Análise espaço-temporal de mensagens do Twitter

Renata de J. Silva[1], Luis Otavio Alvares[1]

[1]Depto. Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88.040-900 – Florianópolis – SC – Brasil

{renatadej.silva,alvares}@inf.ufsc.br

Abstract. *We live in an era in which access to the Internet becomes increasingly common. Social networks such as Twitter microblog, have recorded a significant increase of posted messages. Some of these messages have the geographical coordinates of the location where they were issued. This paper proposes the analysis of these posts considering the spatio-temporal aspects in order to obtain knowledge about users of Twitter. For this, we propose changes in the Weka data mining toolkit to obtain better results on twitter data analysis. Experiments were performed with real data obtaining good results.*

Resumo. *Estamos vivenciando uma era em que o acesso à internet torna-se cada vez mais frequente. As redes sociais, como o microblog Twitter, tem registrado um aumento significativo de mensagens postadas. Parte destas mensagens possui as coordenadas geográficas do local de onde foram emitidas. Este artigo propõe a análise destas mensagens considerando os aspectos espaço-temporais de modo a obter conhecimento sobre os usuários do Twitter. Para isto, propõe adaptações na ferramenta de data mining Weka de forma a obter melhores resultados. Experimentos foram realizados com dados reais, com bons resultados.*

1. Introdução

Com a popularização das redes sociais na internet, a disseminação e o acesso à informação tornou-se muito mais ágil. Uma dessas redes, o Twitter[1], na verdade mais considerado um microblog do que uma rede social, tem características particulares: suas mensagens são limitadas a 140 caracteres e usualmente são postadas de dispositivos móveis como celulares e *smartphones*; a maioria das mensagens reflete onde o usuário está, ou o que ele está fazendo ou sentindo naquele momento; para receber as postagens de um usuário (ser um seguidor) não há necessidade de concordancia deste usuário.

Com mais de 500 milhões de usuários [UOL Tecnologia 2012] e 500 milhões de mensagens por dia [Olhar Digital UOL 2012], o Twitter é uma fonte impressionante de informações. Entretanto, analisar milhões de dados publicados diariamente no Twitter é muito trabalhoso e inviável manualmente. Uma alternativa é aplicar técnicas de mineração de dados. Alguns estudos já abordam este problema, mas muito pouco existe que considere os aspectos espacial e temporal simultaneamente.

1 http://twitter.com. Último acesso em março 2013.

Este trabalho tem o foco em mineração de dados utilizando a base de dados do Twitter, com *tweets* – mensagens publicadas no Twitter – georreferenciados. São apresentadas adaptações na ferramenta Weka para que as análises de dados espaço-temporais dos *tweets* possam se tornar mais eficazes e eficientes. Mais especificamente, é abordada a técnica de formação de agrupamentos com o algoritmo DBSCAN [Ester et al 2006], cuja saída é incrementada com uma visualização na forma de mapas e a indicação das palavras mais frequentes nas mensagens de cada cluster, de modo a se ter uma ideia geral do conteúdo das mensagens.

O restante do artigo está organizado como segue: a seção 2 apresenta alguns trabalhos relacionados; a seção 3 apresenta o que é a ferramenta Weka; a seção 4 apresenta o que foi adaptado nesta ferramenta a fim de melhorar a capacidade de resposta às análises; na seção 5 são apresentados alguns experimentos realizados; e por fim a seção 6 expõe a conclusão e trabalhos futuros.

2. Trabalhos Relacionados

As redes sociais na internet são relativamente recentes e o volume de seus dados tem crescido exponencialmente nos últimos anos. A descoberta de conhecimento neste novo tipo de dado tem suscitado muito interesse e vários trabalhos tem abordado o tema. Entretanto, trabalhos considerando os aspectos espaço-temporais das mensagens postadas são bem menos numerosos. Por exemplo, no Twitter, a localização geográfica do local de postagem das mensagens passou a ser disponibilizada apenas em 2010. Alguns trabalhos que abordam a descoberta de conhecimento espaço-temporal em dados do Twitter são mencionados a seguir.

Como os usuários usam bastante o Twitter para informar a seus seguidores o que estão fazendo no momento, esta rede tem características de tempo-real. Sakaki em [Sakaki et al 2010] usou esta característica para a detecção de eventos naturais como terremotos e tufões, usando as mensagens do Twitter como sensores, analisando as palavras das mensagens.

