REAI
Revista de Estudos Acadêmicos Interdisciplinar
UNIFEBE

AGRUPAMENTO GEOESPACIAL DE EVENTOS DE SEGURANÇA CLASSIFICADOS COMO FURTO NA REGIÃO DE BRUSQUE

GEOEPACIAL GROUPING OF SAFETY EVENTS CLASSIFIED AS A FURT IN THE BRUSQUE REGION

Gabriel Civinski[1]
Cláudio Ratke[2]

RESUMO: Nesse artigo, utilizou-se os dados provenientes do sistema captura site de notícias relativos a segurança, acidentes, e outras ocorrências de segurança. Esse sistema faz a captura das notícias de segurança das mídias de notícias da região de Brusque. Nesse estudo foram utilizados dados do período de 01/03/2017 a 31/06/2017. Utilizou-se técnicas de mineração de texto para impetrar o endereço do incidente e a API do Google para obter as coordenadas (latitude e longitude) dos incidentes. Nesse trabalho empregou-se apenas os incidentes classificados como: Furto. Nessa base de dados aplicou-se uma técnica de análise de agrupamentos da mineração de dados conhecida como: particionados (*k-means*) nas coordenadas geoespaciais, que tem a finalidade de agrupar dados mais homogêneos entre si e heterogêneos entre os grupos. Com isso, se obteve sete agrupamentos ou regiões que demostram um mapa das áreas de risco com sete grupamentos. Onde pôde-se observar que a região central onde se concentra os Furtos.

Palavras-chave: Informações geoespacial, Mapa de risco, Agrupamentos.

ABSTRACT: *In this article, we used the data from the system capture news site regarding safety, accidents, and other security occurrences. This system captures news security news from the region of Brusque. In this study, data were used from 01/03/2017 to 06/31/2017. Text mining techniques were used to get the incident address and Google API to get the latitude and longitude coordinates of the incidents. In this work, only the incidents classified as: Theft were used. In this database was applied a technique of analysis of data mining groupings known as: k-means in the geospatial coordinates, which has the purpose of grouping data more homogeneous among them and heterogeneous between the groups. With that, seven groupings*

[1] Acadêmico do Centro Universitário de Brusque. E-mail: gabriel.civinski@unifebe.edu.br
[2] Professor do Centro Universitário de Brusque. E-mail: claudioratke@unifebe.edu.br

or regions were obtained that show a map of the risk areas with seven clusters. Where it can be observed that the central region where the thefts are concentrated.

Keywords: *Geospatial information, Risk map, Clustering.*

1 INTRODUÇÃO

A região do Vale do Itajaí assim como toda Santa Catarina e o Brasil sofre com eventos de segurança. A segurança é um dos maiores desafios enfrentados pelos governantes brasileiros. A sensação de insegurança permeia todas as classes sociais e a falta de segurança é considerada, pesquisa após pesquisa, como uma das grandes preocupações dos brasileiros.

Por outro lado, o acesso às notícias de maneira eletrônica, via internet, tais como: páginas de notícias dos veículos de transmissão, blogs, redes sociais, sites oficiais, são fontes de informações sobre os eventos de risco, como por exemplo os assaltos, roubos, acidentes, desmoronamentos, assassinatos, etc. Porém, essas informações são dispersas e não são agrupadas de maneira a produzir uma visão consolidada.

Esse trabalho faz parte de um projeto em desenvolvimento que visa capturar os incidentes de segurança, armazená-los para sua posterior análise. Nesse trabalho utilizou-se as técnicas de agrupamento para agrupar localizações dos incidentes de segurança no caso, eventos de segurança classificados como: Furto.

2 REFERENCIAL TEÓRICO

A Insegurança pode ser confirmada pelos crescentes aumentos de roubos e furtos, o jornal Diário do Vale (2015) já indica um crescimento na ordem de 16,5%.

