### 3.7 Proposta de Protótipo de Pós-processamento

Por fim de toda a extração e análise dos dados, outra etapa muito importante do *KDD* é a apresentação dos resultados. Assim como qualquer outra etapa da Descoberta de Conhecimento, a apresentação dos resultados é bem complexa de ser efetuada. Como o foco do presente trabalho não é tratar plenamente da apresentação mas sim da mineração e extração dos dados, esta etapa será abordada de forma bem superficial.

Para apresentar os resultados obtidos pela mineração de dados, serão gerados pequenos relatórios apresentados na tela do computador. Nestes relatórios serão demonstrados apenas as informações principais e mais importantes, como os perfis encontrados de cada um dos períodos selecionados.

## 4 IMPLEMENTAÇÃO

Para iniciar a aplicação dos processos do *KDD*, é necessário seguir o fluxo de etapas exigidas por este método de descoberta de conhecimento. Dentre estas etapas, o ponto inicial é extrair e coletar as informações que serão utilizadas para efetuar a análise dos dados. Para tal, foi desenvolvido um protótipo de extrator de dados que possui a principal funcionalidade de coletar as informações relacionadas a determinados assuntos e armazenar os mesmos em um banco de dados *NoSQL*, para facilitar as posteriores consultas e servir como base de dados históricos.

Este protótipo, além de coletar e armazenar as mensagens extraídas do *Twitter*, recebeu algumas outras funcionalidades de visam otimizar o processo de busca. Para garantir a filtragem dos dados por localização e por termos de busca, foram desenvolvidos no protótipo de extrator de dados um cadastro de coordenadas, onde são informadas as coordenadas da localização inicial e final da origem das mensagens que se deseja efetuar a coleta, assim como para a coordenada central de busca de informações, que é utilizada pela *API REST*. Para efetuar a filtragem dos dados foi desenvolvido um cadastro de termos de busca onde o usuário pode efetuar o cadastro dos termos que se deseja trabalhar na análise, funcionando como uma forma de gerenciador de palavras-chave.

Como a principal fonte de dados é um banco de dados cuja interação é através de uma página WEB, como é o caso do *Twitter*, foi definido que o extrator também segue este conceito de plataforma e portanto foi desenvolvida uma aplicação WEB para que a integração entre a fonte de dados e o extrator fosse o mais simples possível. Para o desenvolvimento das regras de negócio do protótipo, foi utilizada a linguagem *PHP*, para a apresentação dos dados e troca de informações entre as regras de negócios. A parte de apresentação da interface da aplicação foi desenvolvido utilizando *JavaScript*, *jQuery* e *Ajax*, e para armazenar os dados foi utilizado o banco de dados *MongoDB*.

### 4.1 Banco de Dados Mongo DB

O *Mongo DB* é um banco não relacional, orientado a documentos e que pode ter numa mesma entidade registros com estruturas diferentes, ou seja, as entidades não exigem uma estruturação formal, coisa que ocorre com os bancos relacionais. Devido a esta flexibilidade do banco de dados, o modo como ele trabalha com os

dados e com as entidades é bem mais simples. Desta forma o banco de dados não necessita ter a estrutura das entidades definidas previamente, sendo necessário apenas popular o banco de dados, já que o mesmo se encarrega de efetuar esta tarefa para desenvolvedor. Esta estrutura implica num retorno mais ágil quando efetuada a seleção de um grande volume de dados.

Como o MongoDB se baseia em documentos, estes podem possuir uma estrutura muito similar aos documentos *XML* e aos documentos *JSON*, o que facilita muito na forma de trabalho com este banco utilizando aplicações WEB. A estrutura de dados que este banco utiliza é o JSON, o que o torna compatível com praticamente qualquer aplicação WEB, sem ter a necessidade de converter dados ou montar arquivos do tipo *XML*. Como as *APIs* do Twitter retornam as requisições no formato JSON, a sua integração foi feita com o MongoDB.

#### 4.1.1 Instalação

Este banco de dados preza pela facilidade de uso, característica que aparece desde o processo de instalação do mesmo, pois a instalação dele é tão simples quanto o seu uso. Se for efetuada a comparação da instalação deste banco de dados com os demais de estrutura relacional, o processo é muito mais simplificado e que não leva muito tempo e nem requer muitas configurações para se ter um banco de dados rodando.

O Sistema Operacional utilizado no desenvolvimento e testes foi um Mac OS X Yosemite. Para efetuar a instalação, foi necessário efetuar o download do programa de sua página oficial (<http://www.mongodb.org/downloads>) e descompactar o conteúdo do arquivo em um determinado diretório. No caso foi utilizado o diretório */var/mongodb* onde ficam os aplicativos de linhas de comando do sistema operacional em questão, e por fim foi necessário adicionar o caminho deste diretório ao *PATH* do sistema operacional.

A mesma facilidade que está em efetuar a instalação do banco de dados neste sistema operacional está também em iniciar o banco de dados. Para tal é necessário executar o comando de inicialização do banco de dados, setando o local onde o banco de dados irá armazenar os arquivos referentes ao banco de dados. O comando necessário para iniciar o banco de dados é: *mongod -dbpath nomedobanco*. Este comando efetua a criação do banco de dados e sobe o serviço de servidor na porta 27017.

Para que a aplicação funcione corretamente, é necessário criar o usuário *edipo* com a senha *edipo*. Para isto é necessário utilizar os comandos: *mongo*: comando para abrir a conexão com o banco de dados e ficar disponível o acesso de comandos do banco *use admin*: seleciona a base de dados "admin" *db.createUser(user: "edipo", pwd: "edipo", roles: [role: "userAdminAnyDatabase", db: "admin"])*: efe-



tua a criação de um novo usuário

#### 4.1.2 Modelo de Banco de Dados

O MongoDB, assim como qualquer outro banco de dados, armazena todos os dados necessários em uma série de arquivos, que ficam armazenados no diretório que foi definido ao iniciar o processo do banco de dados. Dentre estes arquivos, um deles que possui extensão *.0* que armazena os dados do banco de dados em si, enquanto que o arquivo com extensão *ns* armazena os índices de busca do banco de dados.

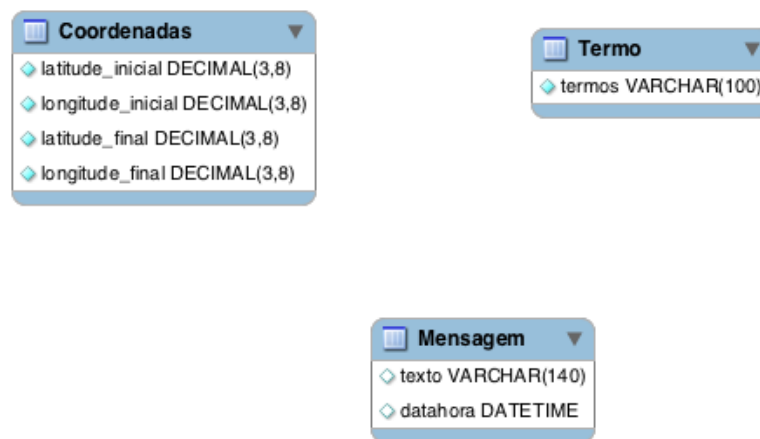


Figura 4.1: Modelo de banco de dados

O sistema de extração de informações do Twitter é composto por três entidades de comportamento individual, como é apresentado na imagem 4.1, porém devido à implementação das regras de negócio da aplicação estas entidades conseguem trabalhar harmonicamente, conseguindo trabalhar com uma grande quantidade de dados em um curto espaço de tempo. Para armazenar todas as informações pertinentes ao sistema como um todo, foram definidas as seguintes entidades que compõem o banco de dados, que estão apresentadas na lista a seguir.

