Um Método de Agrupamento Incremental para a Detecção de Eventos em Redes Sociais

Alice A. F. Menezes[1], **Carlos M. S. Figueiredo**[2,3]

[1]Universidade Federal do Amazonas (UFAM) – Manaus, AM – Brasil

[2]Universidade do Estado do Amazonas (UEA) – Manaus, AM – Brasil

[3]Samsung Ocean – Manaus, AM – Brasil

alice.menezes@icomp.ufam.edu.br, cfigueiredo@uea.edu.br

Abstract. *Studying useful information that are provided by users of social networks is the aim of Social Sensing. Several works in literature present studies with applications in natural disaster detection, traffic monitoring or analyzing dynamic of cities. Such studies normally apply mining and machine learning techniques in previously collected databases for analysis, which are not proper due to real-time nature of posts in social networks. In this paper, we present a method to detect events based on incremental clustering, which proves to be more efficient in terms of processing time and delay over the traditional one. As case study, we detected and analyzed traffic accidents data of New York city, which are events that occurs daily.*

Resumo. *Estudar as informações úteis que são fornecidas pelos usuários de redes sociais é o objetivo do Sensoriamento Social. Diversos trabalhos na literatura apresentam estudos com aplicações de detecção de desastres naturais, monitoramento do trânsito ou análise da dinâmica das cidades. Estes estudos normalmente aplicam técnicas de mineração de dados e aprendizagem de máquina para a análise de dados previamente coletados, o que não é adequado devido à natureza em tempo real das postagens em redes sociais. Neste artigo, apresentamos um método para a detecção de eventos que utiliza agrupamento incremental, o que se revela mais eficiente em termos de tempo de processamento do que a abordagem tradicional. Como estudo de caso, detectamos e analisamos dados de acidentes de trânsito da cidade de Nova Iorque, que são eventos que ocorrem diariamente.*

1. Introdução

Nos dias atuais, redes sociais como o Twitter[1], Facebook[2], Foursquare[3] e Instagram[4], tornaram-se ferramentas importantes para que pessoas possam compartilhar informações relacionadas ao contexto no qual estão inseridas (negócios, ocorrências diárias, vida pessoal e notícias) [Becker et al. 2011]. Devido a isto, o Sensoriamento Social, uma nova área de pesquisa, emergiu.

[1]https://twitter.com
[2]https://facebook.com
[3]https://foursquare.com
[4]https://instagram.com

O Sensoriamento Social utiliza informações disponíveis em redes sociais, que podem ser coletadas e analisadas para a detecção de eventos relevantes em uma área habitacional [Silva et al. 2014]. Neste sentido, trabalhos recentes buscam processar as informações de *stream* do Twitter para a detecção de notícias [Sankaranarayanan et al. 2009, Phuvipadawat and Murata 2010], descoberta de eventos desconhecidos [Becker et al. 2011, Li et al. 2012] e detecção de eventos específicos, como terremotos e trânsito [Sakaki et al. 2010, Nguyen et al. 2016].

As soluções nesta área devem considerar limitações como confiabilidade das informações e dados não estruturados com ruídos ou redundantes [Sakaki et al. 2010]. Desta forma, é comum utilizar técnicas de aprendizagem de máquina tanto no segmento supervisionado quanto não-supervisionado para a análise das informações. Particularmente, técnicas de agrupamento são comumente aplicadas para a detecção de eventos, pois resolvem o problema de redundância dos dados através do agrupamento de informações semelhantes relacionadas ao mesmo evento.

Trabalhos na literatura normalmente utilizam métodos para o processamento e análise de uma grande massa de dados de redes sociais. O problema é que os dados destas redes são dinâmicos, de forma que as informações podem ser publicadas por diversos usuários a qualquer momento e em diferentes formatos. Por exemplo, na detecção de acidentes de trânsito, os usuários podem reportar os eventos conforme passam na mesma área da ocorrência e de acordo com o impacto do mesmo. Assim, com o objetivo de capturar eventos, estes trabalhos vão sacrificar o tempo de detecção do evento para esperar até que uma quantidade significativa de dados possa ser processada. Outra possibilidade é que ocorra alta demanda de processamento devido à utilização sucessiva dos algoritmos, conforme a atualização da base de dados.