Um sistema para a descoberta de atividades sociais fora do padrão é proposto em [Lee et al 2011]. É utilizado o algoritmo K-means. Cada grupo formado é analisado considerando comportamentos de agregação (usuários que estavam em outros locais e agora estão neste) e dispersão (usuários que estavam neste local e agora estão em outros). Um pico nos dados de agregação é um indício de um evento social. Outro trabalho nesta área, mas que refina o processo com uma análise visual interativa foi proposto recentemente [Chae et al 2012]. O artigo [Lee 2012] vai mais além, pois preve a possível evolução e impacto dos eventos detectados.

Com o presente trabalho, aplicando técnica de mineração de dados, foram detectadas regiões de grande concentração de *tweets*. Além disto, para obter conhecimento sobre o que os usuários do Twitter estão fazendo nestas regiões densas, foi utilizado um sistema de busca em texto para capturar as palavras mais frequentes.

3. A ferramenta Weka

Para a mineração e análise dos dados, utilizou-se o Weka [Witten & Frank 2005]. O Weka é uma ferramenta criada na Universidade de Waikato, Nova Zelândia, de código aberto, desenvolvido na linguagem de programação Java e muito utilizada nos meios acadêmicos. Esta ferramenta possui uma coleção de algoritmos para execução das tarefas de mineração de dados.

A técnica utilizada neste trabalho foi a de Agrupamento (*Clustering*) e o algoritmo aplicado foi o DBSCAN [Ester et al 2006], que é um algoritmo baseado em densidade, isto é, as regiões densas formam os *clusters*. Para ser considerada densa, uma região deve ter um número mínimo de pontos (parâmetro "minPoints") dentro de um círculo (parâmetro "epsilon", raio do círculo).

Na ferramenta Weka, após a execução do algoritmo DBSCAN, é possível visualizar os resultados como mostra o exemplo da Figura 1.

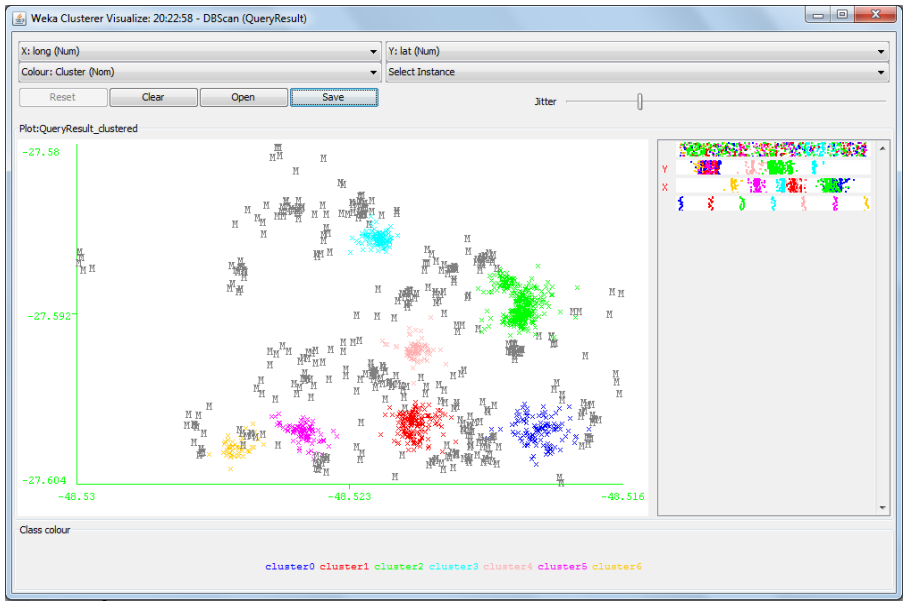

Figura 1. Visualização do DBSCAN na ferramenta Weka

Neste exemplo, é possível identificar os *clusters* que o algoritmo formou, neste caso 7. Os *clusters* são identificados pelas diferentes cores e, pode-se visualizar a posição de cada *cluster*, pois no eixo X foi plotado o atributo longitude e no eixo Y a latitude do ponto em que cada mensagem foi postada. Os pontos (*tweets*) que não pertencem a nenhum *cluster* aparecem na cor cinza e são considerados "ruído" ou *noise* pelo algoritmo.

O algoritmo DBSCAN, na ferramenta Weka, não possui recursos para trabalhar com dados geográficos. Desta maneira, é possível notar que seria difícil analisar este tipo de dado, pois não há informações geográficas, ou seja, não se tem como saber em que parte de uma cidade ou país está cada *cluster*. Para melhorar as análises, foram realizados melhoramentos no Weka, descritos na próxima seção.