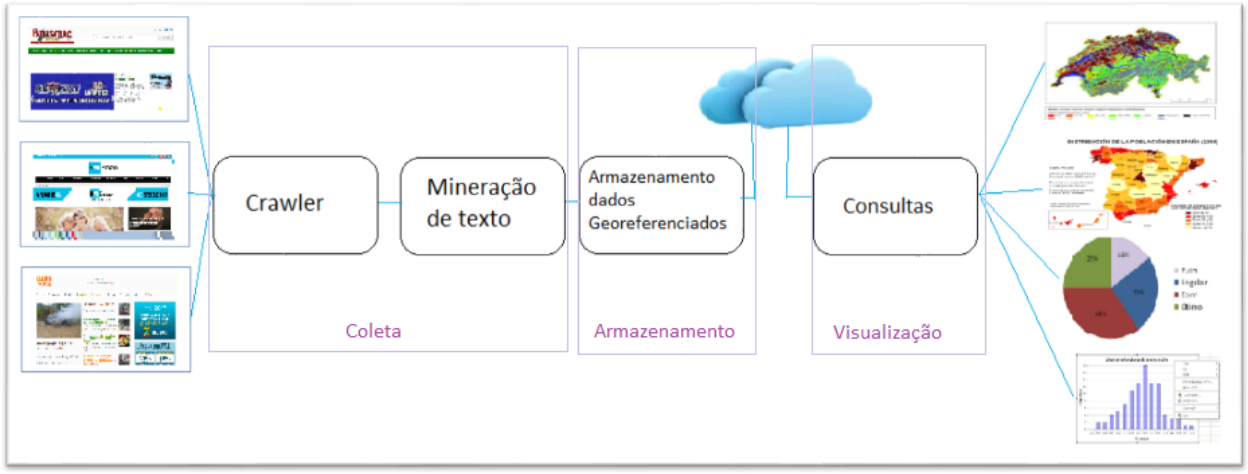
Por outro lado, a Internet está revolucionando as mídias, e é a mídia a mais promissora desde a implantação da televisão. É a mídia mais aberta, descentralizada. Veículos de transmissão de notícias já estão usando essa mídia como principal fonte de informação. Segundo Ferrari (2004, p. 25), o primeiro site jornalístico do Brasil foi o do Jornal do Brasil, criado em maio de 1995, seguido pela versão eletrônica do jornal O Globo. Hoje o Jornal do Brasil circula apenas pela internet

Porém, com a facilidade da geração de conteúdo, tem proporcionado um volume de dados e informações que dificulta a segregação e consolidação da informação. Para ajudar nesse tipo de trabalho estão surgindo tecnologias de coleta e consolidação que facilitem a consolidação da informação.

A solução proposta irá fazer uma coleta dos sites da web, convertendo dados não estruturados em dados estruturados, minerando o seu conteúdo procurando informações

específicas sobre eventos de segurança e sua localização. A macro arquitetura do sistema pode ser vista na Figura 01.

Figura 1-Arquitetura da solução



Segundo Cho e Garcia-Molina (2000) Crawler é um programa que coleta automaticamente páginas da web para criar um índice local e/ou uma coleção local de páginas web e o ponto de partida, porém deve respeitar normas de bom comportamento, tais como (KOSTER,2016):

- Identificar o crawler, usando os campos disponibilizados pelo protocolo HTTP para esse fim;
- Não sobrecarregar servidores Web, evitando pedidos simultâneos ou sequenciais a um mesmo servidor;
- Não visitar servidores ou partes de servidores que não pretendam ser visitados por crawlers, respeitando o protocolo REP (Robot Exclusion Protocol).

A mineração de dados é o processo que consiste na descoberta de informações relevantes em grandes bases de dados. As técnicas de mineração de dados são realizadas sobre depósitos de dados de modo a encontrar padrões úteis e recentes, que poderiam passar despercebidos. Além disto, fornecem a capacidade de se prever resultados de uma observação futura, como, por exemplo, a previsão de quantos desligamentos voluntários futuros irão ocorrer (TAN; STEINBACH; KUMAR, 2009).