- Termos de Busca: Termos de buscas que são utilizados para filtragem das informações do Twitter
- Coordenadas: Coordenadas de referência para definição do ponto central da busca
- Mensagens: Armazena as mensagens carregadas e filtradas pelo processo de extração dos dados.

##### 4.1.2.1 Termos de Busca

A entidade Termos de Busca é encarregada de armazenar todas as palavras-chave e conjunto de palavras, que posteriormente são utilizadas para efetuar a filtragem

das mensagens. Esta entidade é composta simplesmente de um *ID*, que é atribuído pelo próprio banco de dados e pelo campo *Termos*, que armazena o conteúdo de cada um dos termos de busca cadastrados.

#### 4.1.2.2 *Coordenadas*

A entidade Coordenadas é encarregada de armazenar os dois pares de coordenadas, que quando for efetuada a extração dos dados do Twitter servem para montar a área retangular das coordenadas de origem das mensagens através da localização. Esta entidade é composta de um *ID* e quatro campos que armazenam as Latitudes e Longitudes iniciais e finais.

#### 4.1.2.3 *Mensagens*

A entidade Mensagens é encarregada de armazenar o conteúdo das mensagens, que envolve o texto da mensagem já devidamente tratado, a data e a hora que esta mensagem foi postada. O conteúdo da mensagem é derivada da filtragem efetuada pelos termos de busca e pelas coordenadas.

## 4.2 *API Twitter*

As *APIs* de coleta de informações do *Twitter* possuem a finalidade de facilitar o processo de coleta das informações da rede social, sendo que exige uma determinada curva de aprendizagem para que seja possível efetuar algumas atividades entre a *API* e a rede social. Dentre as exigências necessárias, a principal é na forma de como a *API* funciona para conseguir recuperar as informações, como é o caso da localização correta da rede social. Foram utilizados os dois tipos de *APIs* disponibilizados pelo *Twitter* para que fosse possível recuperar a maior quantidade de mensagens possíveis. Dentre os parâmetros utilizados para filtragem, segue a lista.

- Busca: o sistema de busca da *API* efetua as buscas a partir dos termos de busca determinados nos parâmetros da *API*.
- GeoLocalização: o sistema de busca que envolve o sistema de GeoLocalização pode ser efetuado de diversas formas. Existe a busca por proximidade de um determinado ponto, trabalhando sobre um raio de de busca de uma determinada quantidade de quilômetros do ponto central definido.
- Streaming: A busca por streaming é efetuada em tempo real, onde a aplicação abre uma janela de conexão e as mensagens que são postadas na rede social neste intervalo serão coletadas.

#### 4.2.1 Aplicação do *Twitter*

Uma aplicação que trabalha com as informações do *Twitter* exige alguns requisitos para que a rede social permita o acesso dos dados. Dentre eles, uma aplicação deve estar cadastrada no ambiente de desenvolvimento da mesma para que o ambiente gere as identificações e *tokens* de acesso.

Ao criar uma aplicação na página de desenvolvimento do *Twitter*, existe uma série de parâmetros a serem informados, partindo do nome e descrição da aplicação, e também para criar as chaves necessárias para acessar as informações do *Twitter*. Além das chaves e dos *Tokens* de acesso à rede social existe a opção que delimitar qual será o tipo de acesso que a aplicação deve possuir em relação à rede social, variando entre acesso de somente leitura até gravação de informações na rede. Como o extrator desenvolvido necessita apenas de leitura das informações do *Twitter*, as permissões de acesso foram definidas desta forma.

Dentre os *Tokens* criados pelo *Twitter*, são necessários 2 tipos de chaves de autenticação e 2 tipos de *Tokens*.

- *Consumer Key*: é a chave de identificação da aplicação para coletar dados da rede social
- *Consumer Secret*: é a chave de autenticação da aplicação para liberar o processo de coleta de dados
- *oAuth Token*: é o Tolken de identificação da aplicação para gravar dados na rede social
- *oAuth Token Secret*: é o Tolken de autenticação para gravar as informações na rede social

#### 4.2.2 Parâmetros para Busca

Sabendo que o *Twitter* possibilita diversas formas para efetuar as buscas das informações, no desenvolvimento da aplicação foi utilizado dois tipos de busca para que os dados fossem filtrados e coletados conforme desejado. Os dados deviam ser originados das proximidades da serra gaúcha e o assunto deveria estar relacionado com o trânsito, e para isto foram utilizadas regras que suprem exatamente estas necessidades.

Ao utilizar a *API REST* a coleta de informações foi utilizada com a filtragem por localização através do parâmetro *geocode*, onde foi passadas as coordenadas da Praça Dante Alighieri, seguido do parâmetro que indica o raio de busca que foi efetuada a coleta. Desta forma, a aplicação efetuará a coleta apenas das mensagens que estão dentro deste raio de busca que está definido em 300km, abrangendo a serra gaúcha, a capital e algumas cidades de Santa Catarina próximas à divisa com o Rio Grande do Sul. Em conjunto, foi utilizada a busca por termos de busca foi utilizada através

do parâmetro de busca  $q$ , que define quais são as palavras que devem ser buscadas na rede social.

Como o *Twitter* limita o processo de coleta de dados em torno de 750 mensagens durante um período de 1 hora, foi desenvolvido um processo que repete todo o processo de coleta das informações a cada hora. Assim os dados postados são coletados a cada 750 mensagens, visando coletar a maior quantidade de mensagens possível.

Durante o intervalo entre as coletas pelo processo anterior, entra em ação o processo de coleta através da *API Stream*, que utiliza o sistema de localização um pouco diferente. Para determinar a área de filtragem, é necessário utilizar dois pares de coordenadas que representam o vértice superior esquerdo e o vértice inferior direito de um retângulo que delimita a área onde será coletada. A área definida é de toda a serra gaúcha e de parte de região metropolitana, ou como for definido no cadastro.

### 4.3 Desenvolvimento do Protótipo de Extrator de Dados

O Extrator de Dados desenvolvido e utilizado no presente trabalho é um protótipo que possui a principal funcionalidade de coletar as informações da rede social e armazenar estes dados no banco de dados. Para compor o funcionamento da aplicação, existem outras sub-processos que implicam diretamente no funcionamento principal que permitem o cadastramento dos termos de busca e das coordenada. Para complementar o processo da aplicação e permitir gerar os arquivos para análise pelo *Weka*, foi desenvolvida uma opção que permite exportar as informações para arquivos *CSV*.

A aplicação está estruturada nas seguintes funcionalidades.

- Extrator
- Termos de Busca
- Coordenadas
- Exportação de Dados

#### 4.3.1 Tecnologias aplicadas

Dentre as tecnologias utilizadas no desenvolvimento da aplicação, foi utilizado o *PHP* como linguagem de programação principal devido à sua facilidade de aprendizado, simplicidade de desenvolvimento, grande quantidade de documentação e atividade da comunidade no desenvolvimento. Outros fatores importantes que influenciaram na escolha desta linguagem foi a grande quantidade de *APIs* do *Twitter* disponíveis para esta linguagem, independente do tipo da *API* selecionada.

Algumas outras rotinas foram desenvolvidas em *JavaScript*, que se apresenta em processos de controle de tela e comportamento da página. A partir dele e em

conjunto estão sendo utilizados o *Ajax* e *JQuery* que auxiliam no controle *Client-Side*. Para armazenar as informações para pesquisas posteriores, foi utilizado o banco de dados não relacional *MongoDB*, que possui uma facilidade de integração e funcionamento com o PHP e com as *APIs* do *Twitter*.