Neste contexto, a principal contribuição deste trabalho é a utilização de um algoritmo de agrupamento incremental, que é viável de ser utilizado no cenário de tempo real das redes sociais, em contraposição aos métodos estáticos tradicionais. Além disso, o método proposto diferencia-se de trabalhos anteriores por permitir a detecção de diferentes ocorrências de um evento relacionado ao mesmo assunto, em vez de agrupar os dados sociais em diferentes categorias (exemplo: notícias e tendências de tópicos). Também realizamos um estudo de caso relacionado a detecção de ocorrências de acidentes de trânsito e apresentamos uma similaridade de 90%, enquanto reduzimos o tempo de processamento dos dados.

2. Trabalhos Relacionados

Uma das áreas relacionadas ao Sensoriamento Social é a detecção de eventos, que estuda os fenômenos do mundo real reportados por meio de usuários de redes sociais e que contém informação espaço-temporal [Valkanas and Gunopulos 2013].

As redes sociais tornaram-se uma importante fonte de informação para estudar os aspectos do contexto dos usuários. O Twitter é a rede social utilizada nos trabalhos apresentados nesta seção. Os trabalhos que utilizam esta rede social para obter informações estão relacionados com processamento de dados massivos, providos por *tweets* e *retweets* de usuários em tempo real [Atefeh and Khreich 2015, Li et al. 2012].

Neste sentido, alguns estudos realizam agrupamento de dados para detectar eventos não especificados, como notícias recentes [Sankaranarayanan et al. 2009] ou situações

de desastre natural [Toriumi and Baba 2016]. Outros estudos utilizam a abordagem de agrupamento para diferenciar eventos do mundo real de mensagens não relacionadas a eventos [Becker et al. 2011].

[Sankaranarayanan et al. 2009] propôs um sistema chamado *TwitterStand*, que captura *tweets* relacionados a notícias recentes. Na solução, um classificador baseado em Naive Bayes é responsável por separar notícias de outras informações. Em seguida, um algoritmo de agrupamento forma *clusters* de notícias utilizando *hashtags* para reduzir erros de agrupamento. [Phuvipadawat and Murata 2010] apresentam um método para o rastreamento de notícias recentes do Twitter utilizando palavras-chave em na coleta dos dados. Em seguida, mensagens com termos similares, *hashtags* e nomes de usuários são agrupados. A solução considera o número de seguidores e o número de *retweetes* para ranqueamento dos *clusters*. [Becker et al. 2011] propuseram uma técnica de agrupamento em redes sociais que assimila *tweets* através do TF-IDF. Posteriormente, os *clusters* são classificados em eventos do mundo real e não-eventos através de um algoritmo de *Support Vector Machine* (SVM).

Existem também estudos que analisam apenas um tipo de evento utilizando aprendizagem supervisionada. Nestes estudos, é necessário que informações preliminares sobre os eventos sejam disponibilizadas. Além disso, se o tipo de evento mudar, novas informações devem ser adquiridas para a solução.

[Sakaki et al. 2010] propuseram um modelo para detectar um tipo específico de evento. Os autores treinaram um algoritmo de SVM através de dados do Twitter rotulados manualmente com informações de terremotos e tufões. Em seguida, estimaram a trajetória dos tufões aplicando Filtros de Kalman e Filtros de Partícula. [Nguyen et al. 2016] apresentaram um sistema chamado *TrafficWatch*, que coleta, filtra e analisa *tweets* relacionados a incidentes na Austrália. O objetivo era utilizar as redes sociais como um canal adicional de monitoramento do trânsito e gerenciador de incidentes. Para isto, os autores aplicaram processamento de linguagem natural para extrair informações dos *tweets*. Em seguida, foram utilizados os algoritmos de SVM e Árvores de Decisão para definir eventos relevantes e não relevantes.