4. Adaptações na Ferramenta Weka

Para que os resultados pudessem ser analisados de maneira mais ágil, sem que fosse necessário grande esforço humano, a ferramenta Weka foi adaptada. Para isto, foram desenvolvidos 2 recursos novos na ferramenta: (i) geração de um mapa onde cada marcador representa o centróide (centro de gravidade) de um *cluster*; (ii) com o clique de mouse em um marcador, podem ser visualizadas as palavras mais frequentes do *cluster* correspondente.

4.1. Geração de mapa com a API Google Maps

Para facilitar a análise de dados geográficos, optou-se por implementar a geração de um mapa real. Foi adicionado um método responsável por esta ação que é automaticamente executado durante a execução do algoritmo DBSCAN do Weka.

Para o desenvolvimento da geração do mapa, foi utilizada a API do Google Maps. Para a inserção de múltiplos pontos com ícones personalizados, foi utilizado como base o *script* do site *Link Nacional* [Link Nacional 2011]. Com isso, foi possível indicar latitude, longitude, ícone/marcador e descrição para cada *cluster*.

No mapa desenvolvido, cada marcador representa o centróide de um *cluster*. Isto foi feito porque plotar todos os pontos de um *cluster* iria poluir muito o mapa e, com isso, dificultaria a análise. Além disso, o conjunto dos pontos de um *cluster* já pode ser visualizado na interface padrão do Weka, se houver necessidade de se conhecer melhor a distribuição dos pontos do *cluster*.

O resultado da geração do mapa é um arquivo HTML. A Figura 2 é um exemplo de como os *clusters* são visualizados na interface que foi desenvolvida neste trabalho. A interface mostra os centróides dos *clusters* gerados, representados pelos marcadores, que também identificam o período da mensagem (manhã, tarde ou noite) conforme a sua cor, e se foram postados em dias de semana ou nos fins de semana. No exemplo da Figura 2, todos os *clusters* são de mensagens postadas no período da manhã nos dias de semana.

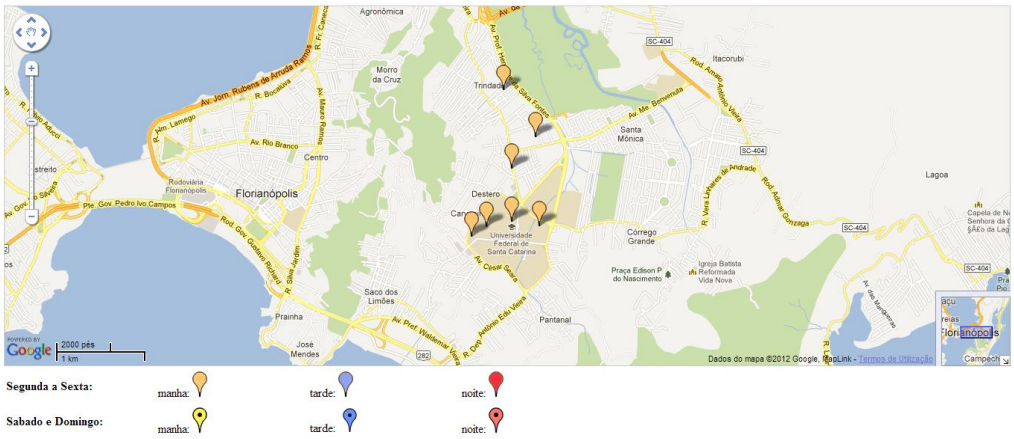

Figura 2. Visualização do mapa gerado

Conforme pode ser observado na Figura 2, o mapa desenvolvido facilita a análise de dados geográficos, pois é possível identificar onde cada *cluster* está situado no espaço, ou seja, é possível visualizar sobre qual local o *cluster* está localizado e também verificar nomes de ruas e bairros, a existência de rios, morros, etc.

No mapa gerado, os recursos do Google Maps podem ser utilizados (por exemplo, utilizar o *zoom*) e, além disso, foi implementada uma legenda com informações dos marcadores.

4.2. Obtenção de Palavras Frequentes

Para conhecer melhor o que os usuários do Twitter estão fazendo, foi implementada uma funcionalidade que captura as palavras mais frequentes das mensagens de cada agrupamento formado. Para isto, foi utilizada a biblioteca de tratamento de texto Tsearch2 [Bartunov & Sigaev 2012], que é uma extensão do PostgreSQL, desenvolvida

na Universidade de Moscou. Optou-se por adaptar esta biblioteca em vez de desenvolver a funcionalidade, pois é uma tarefa complexa e que deve ser computacionalmente eficiente.