#### 4.3.2 Metodologias aplicadas

A aplicação está estruturada lógica seguindo o modelo *MVC*, onde a camada de visualização fica separada da camada das regras de negócio da aplicação. Além disso, os componentes de conteúdo estático, como arquivos *CSS* (*Cascading Style Sheet*), arquivos *JavaScript*, imagens e outros tipos de arquivos ficam armazenados em um diretório separado das regras de negócio e visualização da aplicação. A pasta de *Util* se responsabiliza por armazenar os arquivos que compõem os conteúdos padrões das páginas, como o cabeçalho e rodapé.

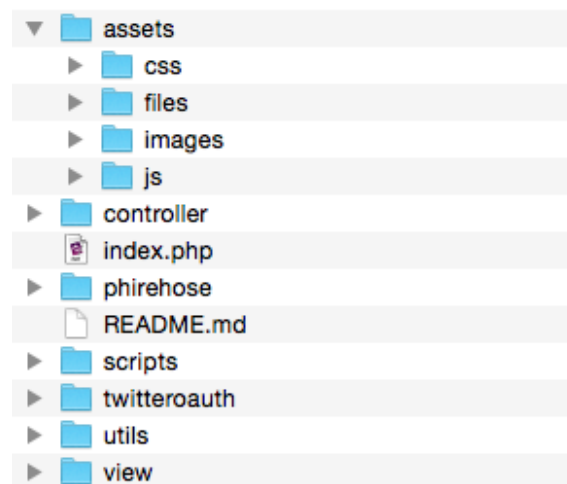


Figura 4.2: Estrutura da Aplicação

A imagem 4.2 ilustra a estrutura utilizada na aplicação onde pode ser observado que dentro da pasta *Assets* ficam armazenados os arquivos estáticos utilizados para complementação e estilização da página. Na pasta *Controller* ficam todos os arquivos relacionados com as regras de negócio e banco de dados, onde são processados os armazenamentos no banco de dados e efetuadas as filtrações da mineração de dados. Na pasta *Utils*, assim como ocorre com a pasta *Assets*, ficam armazenados os conteúdos estáticos das páginas em si, e na pasta *View*, fica a camada de apresentação da aplicação, que possuem as telas da aplicação. Além de toda a estrutura da aplicação, pode ser observada nas pastas *twitteroauth* e *phirehose*, que armazenam todo o código fonte das *API* que efetua a tarefa de se comunicar com a rede social e efetuar os devidos procedimentos.

### 4.3.3 Extrator

O protótipo de extrator de possui a finalidade de coletar, filtrar e armazenar as informações coletadas da rede social e armazenar no banco de dados. A coleta é efetuada utilizando-se de coordenadas e em um segundo momento é efetuada uma busca secundária através de termos de busca. É apenas neste momento em que existe a interação com a rede social, pois nas demais funcionalidades do sistema servem apenas de apoio para que esta etapa ocorra corretamente. Por conta disso, é apenas no controlador do extrator *contExtrator.php* e do script coletor *coletor.php* que estão definidos os parâmetros da aplicação, com as chaves e *tokens* de acesso a rede social que foram criados anteriormente.

Antes de efetuar a coleta das informações do *Twitter*, o sistema carrega todas as informações necessárias para efetuar a busca antes dela acontecer. Nesta etapa de pré-seleção dos dados, o sistema busca os pares de coordenadas cadastrado no sistema, assim como todos os termos de busca cadastrados. Após isto, o sistema efetua a comunicação e coleta dos dados da rede social seguindo os dados pré-cadastrados no sistema.

Com os dados coletados da rede social, é realizada a verificação das mensagens que estão sendo importadas com as mensagens que já haviam sido importadas anteriormente, para que não existam duplicidade de mensagens. No banco de dados são armazenados a mensagem, a data, a hora de quando esta mensagem foi postada na rede e quem foi o usuário que postou aquela mensagem na rede social.

Quando esta coleta está sendo efetuada através da interface Web, é apresentado na tela inicial as mensagens que os dados estão sendo coletados da rede social, ou é apresentada uma tela com a quantidade de mensagens existentes no banco de dados. Na imagem 4.3 pode ser observada esta segunda tela.



Figura 4.3: Estrutura da Aplicação - Aguardando a próxima extração

Durante a coleta dos dados utilizando a interface Web, foi percebido que depois de um tempo de coleta, as requisições com a rede social começavam a falhar. Para resolver este problema foi desenvolvido um script em *PHP* que possui as mesmas

regras que a página principal do extrator, porém com um laço de repetição infinito que, mesmo que alguma requisição falhar, a aplicação não ficará travada. Segue abaixo o trecho de código onde está implementado este laço de repetição.

```
<?php
...
// definição das chaves da Aplicação
echo "Iniciando coleta..." . PHP_EOL;
try {
    $mongo = new MongoClient();
    $db = $mongo->twitterdb;
    $collection = $db->createCollection('coordenada');
    $dados = $collection->find();

    foreach ($dados as $document){
        // Carrega coordenadas do banco
        ...
    }
    $sc = new FilterTrackConsumer(OAUTH_TOKEN, OAUTH_SECRET, Phirehose::
        METHOD_FILTER);
    $sc->setLocations(array(
        array((float)$longitude_inicial,(float)$latitude_inicial,(float)
            $longitude_final,(float)$latitude_final),
        ));
    while(true){
        restApiConsumer();
        $sc->consume();
        $collData = $db->createCollection('mensagem');
    }
}
?>
```