Neste trabalho, propomos um método genérico para a detecção de eventos específicos em redes sociais. Diferente dos trabalhos citados nesta seção, utilizamos um algoritmo incremental de aprendizagem não-supervisionada para processar os dados de acordo com a natureza de tempo real da rede social Twitter. O objetivo é melhorar a eficiência no procesamento de informações, atualizando apenas os dados pertinentes a cada nova informação recebida da rede social.

3. Método proposto

Para a detecção de eventos, propomos um método que utiliza um algoritmo de agrupamento incremental para processamento de *streams* de dados provenientes de redes sociais (Figura 1). Esta abordagem é viável em relação a tradicional, pois apenas uma parte específica da base de dados é processada de acordo com a publicação de novas informações nas redes sociais e com o funcionamento do modelo *Sliding Window*. Além disso, esta abordagem permite a detecção de diferentes ocorrências relacionadas ao mesmo evento em estudo, mesmo que este aconteça em locais e horários aproximados.

Primeiramente, é realizada a coleta e filtragem dos dados por meio de palavras-

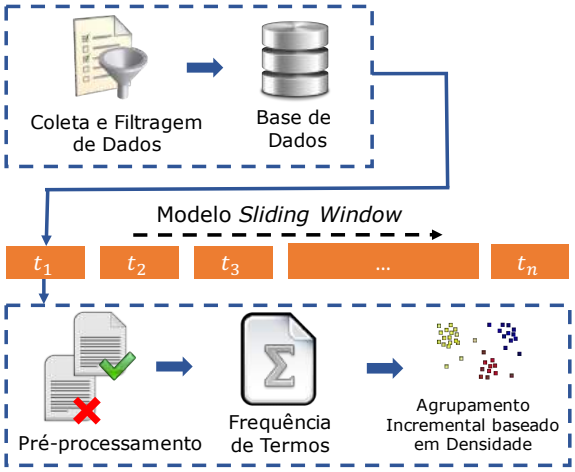


Figura 1. Método proposto para agrupar dados similares em *streams* de redes sociais.

chave relacionados ao tópico em análise. No caso deste trabalho, as palavras-chave são referentes ao estudo de caso de acidentes de trânsito na cidade de Nova Iorque, como por exemplo "*accident*", "*car*", "*injury*" e assim por diante. A coleta de dados é realizada através do mecanismo de busca da API da rede social Twitter ou *Web Crawlers*, resultando em dados massivos de texto que aumentam constantemente. A seguir, a filtragem dos dados é realizada, o que é importante devido os dados de redes sociais possuírem ruídos como gírias e aglutinações [Sankaranarayanan et al. 2009, Sakaki et al. 2010].

Após a remoção dos ruídos, é realizada a extração de catacterísticas dos *tweets*. Apesar de ser possível coletar dados como imagens, fotos de usuários e geolocalização de locais, a maioria das informações provenientes de redes sociais são em forma de texto. Assim, nós extraímos apenas os dados de texto das ocorrências coletadas e aplicamos um algoritmo para determinar a frequência dos termos na extração de características. Assim, os termos mais relevantes para o algoritmo não-supervisionado recebem um peso maior do que os não relevantes.

Por fim, aplicamos o algoritmo de agrupamento incremental para processar apenas novos dados que são adicionados à base, além dos dados que são afetados por esta atualização. Desta forma, o tempo de agrupamento de novas informações reduz, tornando possível a utilização do método para *streams* de dados em constante mudança.

4. Agrupamento Incremental

O processo do algoritmo de agrupamento incremental (Algoritmo 1) é baseado em densidade e consiste em verificar a distância entre os dados e unificar dados próximos. A distância máxima entre dois dados está presente na variável *eps-neighborhood*.

Quando os *clusters* são formados, os dados que não pertencem a nenhum *cluster* são considerados ruídos pelo algoritmo. Em outras palavras, o algoritmo verifica a densidade dos *clusters*. Os de alta densidade são considerados informações de um mesmo evento, enquanto os de baixa densidade são considerados ruídos. Esta característica do algoritmo faz com que o mesmo torne-se viável para ser utilizado na abordagem de Sensoriamento Social, dado que as informações podem conter palavras-chave relacionadas ao estudo de caso, mas podem não pertencer a um evento.