Mineração de Texto faz parte da mineração de dados, porém, conforme Tufféry (2011, p. 627), é o conjunto de métodos e técnicas utilizadas para o processamento eletrônico de

grandes volumes de dados em linguagem natural de texto, os quais estão armazenados em sistemas informatizados, para fins de extração e estruturação de conteúdos e temas. Esse processo visa auxiliar a descoberta de dados escondidos nos textos. Pode-se dizer esquematicamente que Mineração de Texto é a junção de Lexicometria com Mineração de dados.

A mineração de texto possui uma série de algoritmos para extração de conhecimento dentro de textos para os mais distintos objetivos. O algoritmo de mineração de opinião em um texto trabalha sobre o contexto de opiniões e entrevistas de usuários, no qual o mesmo minera sobre as opiniões para revelar e sumarizar opiniões sobre o tópico mais discutido, desta forma otimizando decisões e *business inteligence* (AGGARWAL; ZHAI, 2012).

Para que haja uma melhor classificação e análise sobre os textos, a realização de estruturas preparatórias se torna importante, como a retirada de preposições, singularização das palavras, criação de uma tabela de sinônimos que visa agrupar uma série de palavras em um mesmo grupo. Conforme Marcacini, Moura e Rezende (2011):

Para a extração e organização não supervisionada de conhecimento a partir de dados textuais, o diferencial está na etapa de extração de padrões, na qual são utilizados métodos de agrupamento de textos para organizar coleções de documentos em grupos. Em seguida, são aplicadas algumas técnicas de seleção de descritores para os agrupamentos formados, ou seja, palavras e expressões que auxiliam a interpretação dos grupos. (MARCACINI; MOURA; REZENDE, 2011, p.8).

Segundo Wives (2002), as fases para realizar a busca automática de palavras relevantes e similaridades, são a identificação de termos, a remoção de stop words, a normalização morfológica e seleção de termos. As características de cada etapa propostas por Wives (2002) são:

a) identificação de termos: nesta fase é aplicado um analisador léxico que identifica as palavras e ignorados símbolos, caracteres de controle ou de formatação;

b) remoção de palavras irrelevantes (*stop words*): consiste no processo de eliminar palavras que funcionam apenas para realizar a ligação entre frases, sendo que estas não necessitam ser incluídas. Por exemplo: retirada de palavras como “nas”, “das”, “ou”, “seja”, entre outras;

c) normalização e padronização de vocabulário: este processo visa eliminar as variações morfológicas de uma palavra, através da identificação do radical livre desta palavra, onde os prefixos e sufixos são eliminados e os radicais resultantes são utilizados. “Assim, uma ideia, independentemente de ter sido escrita através de seu

substantivo, adjetivo ou verbo, é identificada por um mesmo (e único) radical." (WIVES, 2002, p.53);

d) seleção de termos relevantes: essa etapa consiste na exclusão dos termos com menor importância, existe uma série de técnicas para a seleção de termos que podem se basear na posição dos termos ou na sua posição quanto a sintaxe.

Após a aplicação das etapas o resultado será o conjunto de palavras que possuíram a maior importância dentro do contexto analisado. Sendo que, através destas palavras pode-se detectar pontos negativos e positivos, permitindo assim, que sejam tomadas decisões a partir do resultado gerado (WIVES, 2002).

É a parte do trabalho em que se dá um referencial teórico para situar o assunto. Tratamos de expor, de modo sintético, o que já se escreveu sobre o assunto, por meio de um resumo fiel da ideia central dos materiais lidos (livros, artigos de periódicos, dissertações, entre outros).

As informações obtidas após a mineração textual serão armazenadas neste banco com a adição das informações de latitude e longitude, que serão utilizadas para a determinação da geolocalização. Tem-se a necessidade de optar em um banco de dados que permita às realizações de operações geoespacais, como por exemplo o MongoDB (MONGODB, 2017).

Por fim, na visualização utilizaremos um SIG (Sistema de informações Geográficas). Rosa (2004), um SIG pode ser definido como um sistema destinado à aquisição, armazenamento, manipulação, análise e apresentação de dados referidos espacialmente na superfície terrestre, integrando diversas tecnologias. Essa tecnologia automatiza tarefa até então realizadas manualmente e facilita a realização de análises complexas, por meio da integração de dados de diversas fontes.