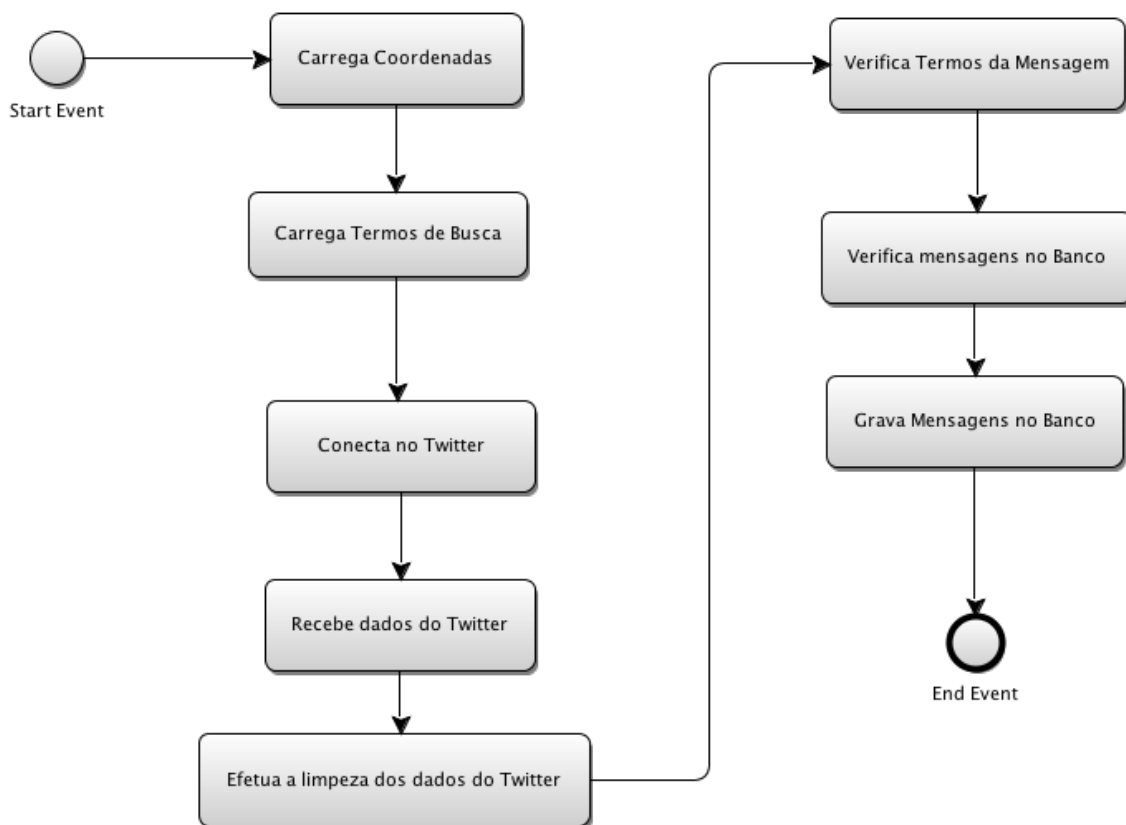


Figura 4.4: Fluxo de coleta de dados do *Twitter*

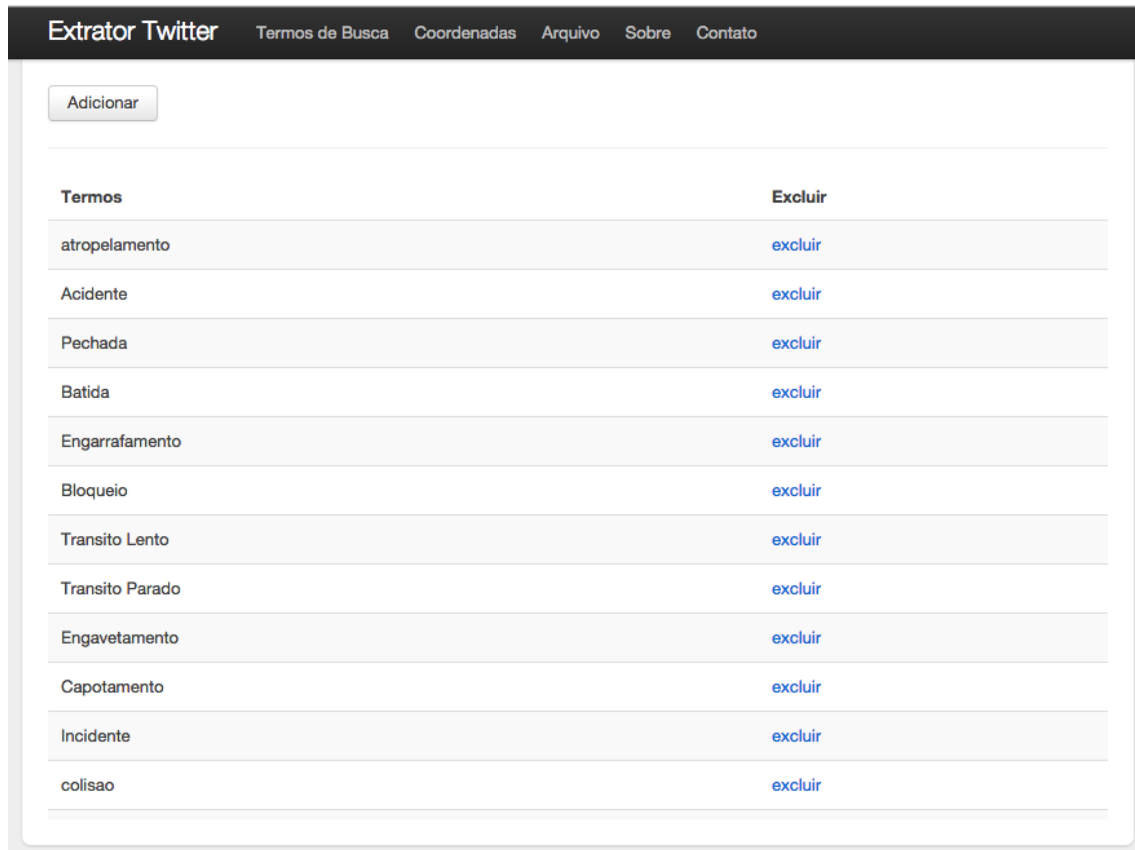
O processo de extração de dados segue o fluxo apresentado na imagem 4.4, onde parte da montagem dos parâmetros para seleção dos dados da rede social que foram cadastrados nas outras etapas. A conexão é iniciada com a rede social e os processos de coleta dos dados, limpeza, classificação e armazenamento dos dados são efetuados dentro de um laço de repetição, possibilitando armazenar as informações do *Twitter* à medida que as mesmas vão sendo carregadas.

#### 4.3.4 Termos de Busca

A funcionalidade de Termos de Busca permite que sejam cadastradas as palavras-chave. Este é um processo de cadastramento básico, mas para a aplicação é essencial para que as buscas sejam direcionadas exatamente para o assunto desejado, no caso da pesquisa deste trabalho o assunto desejado foram buscas relacionadas com as ocorrências do trânsito.

O processo de cadastramento dos termos de busca, o processo foi dividido em duas etapas. O primeiro está relacionado com a listagem do que está cadastrado no sistema. Esta listagem demonstra simplesmente o que foi cadastrado no sistema e permite a exclusão de cada um dos termos relacionados. Um exemplo desta tela pode ser observado na imagem 4.5.





The image shows a web application prototype for 'Extrator Twitter'. It features a dark header with the title and navigation links. Below the header is a button labeled 'Adicionar'. The main content area contains a table with two columns: 'Termos' and 'Excluir'. The table lists various terms related to traffic accidents, each with a corresponding 'Excluir' link.

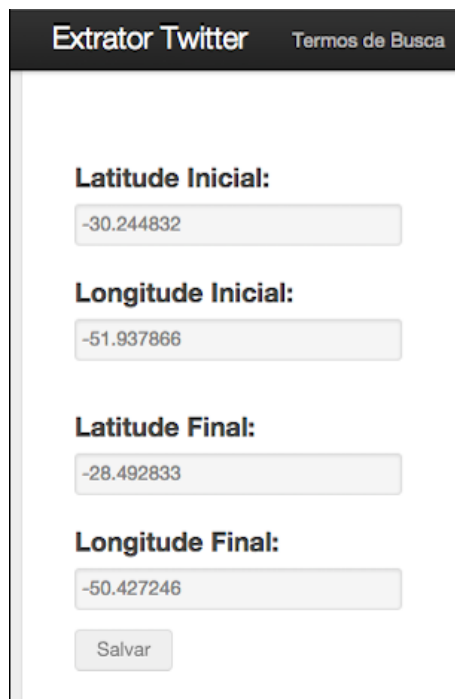
Termos	Excluir
atropelamento	<a href="#">excluir</a>
Acidente	<a href="#">excluir</a>
Pechada	<a href="#">excluir</a>
Batida	<a href="#">excluir</a>
Engarrafamento	<a href="#">excluir</a>
Bloqueio	<a href="#">excluir</a>
Transito Lento	<a href="#">excluir</a>
Transito Parado	<a href="#">excluir</a>
Engavetamento	<a href="#">excluir</a>
Capotamento	<a href="#">excluir</a>
Incidente	<a href="#">excluir</a>
colisao	<a href="#">excluir</a>

Figura 4.5: Protótipo de tela dos termos de Busca

A segunda etapa do processo é responsável pelo cadastramento das palavras-chave no banco de dados. Na tela principal está disponível o botão "Adicionar", que pode ser observado também na imagem 4.5. Através desse botão se tem acesso à tela de cadastramento, que abre sobre principal uma tela com o formulário que possui os campos necessários para cadastrar os termos de busca.