Antes de aplicar a função do Tsearch2, para a análise do texto do *tweet* em si, decidiu-se eliminar a acentuação ortográfica, para que a busca por texto encontrasse mais ocorrências de uma mesma palavra.

O Tsearch2 possui diversas opções para tratamento de texto, como eliminação de *stopwords*, *stemming*, etc. Para este estudo não foi realizado *stemming*, de modo que, por exemplo, os termos "casa" e "casarão" são considerados distintos. Foi utilizado o conjunto de *stopwords* (palavras ignoradas pelo sistema) referente à língua portuguesa, acrescentado de outras palavras observadas no decorrer do trabalho como irrelevantes para o estudo, como por exemplo, as letras isoladas e expressões como *4square*.

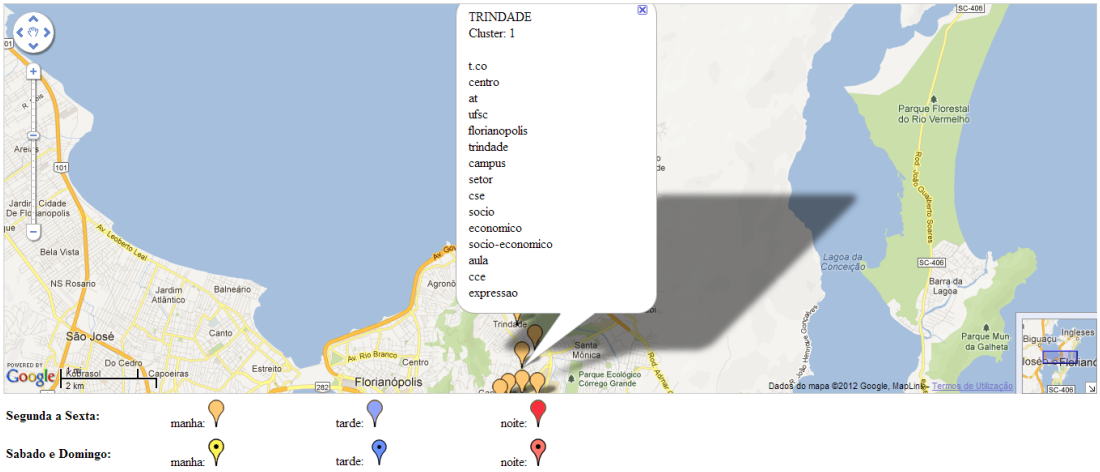

Figura 3. Visualização da interface com as palavras mais frequentes de um *cluster*

A Figura 3 apresenta a interface desenvolvida para a visualização das palavras mais frequentes nos *tweets* de um *cluster*. Basta clicar sobre um marcador para que a lista das palavras mais frequentes no cluster representado pelo marcador seja apresentada.

5. Experimentos Realizados

Para avaliar a eficácia das extensões realizadas foram realizados experimentos com *tweets* postados na cidade de Florianópolis, no período de abril a novembro de 2011, e contendo as coordenadas geográficas do local em que foram postados. Os *tweets* com as coordenadas geográficas corresponderam a aproximadamente 10% dos *tweets* emitidos. O SGBD utilizado foi o PostgreSQL. A escolha deste SGBD foi feita por este permitir a manipulação de dados geográficos por meio da extensão PostGIS, que segue o padrão OGC [OGC 2008].

As informações mais relevantes contidas na base de dados são: latitude e longitude (ambas do tipo "double"), data/hora de postagem da mensagem (tipo "timestamp") e texto do *tweet* propriamente dito (tipo "text").

Como havia a intensão de realizar a análise dos *tweets* por bairro de Florianópolis, uma primeira preparação dos dados foi a determinação do bairro em que os *tweets* foram postados. Para isto, inicialmente, foi criada na tabela de *tweets* uma

coluna de tipo “geometry”, necessária para a utilização das funções espaciais do PostGIS. Assim, para cada *tweet* foi gerado um tipo geométrico “ponto”, por meio da função “ST_MakePoint” do PostGIS, aplicada aos campos latitude e longitude.

Em seguida, foi utilizado um arquivo *shapefile* disponível no *site* do Instituto Brasileiro de Geografia e Estatística (IBGE) para a obtenção dos limites dos bairros de Florianópolis. Desta maneira, foi empregada a função “ST_Contains” no PostGIS para o cruzamento da tabela de bairros e tabela de *tweets* para, enfim, popular a coluna com a informação do bairro em que o *tweet* foi postado.