Agrupamentos (*clustering*) como uma relação objetos de dados no mesmo agrupamento devem ser semelhantes entre si, enquanto objetos de dados em diferentes clusters devem ser diferentes. (Maravalle et al., 1997).

Dentre os diversos métodos de agrupamento tem-se, os métodos de particionamento buscam encontrar, iterativamente, a melhor partição dos n objetos em k grupos. Frequentemente os k clusters encontrados pelos métodos de particionamento são de melhor qualidade (grupos internamente mais homogêneos) do que os k clusters produzidos pelos métodos hierárquicos.

Os métodos de particionamento mais utilizados são baseados em um ponto central (média dos atributos dos objetos - k-médias) ou em um objeto representativo para o cluster (k-medoids) (Kaufman e Rousseeuw, 1990)

Algoritmos da família k-means compartilham o mesmo princípio de operação básica que pode ser Delineou o seguinte:

1. O número de clusters é predeterminado e referido como k (daí o algoritmo "k-" Nomes)

2. Os clusters são representados por vetores de valor de atributo único, genericamente chamados de (Centroides) centros de cluster.

3. O processo é realizado através da atribuição interativa Instâncias de treinamento para clusters com os centros mais próximos (ou seja, menos dissimilares) e, em seguida, deslocando os centros para refletir o conteúdo real de clusters particulares.

3 PROCEDIMENTOS METODOLÓGICOS

Utilizou-se o método indutivo para o desenvolvimento do protótipo através de pesquisas bibliográficas. Uma pesquisa exploratória baseada em estudos de caso e exemplos foi usada para nortear esse trabalho. Utilizou-se técnicas de análise de agrupamento, técnicas de mineração de dados (*Data mining*) para validar os dados e protótipo.

4 ANÁLISE DOS RESULTADOS

Nesse trabalho foram analisados 87 incidentes de segurança do tipo Furto, ocorridos durante o período de 23/03/2017 a 16/06/2017 de 2017. Todo obtidos nas mídias de notícias da região. Após sua obtenção o dado a mineração de dados segrega o endereço e obtém a coordenadas (latitude e longitude) junto a API do Google a latitude e longitude. Um Fragmento desses dados pode ser visto no Quadro 01.

Quadro 1- Fragmento dos incidentes de Furto

Número	Data	Classe	Latitude	Longitude	SuperClasse
3	17/06/17 09:28	Furto de Loja	-27,10288000	-48,91773000	Furto
12	13/06/17 18:06	Furto	-27,08704800	-48,90499900	Furto
13	13/06/17 10:14	Furto de carro	-27,09531000	-48,91875100	Furto
14	13/06/17 10:04	Furto de telefones	-27,09755300	-48,91337600	Furto
16	12/06/17 18:51	Furto televisão	-27,04918000	-48,87824000	Furto
17	12/06/17 18:42	Furto de mochila	-27,14011000	-48,96060200	Furto
19	12/06/17 09:47	Furto residência	-27,10414000	-48,92693000	Furto
20	12/06/17 09:46	Furto residência	-27,10570100	-48,92731100	Furto
25	11/06/17 16:54	Furto carteira no carro	-27,17183000	-48,90278000	Furto
31	08/06/17 19:02	Furto	-27,10676500	-48,94372700	Furto
33	08/06/17 09:57	Furto	-27,12747000	-48,96088200	Furto
36	07/06/17 10:03	Furto	-27,09301000	-48,90097100	Furto