#### 4.3.5 Coordenadas

Para completar e complementar o processo de montagem dos parâmetros para efetuar a busca, foi desenvolvido o cadastro de coordenadas. Este cadastro permite definir os dois pares de coordenadas que servem para delimitar a área de busca das mensagens, porém só se pode cadastrar o par inicial e o par final das coordenadas, de forma que somente exista uma área de coleta, como pode ser observado na imagem 4.6.



O protótipo da interface do Extrator Twitter apresenta uma barra superior com o título 'Extrator Twitter' e um link 'Termos de Busca'. Abaixo, há quatro campos de entrada para coordenadas geográficas, cada um com um rótulo em negrito: 'Latitude Inicial' (valor: -30.244832), 'Longitude Inicial' (valor: -51.937866), 'Latitude Final' (valor: -28.492833) e 'Longitude Final' (valor: -50.427246). Um botão 'Salvar' está posicionado na base dos campos.

Figura 4.6: Protótipo de tela dos dois pares de coordenadas

#### 4.3.6 Exportação de Dados

Para finalizar o ciclo do processo do extrator, é necessário gerar os arquivos para serem processados pelo software *Weka*. Para isto foi necessário desenvolver uma funcionalidade no extrator para que fosse possível exportar os dados do banco, e de alguma forma e este estivesse organizado para que o *software* de mineração conseguisse entender o que está presente neste arquivo gerado. Esta funcionalidade desenvolvida permite que a aplicação exporte os dados do banco em um arquivo do tipo *CSV*, que é aceito pelo software assim como os arquivos *ARFF*, conforme observado na imagem ??.

O procedimento que efetua a exportação dos dados para um arquivo do tipo *CSV*, o sistema utiliza uma estrutura específica para que os dados da análise sejam gerados de forma correta e compreensível. Dentre esta estrutura ficam a data, a hora, o dia da semana e o tipo de ocorrência, que é derivado do tipo de ocorrência. Ao selecionar a opção "Arquivo" no menu da aplicação, a mesma reúne todo o conteúdo necessário e gera um arquivo do tipo *CSV* na pasta *assets/files*, além de apresentar na tela um *link* para efetuar o *download* deste arquivo e poder salvar em qualquer outro diretório e em qualquer outro computador, não ficando restrito somente ao servidor.

Pode ser observado na imagem 4.7 um exemplo do arquivo exportado pelo protótipo de extrator de dados do Twitter. Nele cruzam-se as informações de ocorrências do trânsito, o tipo de veículo que se envolveu em determinada ocorrência,

o dia da semana e o período do horário da ocorrência.

<b>Dia Semana</b>	<b>Periodo</b>	<b>Ocorrencia</b>	<u>automovel</u>	<u>bicicleta</u>	<u>caminhao</u>	<u>caminhonete</u>	<u>moto</u>	<u>onibus</u>	<u>pedestre</u>
qui	21-23	Acidente	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>
qui	21-23	Acidente	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>
qui	21-23	Nenhum	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>
qui	21-23	Batida	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>
qui	21-23	barreira	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>
qui	21-23	Nenhum	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>
qui	21-23	Bloqueio	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>
qui	21-23	Nenhum	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>	<u>nao</u>

Figura 4.7: Exemplo do layout do arquivo gerado pelo Extrator

## 5 ANÁLISE E AVALIAÇÃO DAS INFORMAÇÕES

Com o desenvolvimento do extrator finalizado, o foco se direciona em coletar e analisar as informações do trânsito provenientes do *Twitter*. O processo de extração das informações, classificações e limpeza das informações fica por conta do extrator de dados implementado neste trabalho e descrito no capítulo anterior. A mineração dos dados fica por conta do *Weka*, o qual possui a implementação de diversos algoritmos de Regras de Associação implementados. A mineração é através do carregamento de um arquivo e efetuada a solicitação de execução da análise, entretanto a interpretação dos resultados ainda deve ser efetuado de forma manual, de forma que o volume de dados foi reduzido e tornou esta tarefa humanamente possível.

Com a filtragem por termos de busca e por coordenadas, foi possível determinar qual o assunto que uma mensagem está tratando e qual a origem de sua localização. Se a mesma não suprir estes dois requisitos básicos que uma mensagem deve possuir, a mensagem que está sendo analisada no momento será desconsiderada.

### 5.1 Parâmetros Cadastrados

Como o objetivo principal é tratar as ocorrências de trânsito da Serra Gaúcha, foram cadastrados os pares de coordenadas apresentadas na imagem 4.6. As mensagens postadas no *Twitter* e que possuem estas mediações serão consideradas e levadas para o próximo nível de filtragem. Nas mensagens do *Twitter* também existem localizações por cidades, estados e países, porém como optou-se por utilizar as coordenadas geográficas, as mensagens que não tiverem Latitude e Longitude serão desconsideradas da análise de dados, mesmo que estejam entre as cidades da Serra Gaúcha.

O segundo nível de filtragem está relacionado com os termos de busca e trechos de texto que devem estar contido nas mensagens. Para efetuar a extração das mensagens foram cadastrados os termos relacionados às ocorrências de trânsito. Todas as mensagens que possuírem em qualquer trecho os termos presentes na lista apresentada na imagem 4.5 devem ser considerados para análise, caso contrário a mensagem será descartada.

## 5.2 Coleta dos Dados

A coleta dos dados pode ser efetuada de duas formas distintas: a primeira é através da página principal da aplicação e a outra forma é através de um *script* que é rodado através de linha de comando. Em ambas as opções, a aplicação irá efetuar o mesmo processo de comunicação com a rede social, mesmo tipo de filtragem e armazenamento em banco de dados. Em comparativo das duas formas de coleta dos dados, a opção de rodar através de *scripts* é que não existe a limitação do tempo de requisições e de quantidade de requisições que pode ser efetuada no dia.

A extração dos dados da rede social foi efetuada durante um período de 30 dias, envolvendo os meses de Setembro e Outubro. Durante este período, o sistema extrator coletou aproximadamente 1500 mensagens, valor este que pode ser considerado razoável se for levado em consideração os diversos níveis de filtrações que foram efetuados para coletar os dados do presente trabalho. Dentre eles, o filtro que mais restringe as mensagens é aquele que limita a localização pelos motivos de ser baixo o número de mensagens que estão no *Twitter* que possuem as coordenadas de origem da mensagem e como consequência, menor ainda a quantidade de mensagens que estão com as coordenadas dentro da área da Serra Gaúcha.

As mensagens que foram extraídas e filtradas receberam um tratamento de limpeza, onde qualquer caracter especial foi removido, e após esta etapa foi efetuado o armazenamento destas mensagens no banco de dados *MongoDB*. Ao término deste processo de coleta, os dados foram exportados através de uma opção no ambiente WEB do extrator, que gera um arquivo no formato *CSV*, que possibilita a análise de cada uma das métricas contidas na estruturação do arquivo. Todas estas etapas foram pensadas nas limitações e exigências que o *Weka* impõe em relação ao tipo de arquivo e ao tipo de dados que pode estar contido no arquivo. A estratégia de coleta dos dados em um período de aproximadamente um mês, durante todos os dias foi assim determinada para que fosse possível montar uma base de dados capaz de determinar o perfil de trânsito mensal, além de que pelo volume de dados que o extrator retornou, a análise em um período semanal iria distorcer muito as estatísticas do perfil.

Avaliando os dados que foram coletados no segundo semestre do ano de 2014 com os dados que foram coletados no segundo semestre de 2015, foi percebido que a quantidade de mensagens diretamente relacionadas com o trânsito reduziu consideravelmente, o que retrata uma mudança no perfil do usuário do *Twitter*, que tem migrado para outras redes sociais ou utilizado outras redes sociais em paralelo com o *Twitter*. Esta característica pode ser observada pelos dados extraídos, onde a grande maioria dos dados coletados que se enquadram nos filtros utilizados pelo extrator, acabam não retratando devidamente a situação do trânsito, de mesma forma que as

mensagens que foram coletadas não possuem a estrutura correta que era esperada, o que por consequência acaba comprometendo muito no resultado da mineração dos dados, através da execução dos algoritmos de mineração de dados.