Entre os 35 bairros de Florianópolis, 15 foram desconsiderados para a pesquisa em função do pequeno número de postagens realizadas. O total de registros (*tweets*) analisados foi de 152.552.

O conjunto de dados foi filtrado por 3 atributos: bairro, período do dia (manhã, tarde ou noite) e dia da semana (segunda a sexta-feira ou sábado e domingo), totalizando 6 consultas por bairro.

O algoritmo DBSCAN foi executado, inicialmente, de maneira padrão para os 20 bairros de estudo (minPoints = 2,5% dos *tweets* do bairro, epsilon = 0,045). Isto quer dizer que, para cada bairro, o algoritmo DBSCAN foi executado 6 vezes, totalizando 120 consultas no banco de dados aplicadas ao software Weka. Para que o estudo não se tornasse cansativo e para evitar o trabalho manual, o código da ferramenta Weka foi adaptado para automatizar estas execuções das consultas.

Como o algoritmo utilizado neste trabalho foi o DBSCAN, e o mesmo tem por característica gerar grupos em regiões densas, existe uma grande possibilidade de os pontos centrais gerados por *cluster* estejam sobre, ou muito próximos, a locais atrativos. Por exemplo, universidades, restaurantes, bares, shoppings centers, estádios de futebol, centros comerciais, empresas, entre outros. Isto pode ser percebido ao visualizar os marcadores plotados pelo Weka na interface desenvolvida neste trabalho.

Exemplos dos resultados obtidos com essa análise são mostrados nas Figuras 4 e 5. A Figura 4 apresenta os centróides de *clusters* no bairro Trindade. Pode-se observar que muitos *clusters* estão no campus da UFSC durante os dias de semana (marcado pelo círculo) e que nas noites de finais de semana muitos clusters são formados no entorno da UFSC (marcadores com ponto preto), que é uma região de muitos bares.

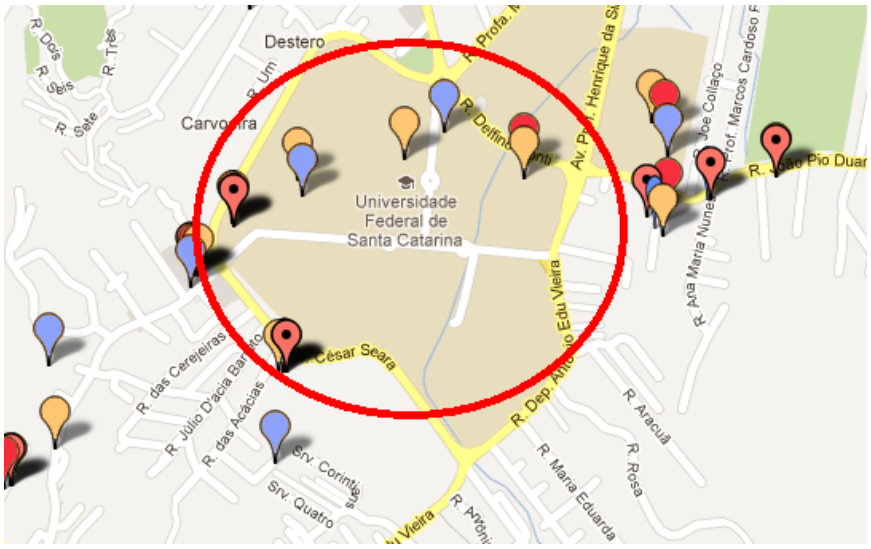

Figura 4. *Clusters* formados na UFSC e entorno

A Figura 5 apresenta *clusters* formados no estádio de futebol Orlando Scarpelli (marcado com o círculo) nas tardes e noites de finais de semana, o que deve corresponder a jogos sábados à noite e domingos à tarde. Além disso, esta figura apresenta as palavras mais frequentes encontradas em um dos *clusters*. Pode-se notar que estas palavras estão relacionadas a futebol.