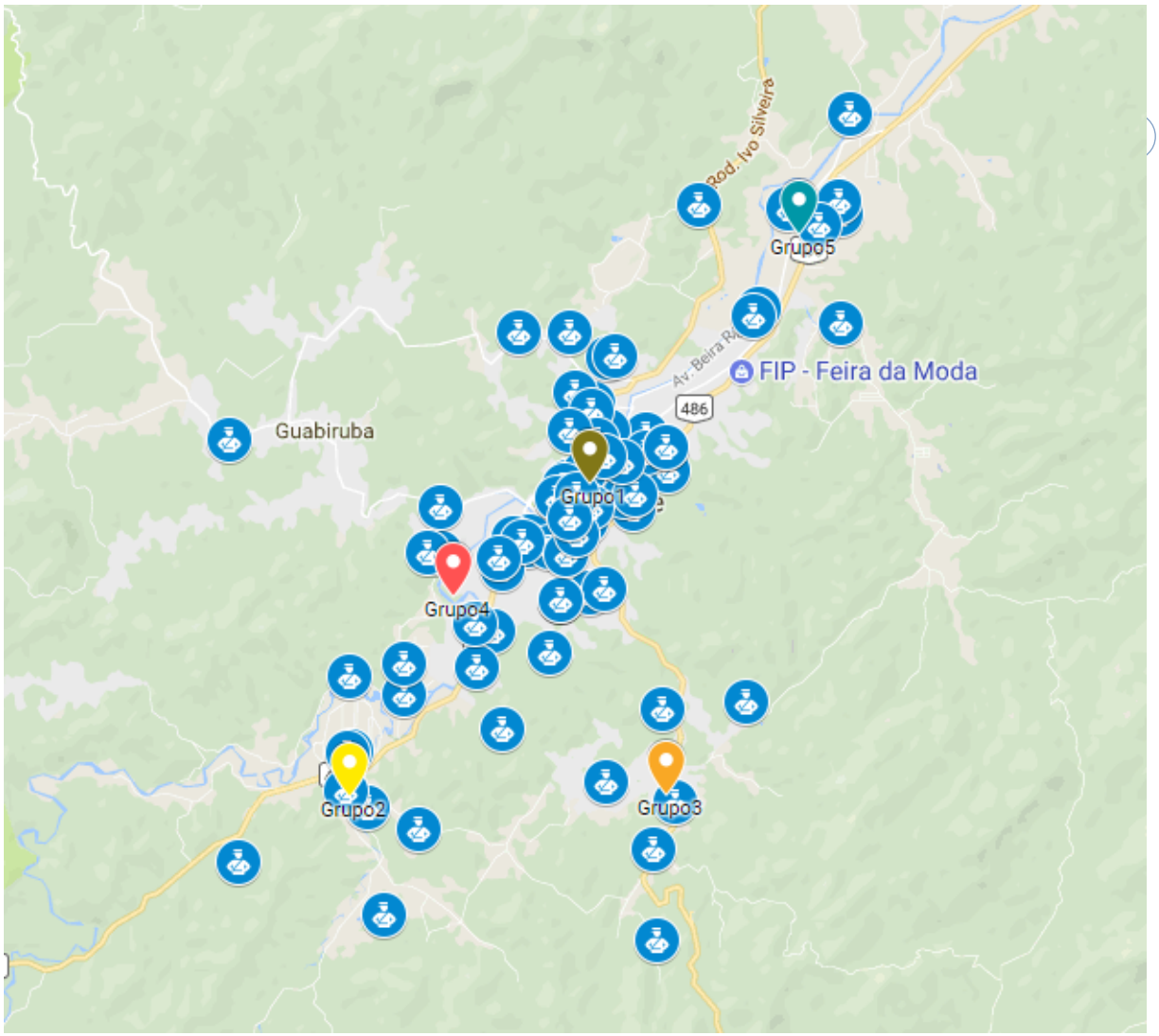
Aplicando a análise de agrupamentos, k-médias, indicando 5 grupos (k=5) obtém-se os seguintes grupos indicados pelos centroides.