### 5.3 Utilização do Software de Mineração

O software *Weka* está disponível na página <http://www.cs.waikato.ac.nz/ml/weka/>, onde pode baixado e instalado seguindo a padronização de cada sistema operacional. Este software funciona sobre a plataforma *Java*, portanto é necessário que este esteja devidamente instalado no sistema. Como exemplos de forma de instalação:

- no sistema *Mac OS X*, a instalação é efetuada descompactando a pasta baixada do site e copiando-a para a pasta *Applications*
- no sistema *Windows*, a instalação é efetuada através de um instalador com algumas etapas até a finalização da instalação, normalmente instalada em **C: Program Files**

O *Weka* é muito utilizada em ambiente acadêmico, pois permite que o algoritmo seja testado manualmente e que os resultados da análise dados sejam rapidamente apresentados na tela, permitindo entender o comportamento do algoritmo. Como o presente trabalho possui o mesmo perfil, onde o objetivo era o estudo de um algoritmo e analisar o comportamento do mesmo frente à análise dos dados provenientes de uma rede social, foi selecionada esta interface para efetuar as análises dos dados.

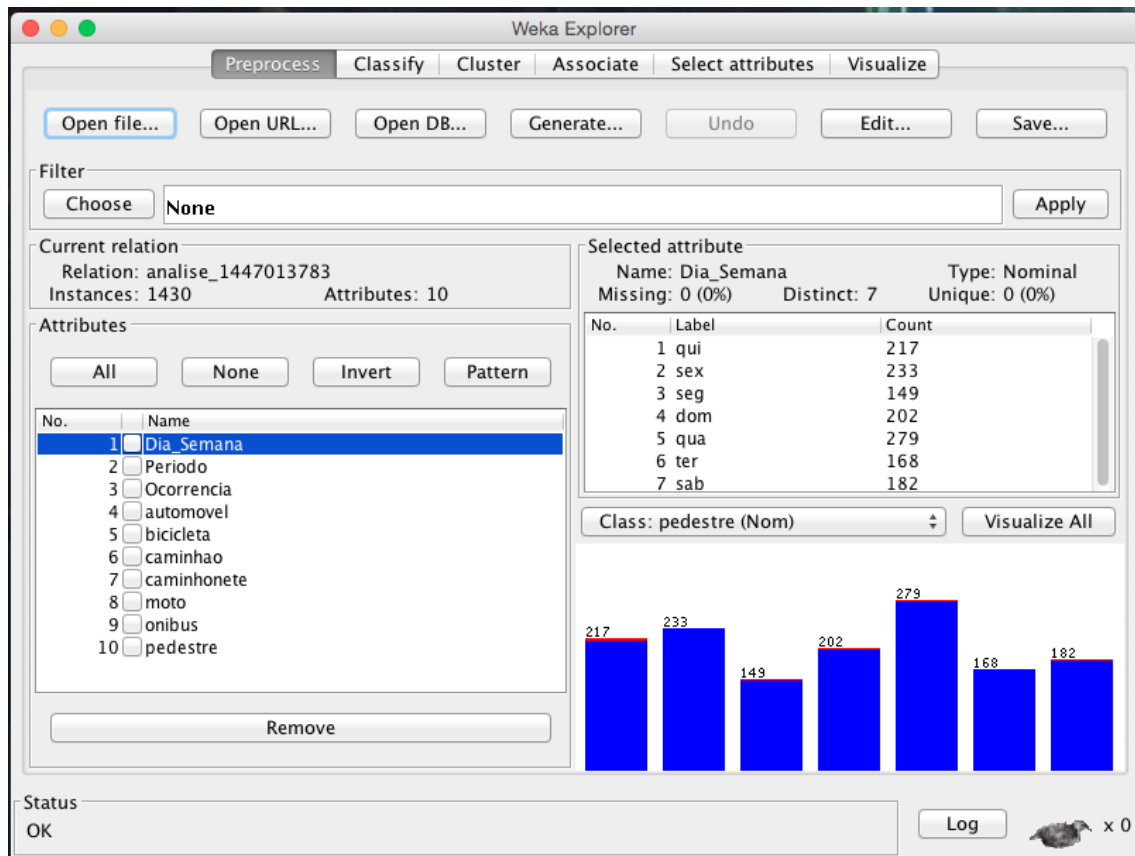


Figura 5.1: Tela do Weka na etapa de pré-processamento

O *Weka Explorer* divide o processo de análise de uma base de dados ou de um arquivo em três etapas, sendo elas treinamento e filtragem, a mineração dos dados em si e a visualização dos resultados. Pode ser visualizado na imagem 5.1 todas as opções que esta funcionalidade do software permite. Esta imagem apresenta o momento em que está sendo realizado o pré-processamento do arquivo com os dados extraídos da rede social.

Na etapa de treinamento, pode ser efetuada a seleção do arquivo *CSV*, um arquivo no formato *ARFF*, ou pode ser uma conexão direta com um banco de dados. Pela facilidade de manipulação dos dados e montagem dos dados, foi utilizado um arquivo *CSV*. Além da seleção da fonte dos dados, é possível também modificar algum parâmetro e/ou métrica através de funções, assim como pode ser efetuada uma filtragem em específico, caso a base de dados contenha um número muito elevado de dados.

Na etapa de mineração dos dados é efetuada a seleção do algoritmo de mineração de dados, de forma que os mesmos estão organizados conforme a sua classificação, sendo Classificação, *Cluster* e Associação.

Ao selecionar o tipo de algoritmo, existe o botão *Choose* para selecionar o algoritmo de mineração e ao seu lado a sequência de parâmetros. Neste campo ao lado,

ao ser clicado, podem ser alterados os valores dos parâmetros conforme a regra de negócio desejada.

Em algumas situações, para identificar a utilidade de cada um dos parâmetros do algoritmo de mineração de dados escolhido, foi necessário efetuar algumas buscas na documentação do *Weka*. Serve como exemplo a utilização do algoritmo *Apriori*, que para conseguir identificar corretamente os parâmetros que deveria ser utilizado foi necessária a consulta na documentação.

## 5.4 Resultados Obtidos

Mesmo sabendo que a base de dados está em de certa forma comprometida, devido à baixa quantidade de dados úteis coletados, os algoritmos foram executados e os resultados analisados. comprometidos. Para executar ambos os algoritmos foi utilizado o mesmo arquivo *CSV*, que possui as seguintes colunas:

- Dia da Semana
- Período
- Ocorrência
- Automóvel
- Bicicleta
- Caminhão
- Caminhonete
- Moto
- Ônibus
- Pedestre

### 5.4.1 Apriori

Com a quantidade de dados úteis para a análise era realmente muito baixa, o algoritmo *Apriori* efetuou algumas análises, entretanto os resultados obtidos são somente de falsos positivos, o que indica que as regras de associação encontradas na base selecionada foram somente com resultados de análises negativas como por exemplo: se um automóvel não se envolveu em um acidente, um caminhão também não se envolveu. Outro fato que indica que o resultado do *Apriori* não foram significativos foi de que o grau de confiança para todas as regras de associação encontradas pelo algoritmo ficaram com 1.

A execução do algoritmo foi efetuada com os parâmetros *car = false*, *classindex = -1*, *delta = 0.01*, *metricType = Confidence*, *minMetric = 0.9*, *numRules=25*, *outputItemSets=False*, *removeAllMissingCols=False*, *significanceLevel=-1.0*, *upperBoundMinSupport = 1.0* e *verbose=false* o algoritmo retornou somente resultados



negativos com grau de confiança 1 em todas elas. Com eles, o objetivo foi de explorar as 25 melhores regras de associação que o algoritmo conseguisse encontrar, o delta foi setado também para efetuar um nível de cruzamento entre os dados mais profundo.

#### 5.4.2 Tertius

Da mesma forma como foi efetuada a mineração com o algoritmo *Apriori*, a análise utilizando o *Tertius* também passou pelo mesmo problema da baixa quantidade e qualidade dos dados, o que gera uma grande probabilidade dos resultados da análise saírem distorcidos e incoerentes com o foco da pesquisa.

Com o intuito de gerar a mesma condição de análise que o algoritmo *Apriori* foi condicionado, foi utilizado o mesmo arquivo e os parâmetros foram configurados para que fosse possível selecionar a mesma quantidade de regras de associação, assim como foi configurado o parâmetro para ir até o mesmo nível de profundidade que o algoritmo *Apriori* foi. Os parâmetros utilizados na análise foram: *classIndex=0*, *classification=false*, *confirmationThreshold=0.0*, *confirmationValues=25*, *hornClauses=false*, *missingValues=MatchAll*, *negation=None*, *noiseThreshold=1.0*, *numberLiterals=4*, *repeatLiterals=false*, *rocAnalysis=False* e *valueOutput=No*.

Durante a execução, foi percebido que este algoritmo levou muito mais tempo para ser executado, e dentre os resultados do relatório do *Weka*, foi identificado que o número de hipóteses exploradas foi de aproximadamente 157826. Esta numeração indica que foram efetuados diversos cruzamentos entre cada um dos registros do arquivo CSV para conseguir efetuar a análise dos dados. Sobre os resultados obtidos, foram possíveis identificar algumas regras que representam determinadas condições do trânsito, como por exemplo o maior número de acidentes ocorreram no período do fim de semana e no horário da madrugada.

Mesmo conseguindo encontrar algumas regras de associação interessantes, estas regras poderiam ser traçadas através de algumas consultas de SQL no banco de dados. Isto por sua vez prova que o processo de mineração de dados, neste determinado momento foi ineficaz devido à pequena quantidade de dados disponível para a análise e pela grande quantidade de ruído nas informações.

#### 5.4.3 Apresentação de Resultados

Avaliando os resultados obtidos pela execução do algoritmo *Tertius*, foi efetuada a análise destas regras de associações geradas por este algoritmo e interpretadas. Nesta análise foram efetuados alguns cruzamentos entre as ocorrências de trânsito e alguns parâmetros pré-definidos

A partir das regras de associações geradas pelo algoritmo *Tertius*, foi possível

efetuar algumas análises simples dos resultados que o mesmo retornou. Foram efetuados alguns cruzamentos entre as ocorrências de trânsito e alguns parâmetros pré-definidos. Como pode ser observado na imagem 5.2, podem ser visualizados os tipos de veículos que mais tiveram destaque no momento de montagem das regras de associação através da execução do algoritmo.



Figura 5.2: Gráfico de relação entre Componentes do Trânsito e as ocorrências das regras de associação levantadas

Na imagem 5.3 pode ser observado o cruzamento das informações geradas pela mineração dos dados e os períodos do dia, e analisando os dados é possível identificar que a grande maioria das ocorrências levantadas ocorreram no período das 3:00 às 6:00 horas da manhã. E no gráfico apresentado na imagem 5.4 é efetuado o cruzamento das ocorrências das regras de associação extraídas com os dias da semana que as mesmas ocorreram, em sua grande maioria está nas quintas e nos sábados, e exatamente no mesmo período que as ocorrências por período, das 3:00 as 6:00.

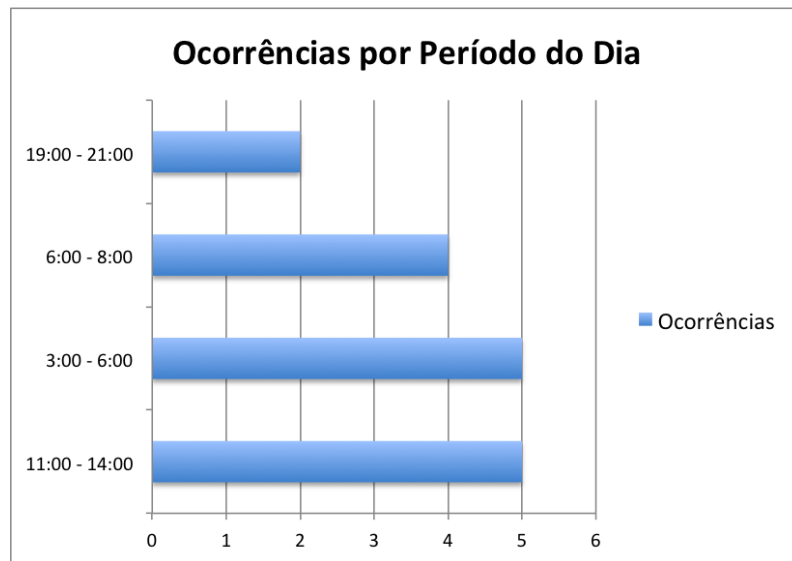


Figura 5.3: Gráfico de relação entre os períodos e as ocorrências das regras de associação levantadas

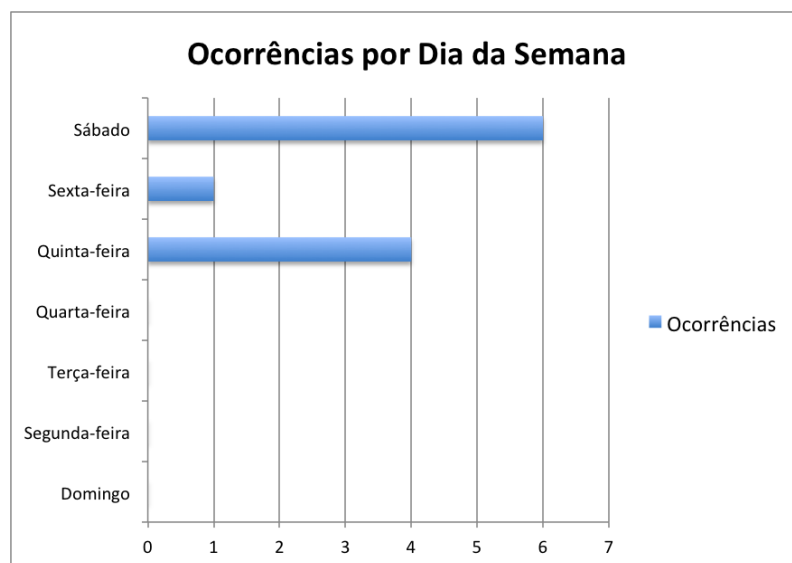


Figura 5.4: Gráfico de relação entre os dias da semana e as ocorrências das regras de associação levantadas

E avaliando de uma forma um pouco mais aprofundada as regras de associações geradas pelo algoritmo, puderam ser traçados alguns perfis do trânsito, pela condição destas regras que foram geradas. Dentre as análises efetuadas, os perfis que acabara se destacando mais foram os seguintes, apresentados na seguinte lista:

- das ocorrências de atropelamentos existentes no arquivo, todas elas foram ocasionadas por automóveis;
- a grande maioria dos acidentes ocorridos, aconteceram no sábado entre às 3 e 8 da manhã.