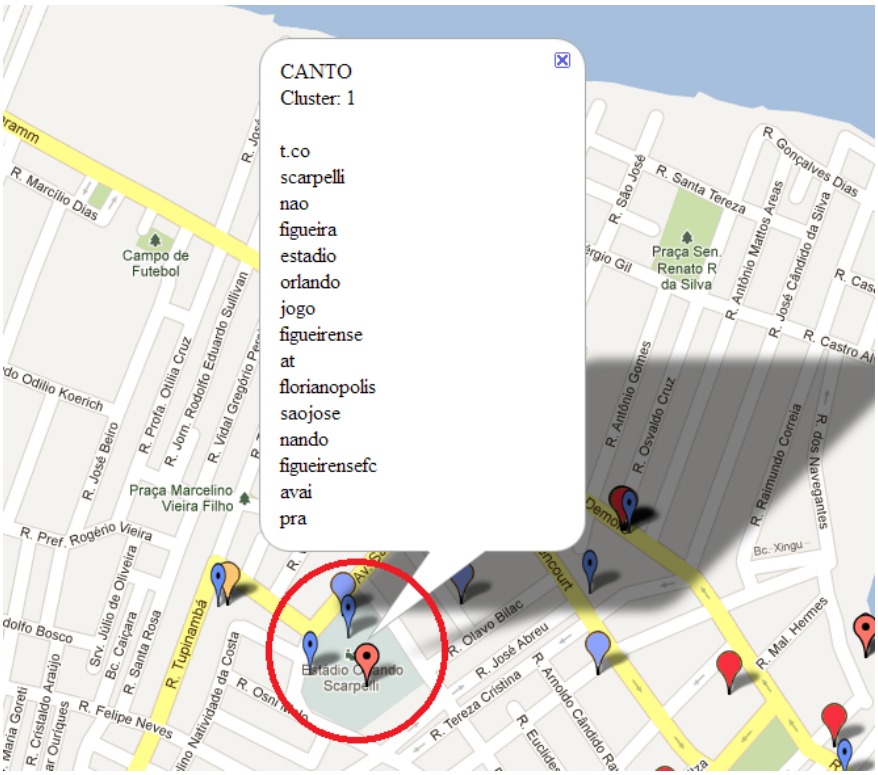

Figura 5. *Clusters* formados no estádio de futebol *Orlando Scarpelli*

Se para os demais bairros os parâmetros utilizados foram razoáveis, para o bairro Centro, a maioria das consultas gerou somente um *cluster* situado no meio deste bairro (Figura 6). Os centróides ficaram aproximadamente no meio do bairro porque os dados eram muito numerosos e geograficamente homogêneos. A Figura 7 apresenta os *tweets* plotados no mapa, visualizado pela ferramenta Quantum GIS (<http://www.qgis.org>). Estes dados são somente do bairro Centro no período da tarde no intervalo de segunda a sexta-feira, totalizando 11.315 registros.

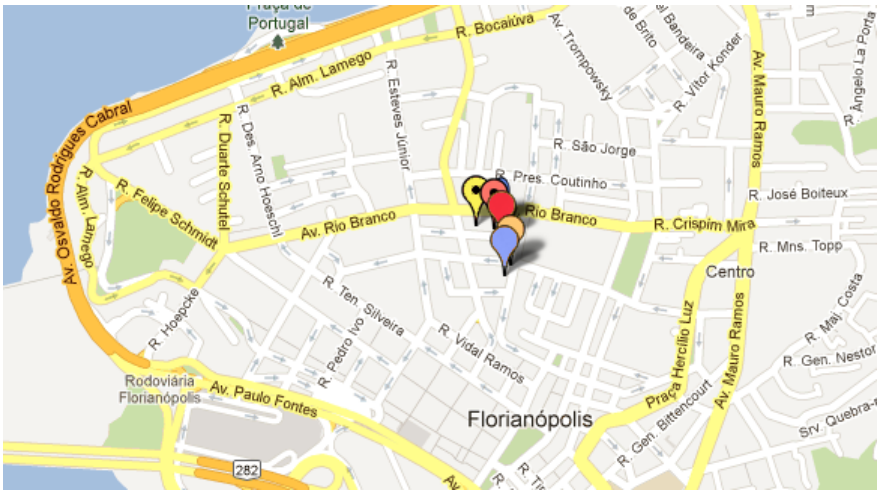

Figura 6. *Clusters* formados no bairro Centro

Para que o algoritmo DBSCAN possa gerar mais *clusters*, neste caso, é necessário diminuir o valor dos parâmetros "epsilon" e "minPoints". Por conseguinte, no bairro Centro, o algoritmo DBSCAN foi executado com diferentes valores de atributos, "épsilon" e "minPoints", até se tornar possível a descoberta de locais de interesse. Dois destes experimentos são detalhados na sequencia.