Quadro 2 - Grupos encontrados utilizandos k-means

Nome	Latitude	Longitude	Qtde
Grupo1	-27.095453516326522	-48.915765395918385	49
Grupo2	-27.148000024999998	-48.960766512499994	8
Grupo3	-27.147892366666667	-48.901121616666664	6
Grupo4	-27.114531785714288	-48.94117579285713	14
Grupo5	-27.05352027777778	-48.87616241111111	9

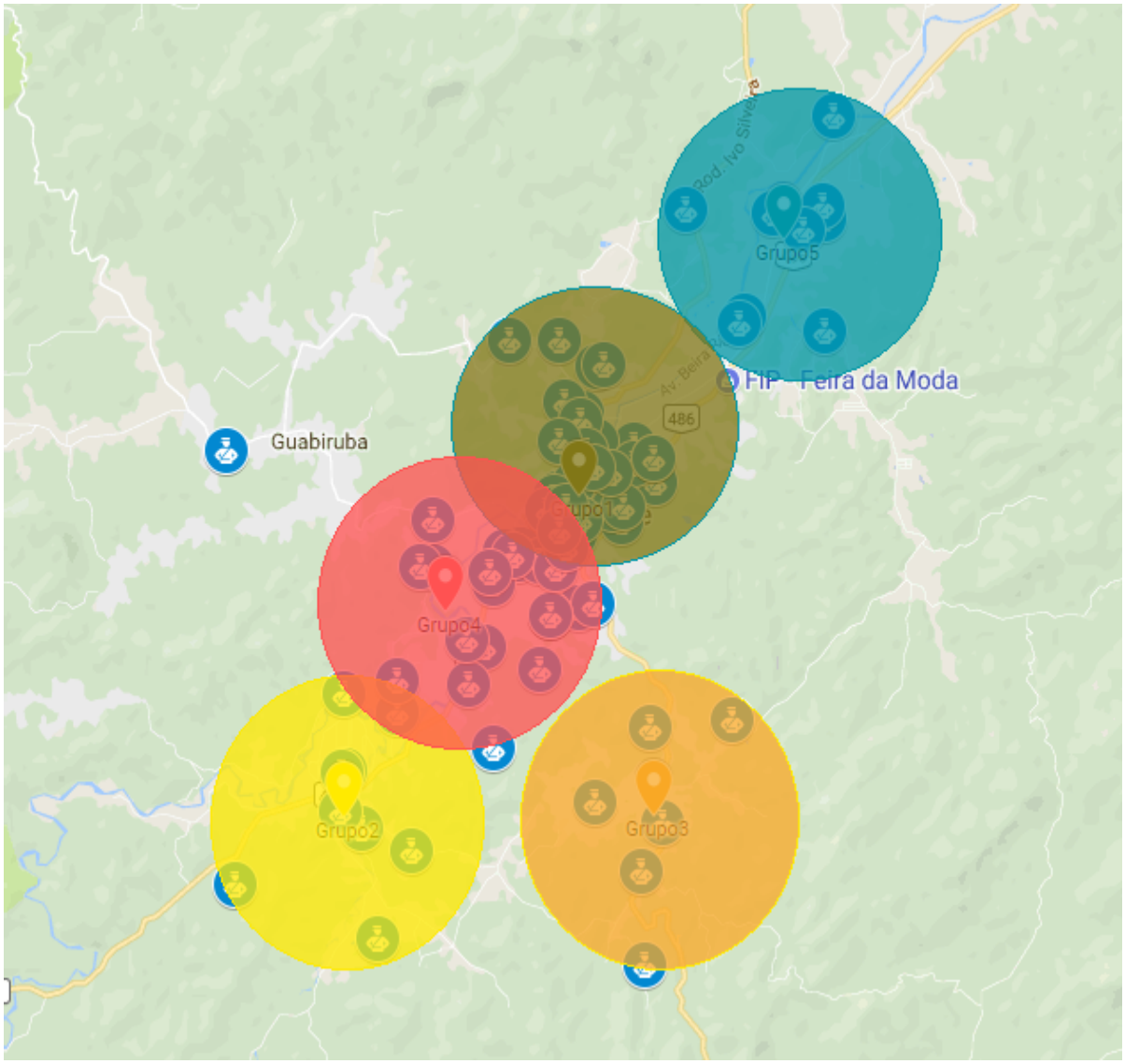
Plotando-se os incidentes e os grupos pode ser visto na Figura 02.