- uma quantidade mais reduzida de acidentes, envolvendo automóveis e motos ocorreram em quintas-feiras.
- a grande maioria das ocorrências ocorridas entre às 11 e 14 horas foram acidentes e atropelamentos, os quais envolveram automóveis e motocicletas

Os perfis de trânsito que foram montados podem representar em partes a realidade das ocorrências que são levantadas pelos órgãos oficiais e apresentados pelos grandes canais de notícias, sendo eles canais televisão, rádio ou internet. Porém como a quantidade restrita de dados "mineráveis" coletados, o resultado da análise pode estar comprometido ou distorcido, de forma que a pesquisa acaba se tornando pouco confiável, como já foi comentado anteriormente em relação aos resultados de cada um dos algoritmos.

## 6 CONCLUSÃO

O principal tema abordado deste trabalho foi de analisar os dados do trânsito da serra gaúcha e da região metropolitana do Rio Grande do Sul, com o foco de tentar ampliar a quantidade de informações e traçar o perfil do trânsito nesta região. Como um dos objetivos era utilizar de uma rede social para servir como fonte de dados, foi selecionado o *Twitter* e a partir destes dados utilizar de algum algoritmo de mineração de dados e extrair algumas regras de relacionamento das ocorrências com as variáveis do trânsito. Inicialmente são apresentados os conceitos básicos do *KDD*, assim como a execução prática de cada uma das etapas do mesmo.

Durante o desenvolvimento do extrator e a extração dos dados da rede social foi verificado que a popularidade da mesma se manteve, porém o perfil dos usuários acabou se alterando de forma que eles passaram a não postar tanto mais mensagens relacionadas com o trânsito e aumentaram as postagens sobre qualquer assunto. Algo que por consequência, acabou contribuindo para estas mudanças no *Twitter* foram a criação de grupos fechados do *Facebook*, onde são publicadas informações de trânsito entre outras coisas, assim como os grupos do *WhatsApp*. O aumento da popularidade do *Waze* na serra gaúcha também contribuiu para a redução na utilização do *Twitter* para postar informações relacionadas com o trânsito.

Além dos fatores citados acima, o processo de filtragem dos dados do *Twitter* para conseguir determinar os assuntos e a localização da origem da mensagem acaba interferindo na quantidade de mensagens que serão classificadas para passarem pelo processo de mineração de dados. Como forma de contornar esta escassez de informações, o extrator de dados foi desenvolvido utilizando as duas formas disponíveis de coleta de dados: por *Streaming* e por *REST API*. De qualquer forma, os termos utilizados para efetuar a filtragem das mensagens acabaram coincidindo com mensagens do uso cotidiano, que tornou a classificar muitas mensagens que não tratam de qualquer assunto relacionado ao trânsito.

Inicialmente foram definidos dois algoritmos de mineração de dados, para poder analisar as informações do trânsito. Entretanto o algoritmo *Apriori* não conseguiu extrair bons resultados, o que gerou as regras de associação completamente distorcidas. O algoritmo *Tertius*, por sua vez conseguiu extrair algumas regras de associação que foram muito úteis para conseguir traçar alguns perfis do trânsito.

Mesmo sabendo da limitação da quantidade de ocorrências coletadas e analisadas, destes dados coletados foi possível traçar alguns perfis do trânsito, como era

esperado pelas hipóteses levantadas no capítulo 3. A análise das regras de associação demonstrou que é possível enriquecer as informações relacionadas com o trânsito a partir da montagem de perfis por meio de redes sociais, e que o uso de regras de associação foi muito útil para conseguir definir estes perfis.

## 6.1 Trabalhos Futuros

Como próximas etapas necessárias para a aplicação deste trabalho ficam a modificação de alguns filtros de coletas para garantir que as mensagens representem melhor o problema abordado, e como complemento da ampliação dos filtros, adicionar algumas outras hipóteses a serem provadas. Com isto, fica necessário efetuar a mineração de dados, explorando melhor cada um dos parâmetros dos algoritmos de mineração de dados de forma que os resultados apresentem informações que consigam provar as hipóteses definidas.

Sabendo que o *Twitter* não detém mais de tanta informação relacionada ao trânsito, como ocorria no início da concepção deste trabalho, assim é interessante ampliar as fontes de pesquisas para outras redes sociais como o *Facebook*, *Instagram* e *Whatsapp*, explorando os grupos fechados que estas redes sociais possuem. Uma outra forma de ampliação da fonte de dados é utilizar outras regiões, que possuem um maior fluxo de informações, o que contribui na coleta dos dados, para aí sim efetuar as análises dos dados, como são os casos das cidades de São Paulo e Rio de Janeiro. Além de aumentar consideravelmente o nível dos dados úteis para serem analisados, outro passo importante seria efetuar um comparativo entre dados estatísticos de um órgão do trânsito com os dados gerados a partir da análise da mineração de dados.

## REFERÊNCIAS

ALECRIM, E. **O que é Big Data?** [S.l.]: Info Wester, 2015. <Disponível em: <http://www.infowester.com/big-data.php>>. Acesso em: 22 de novembro de 2015.

DEVELOPERS, T. **The Streaming APIs.** [S.l.]: Twitter, 2014. <Disponível em: <https://dev.twitter.com/docs/api/streaming>>. Acesso em: 3 de Julho de 2014.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; PADHRAIC, S. **Knowledge Discovery and Data Mining.** [S.l.]: AAAI, 1996. (Data Mining General Overview).

FLACH, P.; MARALDI, V.; RIGUZZI, F. **Algorithms for Efficiently and Effectively Using Background Knowledge in Tertius.** Departament of computer science, University of Bristol: [s.n.], 2006.

FRANCISCHELLI, R. **Estudo Sobre a Influência da Extração de Dados da WEB na Descoberta de Conhecimento Estratégico Relevante.** Caxias do Sul: [s.n.], 2013.

GALVÃO, N. D.; MARIN, H. d. F. **Técnica de mineração de dados: uma revisão da literatura.** São Paulo: [s.n.], 2009.

INWEB. **Observatório da Web.** [S.l.]: INWeb, 2014. <Disponível em: <http://www.observatorio.inweb.org.br/>>. Acesso em: 7 de Julho de 2014.

NAVEGA, S. **Princípios Essenciais do Data Mining.** São Paulo: Inteliwise Research and Training, 2002.

RIBEIRO JR., S. et al. **Observatório do Trânsito: sistema para detecção e localização de eventos de trânsito no twitter.** [S.l.]: Observatório da Web, 2012.

SANTOS, A. D. P. **Descobrendo eventos locais utilizando análise de séries temporais nos dados do Twitter.** Porto Alegre: [s.n.], 2013.

SILVA, A.; BENEVUTO, F.; ALMEIDA, J. **Coleta e Análise d Grandes Bases de Dados de Redes Sociais Online.** Departamento de Ciência da Computação: [s.n.], 2010.

SILVA, A. et al. **Análise de Padrões de Propagação no Twitter**. Minas Gerais: Departamento de Ciência da Computação - Universidade Federal de Minas Gerais, 2010.

SILVEIRA, C. R. d. R. **Utilização do Twitter em Campanhas de Marketing Digital**. Porto Alegre: [s.n.], 2010.

SOUSA, G. L. S. **Tweetmining**: análise de opinião contida em textos extraídos do twitter. Lavras: [s.n.], 2012.

TOMALÉL, M. I.; ALCARÁ, A. R.; CHIARA, I. G. **Das redes sociais à inovação**. Brasília: [s.n.], 2005.

TWITTER. **Twitter**. [S.l.]: Twitter, 2014. <Disponível em: <http://twitter.com>>. Acesso em: 7 de Julho de 2014.

YABING, J. **Research of an Improved Apriori Algorithm in Data Mining Association Rules**. [S.l.]: International Journal of Computer and Communication Engineering, 2013. (Data Mining General Overview). International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013.