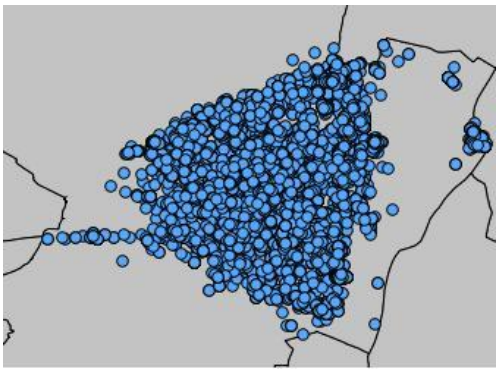

Figura 7. Visualização dos tweets do Centro através da ferramenta QuantumGIS

Para o experimento 1, foram utilizados os parâmetros: (i) minPoints = 1,25%; (ii) Épsilon = 0,011. Em relação às análises dos demais bairros, o número mínimo de pontos utilizado foi reduzido pela metade e o épsilon representou a quarta parte do valor utilizado nos experimentos com os outros bairros.

Com os parâmetros do experimento 1, foi possível identificar locais de interesse como: (i) Terminal de ônibus urbanos (TICEN) – períodos manhã e tarde nos dias de semana e noite tanto de dias de semana quanto de fins de semana; (ii) Instituto Estadual de Educação (IEE) – manhã e tarde de dias de semana; (iii) Praça XV de Novembro – manhã de fim de semana; (iv) Catedral Metropolitana de Florianópolis – tarde e noite de fim de semana; (v) Beiramar Shopping – manhã e tarde de dias de semana e fins de semana; (vi) Boate El Divino – tarde e noite de fins de semana; (vii) Boate 1007 – manhã e noite de fins de semana; (viii) Mercado Público – tarde de fins de semana; (ix) Morro da Cruz – manhã de fins de semana; (x) Instituto Federal de Santa Catarina (IF-SC) – tarde de dias de semana; (xi) Centro executivo localizado na Avenida Mauro Ramos – manhã de dias de semana, etc.

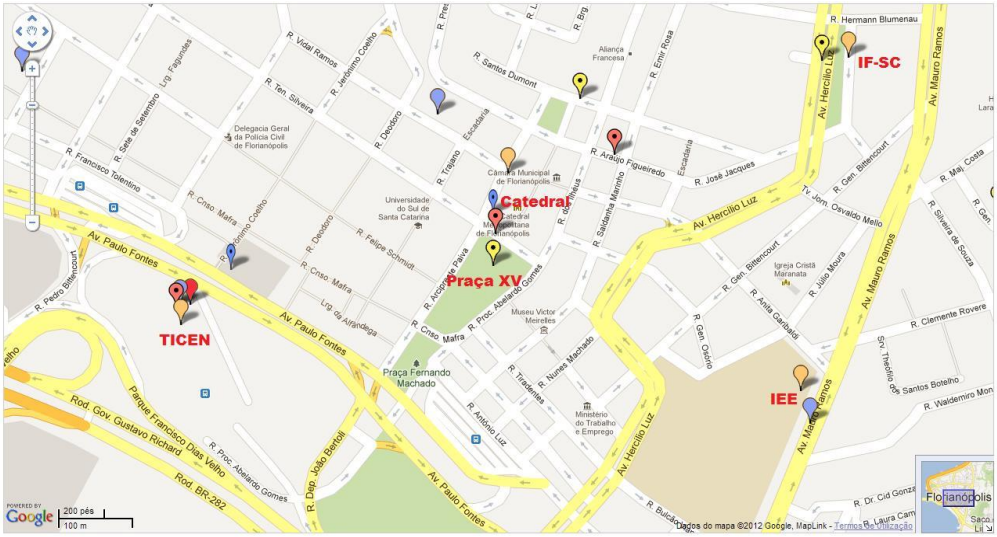

Figura 8. Resultado parcial da execução do experimento 1

A Figura 8 apresenta um *zoom* em parte do Centro com o resultado da execução do experimento 1. É possível identificar alguns dos locais citados, como o TICEN, a Catedral, o IF-SC, o IEE e a Praça XV.

Para o experimento 2, os parâmetros foram reduzidos radicalmente, com o objetivo de detectar um maior número de *clusters* que poderiam ser pequenos, em locais específicos. Foram utilizados os valores minPoints = 45 e epsilon = 0,0005.