Figura 2 - Indicação dos agrupamentos



Plotando-se os incidentes e os grupos e destacando-se a área dos grupos pode ser visto na Figura 03.

Figura 3 - Região dos agrupamentos indicados.



5 CONSIDERAÇÕES FINAIS

Com os 86 eventos de segurança, da amostra, já se pode observar uma concentração furtos na região central da cidade de Brusque, grupos 1 e 4 com 56% e 16% respectivamente (quadro 2), de todos eventos de furtos relatados.

Como o trabalho ainda está incipiente, e ainda não ter um volume de dados desejados, outras análises podem ser possíveis, tais como:

- Evolução dos agrupamentos em relação ao tempo.
- Principais itens furtados.
- Agrupamentos para outros eventos de segurança (roubo, acidentes, etc).
- Correlação de eventos de segurança (furto, roubo, etc).

Informações contidas nos textos dos incidentes (notícias de segurança) não foram analisadas, aplicação da mineração de texto deve proporcionar uma nova perspectiva para análise de dados.

REFERÊNCIAS

AGGARWAL, Charu C.; ZHAI, ChengXiang. **Mining Text Data.** Nova York: Springer Science & Business Media, 2012. 524 p, il.

AGGARWAL, Charu C.; ZHAI, Chengxiang. **Mining Text Data. Nova York:** Springer Us, 2012. p. 415-463. Disponível em: <http://www.cs.unibo.it/~montesi/CBD/Articoli/SurveyOpinionMining.pdf>. Acesso em: 5 set. 2015.

CHO, Junghoo; GARCIA-MOLINA, Hector. The evolution of the web and implications for an incremental crawler. In: INTERNATIONAL CONFERENCE ON VERY LARGE

DATABASES, 21., 2000, Cairo, Egypt. Proceedings... San Francisco: Morgan Kaufmann, 2000. p. 15-19. Disponível em: <http://oak.cs.ucla.edu/~cho/papers/cho-evol.pdf>. Acesso em: 04 de junho de 2017.

FERRARI, Pollyana. **Jornalismo Digital**. 2.ed. São Paulo: Contexto, 2004.

Jornal Diário do Vale. http://www.diariodovalesc.com.br/noticias.php?id=5862 (2015). Acesso em: 04 de Junho de 2017.

KAUFMAN, L.; ROUSSEEUW, P.J. **Finding groups in data: an introduction to cluster analysis**: Jonh Wiley & Sons, 1990.

KOSTER, M. A Standard for Robot Exclusion. http://info.webcrawler.com/mak/projects/robots/norobots.html, Acesso em: 14 fev. 2016.

LIU, Bing; ZHANG, Lei. A Survey of opinion mining and sentiment analysis. In:

MARAVALLE, M.; SIMEONE, B.; NALDINI, R. **Clustering on trees. Computational Statistics & Data Analysis**, v. 24, n., p. 217-234, 1997.

MARCACINI, Ricardo M.; MOURA, Maria F.; REZENDE, Solange O. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. **Revista de Sistemas de Informação da FSMA,** Macaé, v. 1, n. 7, 2011. Disponível em: <http://www.fsma.edu.br/si/edicao7/FSMA_SI_2011_1_Principal_3.pdf>. Acesso em: 14 março. 2017.

MongoDB. Inc, $near Definition. Disponível em <http://docs.mongodb.org/manual/reference/operator/query/near/>. Acesso em: 14 de Março de 2017.

ROSA, R. **Sistema de Informações Geográficas.** Instituto de Geografia, Universidade Federal de Uberlândia, 2004.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Datamining: mineração de dados.** Rio de Janeiro: Ciência Moderna, 2009. xxi, 900 p, il.

TUFFÉRY, Stéphane. **Data mining and statistics for decision making.** The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom: John Wiley & Sons Ltd., 2011.

WIVES, Leandro K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos.** 2004. 136 f. Tese (Curso de Pós-Graduação em Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.