No algoritmo, o *eps-neighborhood* de um dado p, denotado por $N_{Eps}(p)$ é definido $N_{Eps}(p) = q \in D|dist(p,q) \leq Eps$, onde D é a base de dados. Além disso, o algoritmo precisa que para cada dado p em um *cluster* C exista um dado q em C, de forma que a distância entre p e q seja menor que o *eps-neighborhood* e que $N_{Eps}(q)$ contenha no mínimo $MinPts$ dados.

Basicamente, nós inserimos um conjunto de novos dados em uma base de dados previamente agrupada. Devido à natureza de densidade do algoritmo, a atualização influencia apenas dados próximos. Neste processo, dados classificados como ruídos em um agrupamento prévio, podem passar a ser considerados informações, devido às mudanças nos centróides dos *clusters*.

Em nossa solução, aplicamos os conceitos de algoritmo incremental baseado em densidade na biblioteca *Scikit-Learn*[5], modificando a implementação do algoritmo DBSCAN[6] [Ester et al. 1996] seguindo os passos apresentados no Algoritmo 1. Basicamente, inserimos um conjunto de novos dados em uma base previamente agrupada. Em seguida, atualizamos os *clusters* mais próximos aos novos dados. Neste processo, dados classificados como ruídos podem ser classificados como informações de um mesmo evento devido à mudança dos centróides. Ressaltamos que em nossa abordagem outros algoritmos de agrupamento podem ser utilizados, como o BIRCH [Zhang et al. 1996], desde que sejam adaptados para serem incrementais.

Algoritmo 1 Algoritmo de Agrupamento Incremental

1: Entrada: Conjunto de novos dados, eps, $MinPts$
2: Saída: Lista de *clusters* atualizada
3: $D \leftarrow$ Novos dados
4: $C \leftarrow$ Lista de *clusters*
5: **para** cada novo dado d_i em D **faça**
6: Insere d_i na base
7: **se** para cada centróide $dist$(centróide, d_i) $\leq eps$ **então**
8: Atualiza centróide
9: **fim se**
10: Atualiza *clusters* em C
11: **fim para**
12: **para** cada *cluster* atualizado c_i em C **faça**
13: **se** $|N_{Eps}$(dado central em $c_i)| \leq MinPts$ **então**
14: Define c_i como um *cluster* válido
15: **senão**
16: Define c_i como ruído
17: **fim se**
18: **fim para**

[5] http://scikit-learn.org/

[6] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

5. Experimentos e Resultados

Nesta seção, descreventos os experimentos, resultados e avaliações do algoritmo de agrupamento incremental para a detecção de eventos. Também apresentamos um estudo de caso em que aplicamos a solução proposta para a identificação e contagem de eventos relacionados a acidentes de trânsito.

5.1. Descrição dos Dados

Para estudo de caso e validação do método, coletamos 135.078 *tweets* de março de 2015 a setembro de 2016, relacionados a acidentes de trânsito na cidade de Nova Iorque.

Conforme apresentado na Figura 2, informações coletadas de redes sociais possuem diferentes formatos e características. No Twitter, enquanto os *tweets* de usuários comuns não possuem um padrão, *tweets* de canais de notícia possuem sempre a mesma formatação, na qual são reportados o local (regiões e rotas) e o tipo de ocorrência. Em decorrência disso, diversas soluções utilizam apenas os dados provenientes de canais de notícia para a detecção de eventos [Albuquerque et al. 2015].

(a) *Tweet* de canal de notícia (b) *Tweet* de usuário comum

Figura 2. Exemplos de *tweets* relacionados a acidentes de trânsito.

Para a solução proposta, *tweets* de canais de notícia são mais relevantes do que os *tweets* de usuários comuns, devido à etapa de cálculo da frequência dos termos. Na Tabela 1, apresentamos um exemplo do grau de importância de alguns termos. De maneira geral, termos relacionados aos nomes das vias são considerados os mais relevantes, enquanto termos comuns são considerados menos relevantes. Na Tabela 1, por exemplo, verificamos que as rotas interestaduais, como a I-287, e avenidas são consideradas relevantes para o algoritmo. Desta forma, sabemos que as vias são as características mais importantes para o algoritmo de agrupamento incremental.

Termos relevantes	Frequência de Termos	Termos não relevantes	Frequência de Termos
I-287	4.9512	EB (Eastbound)	2.1180
I-495	4.9512	Blocked	2.3486
8th (Ave)	4.9512	Street	2.3486
Route-24	4.9512	Lane	2.4663
14th (Ave)	4.9512	SB (Southbound)	2.5089

Tabela 1. Relevância dos termos presentes nos dados de acidentes de trânsito do Twitter por meio de frequência de termos.

Antes de aplicar o algoritmo que calcula a frequência dos termos, realizamos algumas modificações nos dados, como a remoção de *stop words*.

5.2. Avaliação de similaridade

Com o objetivo de avaliar a similaridade do algoritmo de agrupamento, realizamos o agrupamento manual de uma amostra da base de dados com 3.410 *tweets* de 27/01/2016 a 21/02/2016). Desta forma, foi possível comparar os agrupamento por dia e hora.

Na Tabela 2, apresentamos um exemplo de ocorrência de acidente de trânsito em que os dados foram agrupados manualmente no mesmo *cluster*. No processo de agrupamento manual, consideramos certas características, como as regiões, vias e horário da ocorrência reportada.

Hora	Ocorrências agrupadas
15:20	Accident in #TheBronx:OnTheDeeganExpwy on I-87 NB at W 230th St
16:30	Major accident on I-87 N #NYCTraffic
16:41	Accident in #TheBronx:OnTheDeeganExpwy on I-87 NB between W 230th St and Van Cortlandt Park S
17:01	Major accident on I-87 N #NYCTraffic
17:23	Accident in #TheBronx:OnTheDeeganExpwy on I-87 NB between W 230th St and Van Cortlandt Park S
17:31	Accident on I-87 N #NYCTraffic

Tabela 2. Exemplo de dados sociais agrupados relacionados à mesma ocorrência de evento de acidente de trânsito.

Durante este processo, verificamos que os informes de acidentes eram relacionados com vias expressas e avenidas. Além disso, verificamos que o maior número de informes ocorria na hora do *rush*. Isto pode ser explicado pelo fato de que usuários geralmente reportam eventos em redes sociais que são importantes ou tem impacto em seu contexto. Desta forma, acidentes de trânsito que ocorrem nas vias mais utilizadas e em horários de pico, tendem a ser mais impactantes do que outros acidentes.

Na Figura 3, apresentamos a comparação de ocorrências de acidentes de trânsito entre o agrupamento manual e o agrupamento incremental. Com o objetivo de comparar os agrupamentos, aplicamos a métrica *V-Measure*, obtendo 90% de similaridade.

Após a comparação, analisamos em diferentes períodos do dia o tempo médio utilizado pelo algoritmo para que os *clusters* sejam formados (Figura 4). Basicamente, estes tempos médios são obtidos pela diferença de tempo entre a primeira notificação do acidente, geralmente considerada um ruído por ser diferente dos demais, até a formação do *cluster*. Assim, verificamos que eventos de alto impacto, que em sua maioria ocorrem em horários de pico, são formados em menor tempo que os demais eventos de acidentes de trânsito.

Para exemplificar a formação de um *cluster* ao longo do tempo, apresentamos a Figura 5, em que dividimos a formação em 3 etapas, de forma que cada *cluster* é repre-

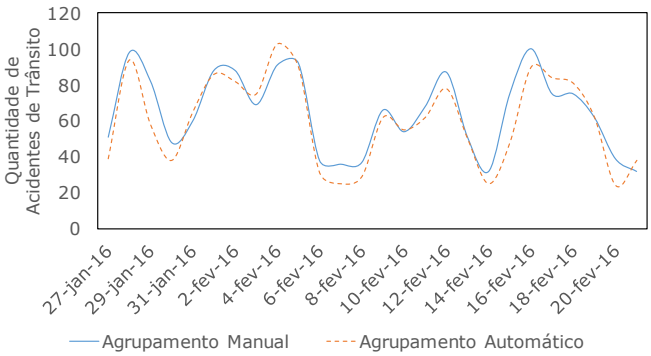


Figura 3. Comparação das ocorrências de acidentes de trânsito identificadas no agrupamento incremental e no agrupamento manual.

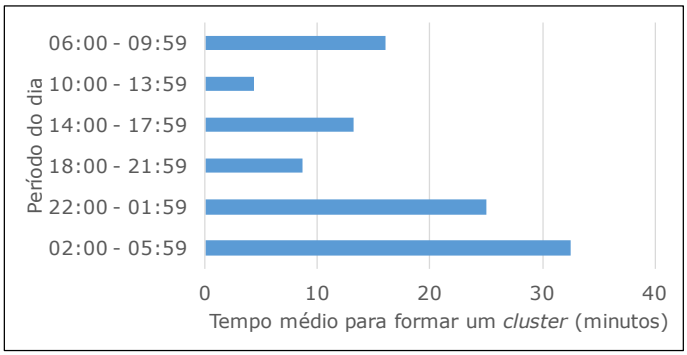
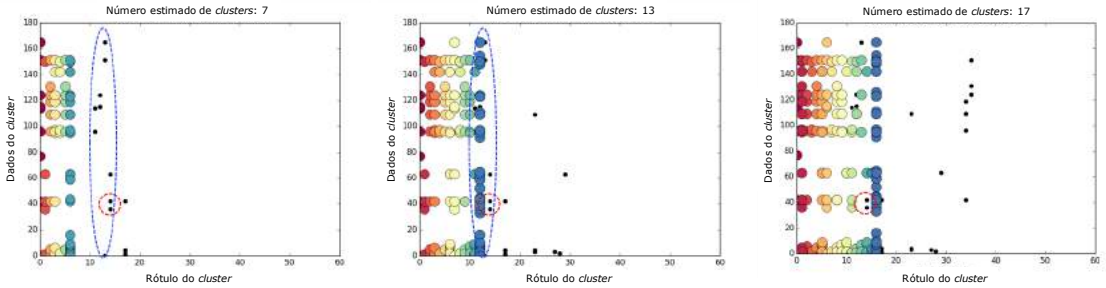


Figura 4. Tempo médio de formação dos *clusters* em diferentes períodos do dia.

sentado por uma cor diferente. Já os dados considerados ruídos são representados por pequenos pontos pretos. Nas Figuras 5(a) e 5(b) é possível verificar o momento em que o algoritmo converte dados considerados ruídos (círculo azul pontilhado), em um *cluster* que representa uma ocorrência de acidente de trânsito após a inserção de novos dados na base. No entanto, verificamos que em alguns casos o dado permanece como ruído mesmo com a inserção de novos dados (círculo vermelho pontilhado), o que ocorre devido à falta de similaridade destes com os demais dados da base (Figura 5(c)).



(a) (b) (c)

Figura 5. Exemplo de como ocorre a formação de *clusters* ao longo do tempo utilizando o método proposto.

5.3. Avaliação do tempo de execução

Para avaliar o tempo de execução do algoritmo, aplicamos tanto a abordagem incremental quanto a abordagem estática no estudo de caso de detecção de acidentes de trânsito. Cada vez que havia inserção de novos dados, o algoritmo de agrupamento estático processava todos os dados da base, enquanto que o algoritmo de agrupamento incremental processava apenas os novos dados e os dados da base afetados pela inserção.

Conforme apresentado na Figura 6, este comportamento impactou no tempo de execução, de forma que, com 135.078 *tweets* na base, o algoritmo de agrupamento incremental executava cerca de 4 vezes mais rápido que o estático. Desta forma, verificamos que para abordagens relacionadas às redes socais utilizar algoritmos tradicionais é custoso, sendo necessária uma abordagem incremental.

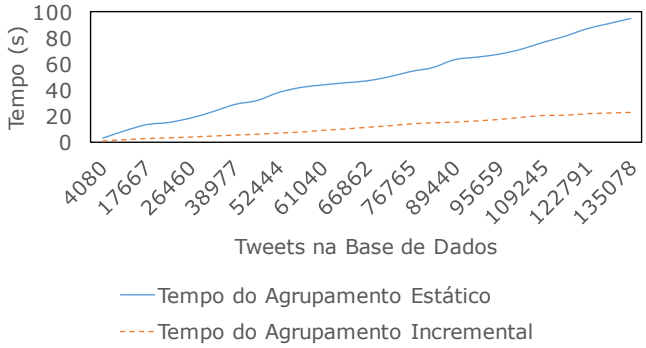


Figura 6. Comparação entre o tempo de agrupamento das abordagens incremental e estática.

6. Conclusão e Trabalhos Futuros

Neste trabalho, investigamos o processo de detecção de eventos em tempo real nas redes sociais através de uma abordagem de agrupamento incremental. Na solução proposta, consideramos cada usuário da rede social Twitter como um sensor que compartilha dados contextuais voluntariamente.

Todos os dados coletados foram filtrados para evitar o máximo de ruídos. As características dos textos presentes nos dados foram extraídas através de um algoritmo que calcula a frequência dos termos. Em seguida, um algoritmo de agrupamento incremental foi aplicado na base de dados de acordo com a inserção de novos dados. Como estudo de caso, executamos a solução proposta em uma base de dados do Twitter para a detecção de acidentes de trânsito, que possuem relevância social devido aos danos e mortes causados por eles. Nossa solução obteve 90% de similaridade e tempo de processamento cerca de 4 vezes mais rápido que a abordagem tradicional.

Como trabalhos futuros, planejamos comparar nossa técnica de agrupamento incremental com outras técnicas da literatura, além de analisar outras bases de dados de redes sociais com diferentes características, com o objetivo de avaliar a performance do algoritmo de agrupamento incremental para a detecção de outros eventos.

7. Agradecimentos

Parte dos resultados apresentados nessa publicação foram obtidos por meio de atividades de Pesquisa e Desenvolvimento do projeto SAMSUNG OCEAN, patrocinado pela Sam-

sung Eletrônica da Amazônia Ltda., apoiado pela SUFRAMA sob os termos da lei federal Nº 8.248/91.

8. Referências

Albuquerque, F. C., Casanova, M. A., Lopes, H., Redlich, L. R., de Macedo, J. A. F., Lemos, M., de Carvalho, M. T. M., and Renso, C. (2015). A methodology for traffic-related twitter messages interpretation. *Computers in Industry*.

Atefeh, F. and Khreich, W. (2015). A survey of techniques for event detection in twitter. *Comput. Intell.*, 31(1):132–164.

Becker, H., Naaman, M., and Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In *International Conference on Weblogs and Social Media*. AAAI.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Li, R., Lei, K. H., Khadiwala, R., and Chang, K. C.-C. (2012). Tedas: A twitter-based event detection and analysis system. In *International Conference on Data Engineering (ICDE)*, pages 1273–1276. IEEE.

Nguyen, H., Liu, W., Rivera, P., and Chen, F. (2016). Trafficwatch: Real-time traffic incident detection and monitoring using social media. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 540–551. Springer.

Phuvipadawat, S. and Murata, T. (2010). Breaking news detection and tracking in twitter. In *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 120–123. IEEE.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. (2009). Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51. ACM.

Silva, T. H., Vaz de Melo, P. O., Almeida, J. M., Salles, J., and Loureiro, A. A. (2014). Revealing the city that we cannot see. *ACM Transactions on Internet Technology (TOIT)*, 14(4):26.

Toriumi, F. and Baba, S. (2016). Real-time tweet classification in disaster situation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 117–118. International World Wide Web Conferences Steering Committee.

Valkanas, G. and Gunopulos, D. (2013). How the live web feels about events. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 639–648. ACM.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: an efficient data clustering method for very large databases. In *ACM Sigmod Record*, volume 25, pages 103–114. ACM.