Em relação à análise anterior (o experimento 1) foi observado, entre outros: (i) na Praça XV de Novembro (que foi considerada uma região densa no experimento anterior), por possuir uma área relativamente grande, não foi identificada como uma região densa, justamente por os *tweets* estarem mais distantes entre si neste local; (ii) alguns locais não detectados com análises anteriores puderam ser encontrados nesta análise, como a Universidade do Sul de Santa Catarina (UNISUL) e Terminal Rodoviário Rita Maria. Alguns clusters em locais residenciais também foram encontrados.

No mapa apresentado na Figura 9, gerado pela execução do experimento 2, é possível identificar alguns dos locais citados, como o Terminal Rita Maria e a UNISUL.

Figura 9. Resultado parcial da execução do experimento 2

6. Conclusão e Trabalhos Futuros

Mineração de dados espaço-temporais, com foco em detecção de agrupamentos, em redes sociais não é um assunto muito explorado. Esta pesquisa buscou conhecer o comportamento dos usuários do Twitter – especificamente na cidade de Florianópolis. De acordo com o conhecimento extraído, podem-se tirar conclusões do interesse da população e, analisar o que estão fazendo em determinados locais e horários. Por exemplo, donos de empresas, tendo acesso a estas informações, podem analisar a satisfação de colaboradores e/ou clientes. Este tipo de pesquisa poderá contribuir, por exemplo, com pesquisas de marketing, e consequentemente, aumentar a segurança dos resultados.

Para atender a proposta deste artigo, o código da ferramenta Weka foi adaptado. Isto tornou o *software* uma excelente ferramenta também para visualização. Desta maneira, o resultado encontrado pelo algoritmo DBSCAN pode ser melhor analisado.

Na implementação atual, pode ocorrer de um *cluster* ser formado apenas por *tweets* de um único usuário. Como trabalho futuro pode ser interessante tratar os dados

para desconsiderar SPAMs, ou impedir a formação de um *cluster* se ele não contiver um número mínimo de usuários distintos.

Também se pretende permitir ações de usuário nos mapas gerados, como limpar *clusters* já populados no mapa e aplicar filtros para visualizar somente *clusters* de interesse. Permitir também mais flexibilidade ao usuário, adicionandos componentes na interface gráfica com este fim.

Referencias

Chae, J. Thom, D ; Bosch, H. ; Jang, Y. ; Maciejewski, R. ; Ebert, David S. ; Ertl, T. (2012) "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition". IEEE Conference on Visual Analytics Science and Technology (VAST). p 143-152.

Bartunov; Sigaev (2012). "Tsearch2 - full text extension for PostgreSQL". http://www.sai.msu.su/~megera/postgres/gist/tsearch/V2/. Acessado em 30 de dezembro de 2012.

Ester, M.; Kriegel, H.-P.; Sander, J. and Xu, X. (1966) A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. M. Fayyad, editors, Second International Conference on Knowledge Discovery and Data Mining, AAAI Press. p. 226-231.

Lee, C-H. (2012) Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams. Expert Syst. Appl. 39(18), p 13338-13356 .

Lee, C-H.; Yang, H.C.; Wen, W-S.; Weng, C-H. (2012) Learning to Explore Spatio-temporal Impacts for Event Evaluation on Social Media. ISNN (2), p 316-325.

Link Nacional. (2011). "Script de Múltiplos Pontos", http://www.linknacional.com.br/criar-site/2011/01/google-maps-api-multiplos-pontos-no-mapa-openinfowindowhtml. Acessado em 30 de dezembro de 2012.

OGC (2008) OpenGIS Standards and Specifications: Topic 5, Features. http://portal.opengeospatial.org/modules/admin/licenseagreement.php?suppressHeaders=0&accesslicense

Olhar Digital UOL. (2012) "Twitter gera meio bilhão de mensagens por dia", http://olhardigital.uol.com.br/jovem/redes_sociais/noticias/twitter-gera-meio-bilhao-de-tuites-por-dia. Acessado em 30 de dezembro de 2012.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes Twitter users: realtime event detection by social sensors. In Proceedings of the 19th international Conference on World Wide Web - WWW '10. ACM, New York, NY, p 851-860.

UOL Tecnologia. (2012) "Twitter passa dos 500 milhões de usuários, mas números mostram queda de microblog no Brasil", http://tecnologia.uol.com.br/noticias/redacao/2012/07/31/twitter-passa-dos-500-milhoes-de-usuarios-mas-numeros-mostram-queda-de-microblog-no-brasil.htm, Julho. Acessado em 30 de dezembro de 2012.

Witten, I. and Frank, E. (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco.