

Observatório do Trânsito: sistema para detecção e localização de eventos de trânsito no Twitter

Sílvio S. Ribeiro Jr., Diogo Rennó, Tatiana S. Gonçalves,
Clodoveu A. Davis Jr., Wagner Meira Jr., Gisele L. Pappa

Universidade Federal de Minas Gerais
{silviojr, renno, tati.sg, clodoveu, meira, glpappa}@dcc.ufmg.br

Resumo. O Twitter se consolidou como uma plataforma popular para fornecer conteúdo gerado por usuários, que varia de simples conversação a informação em tempo real sobre eventos recentes. Muitas pesquisas demonstraram que o conteúdo produzido no Twitter possui alto grau de correlação com o que ocorre no mundo real, o que levou ao desenvolvimento de aplicações em diversas áreas, abrangendo desde epidemias até eleições. Nosso trabalho é baseado no fato de que existe muita informação sobre trânsito disponível no Twitter, principalmente de perfis especializados, criados para coletar e divulgar notícias sobre eventos de trânsito em algumas grandes cidades. Neste artigo, propomos um método para, dado um evento, geolocalizá-lo a partir do conteúdo dos *tweets* assim que são coletados. Os resultados mostram que conseguimos localizar bairros e logradouros com um grau de acerto que varia de 50 a 90%, dependendo do número de lugares mencionados nos tweets.

Categories and Subject Descriptors: H.3 [Information Systems]: Information Storage and Retrieval

General Terms: Experimentation

Keywords: Trânsito, twitter, localização, geocodificação

1. INTRODUÇÃO

O Twitter se tornou uma plataforma popular para usuários que geram conteúdo, que varia desde assuntos corriqueiros até a distribuição de informações sobre eventos em tempo real. Muitas pesquisas já mostraram que o conteúdo produzido no Twitter possui um alto grau de correlação com o mundo real, o que levou ao desenvolvimento de aplicações que cobrem desde epidemias até eleições [Pappa et al. 2010].

Tweets a respeito do trânsito são comuns. Muitos usuários usam o Twitter para informar sobre dificuldades de movimentação dentro da cidade, transmitindo dados sobre as condições do trânsito e reportando problemas como acidentes, obras, veículos com problemas mecânicos, passeatas e outros. Existem perfis no Twitter criados especialmente para informar as condições do trânsito em grandes cidades, alguns inclusive operados por órgãos oficiais de trânsito, constituindo fontes úteis de informação para motoristas que têm acesso às mensagens publicadas por essas contas. Há, portanto, uma grande quantidade de informação sobre trânsito disponível no Twitter, embora de forma desestruturada e espalhada. Nesse cenário, usuários que voluntariamente informam sobre as condições do trânsito estão atuando como sensores de um fenômeno que se desenvolve em tempo real. Existe, portanto, grande interesse no desenvolvimento de métodos, técnicas e ferramentas que consigam captar essas contribuições, organizá-las e publicá-las de forma integrada e consistente.

A informação coletada pode ser usada tanto de forma complementar àquela gerada por câmeras e

sensores físicos, orientando as ações dos agentes públicos a curto e longo prazo, quanto diretamente pelos motoristas, em tempo real ou quase-real, apoiando suas decisões quanto ao deslocamento pela cidade. A crescente popularidade desse tipo de canal de informação indica que, em pouco tempo, informações sensoriadas e transmitidas pelos próprios cidadãos podem se tornar a principal fonte de avaliação da situação do trânsito em tempo real.

Neste artigo, descrevemos a fase inicial de estudo e implementação do Observatório do Trânsito. O Observatório do Trânsito é um sistema de mineração de texto que trabalha sobre o *stream* de tweets, procurando por padrões de texto relevantes que indiquem a situação do trânsito em lugares definidos. Além disso, criamos uma interface Web para reportar a informação coletada para usuários. Para testarmos nosso método, ele foi aplicado na cidade de Belo Horizonte. Os resultados alcançados até o momento mostraram-se promissores na detecção de situações relevantes no trânsito e suas respectivas localizações.

O artigo está estruturado da seguinte forma, A Seção 2 lista trabalhos relacionados ao projeto desenvolvido. A Seção 3 descreve o método proposto. A Seção 4 apresenta os conjuntos de dados utilizados, incluindo as bases de tweets e o *gazetteer* utilizado. A Seção 5 relata os experimentos feitos e resultados obtidos e, finalmente, a Seção 6 apresenta as conclusões e lista trabalhos futuros.

2. TRABALHOS RELACIONADOS

Geocodificação é o processo de transformação de dados descritivos sobre um local, tal como o nome do lugar, em uma referência geográfica absoluta [Goldberg et al. 2007]. Os *gazetteers* são uma importante fonte de informação para a realização de geocodificação, pois têm o propósito de associar nomes de lugares à sua localização. Mesmo utilizando um *gazetteer*, no entanto, a geocodificação é uma tarefa desafiadora, porque existem ambiguidades entre lugares (i.e., muitos lugares com o mesmo nome) ou entre lugares e coisas (i.e., nomes de lugares iguais aos nomes de outras entidades) [Daniel 2010]. Além disso, o uso de abreviações e simplificações, comuns em tweets devido às limitações de espaço para o texto, também complicam o reconhecimento dos nomes de lugares.

Algumas técnicas foram propostas para reconhecer e interpretar nomes de localidades em textos. Em [Twaroch et al. 2008], nomes de lugares são detectados em pesquisas na Web através de frases relacionadas ao contexto de localidades e de um *gazetteer* de referência. [Amitay et al. 2004] propõem um método para resolver a ambiguidade entre o nome de um lugar e os nomes de outros lugares ou palavras comuns. A abordagem utiliza um *gazetteer* hierárquico mundial para determinar uma localidade única com uma certa confiança, através de passos como encontrar possíveis nomes de lugares próximos a um nome já identificado ou procurar por um lugar comum (i.e., um país) relacionado a alguns dos nomes ambíguos (i.e., cidades). Uma abordagem baseada em *scores* também é proposta para identificar o foco de uma página da Web (o principal local sobre o qual ela trata). [Delboni et al. 2005] propõem reconhecer lugares relevantes em um texto encontrando expressões posicionais tais como “próximo a” ou “a cinco minutos de” e procurando ao redor delas na sentença. Visando melhorar a precisão da resposta para uma pesquisa contendo tal texto, a identificação de sinônimos para as expressões é proposta. [Cardoso et al. 2008] apresentam o protótipo de um sistema para recuperação de informações geográficas que busca capturar evidências geográficas implícitas, tais como nomes de empresas ou edifícios, e utilizá-las junto às evidências explícitas para melhorar os resultados do sistema.

[Cheng et al. 2010] propõem uma abordagem para localizar usuários do Twitter baseada no conteúdo dos tweets. Utilizando somente o texto das mensagens, eles desenvolveram um *framework* probabilístico para estimar a localidade de um usuário do Twitter a nível de cidade. Um classificador é utilizado para automaticamente identificar palavras nos tweets que sejam fortemente relacionadas a um escopo geográfico local, e a localidade dos usuários pode ser então estimada através de um modelo suavizador que busca as palavras identificadas nos tweets dos usuários.

Para o presente trabalho, a determinação da cidade de origem dos tweets é insuficiente para resolver a geocodificação. É necessário obter informações que permitam localizar o ponto na cidade ao qual a mensagem se refere. Uma pequena parcela dos tweets é produzida por dispositivos móveis que, caso autorizado pelo usuário, associam coordenadas GPS à mensagem, dentro dos limites de precisão do equipamento. Em muitos casos, no entanto, é necessário obter a localização através da interpretação da mensagem, buscando nela referências a lugares intra-urbanos, como ruas, avenidas e pontos de referência. O uso de um gazetteer, nessa situação, só é possível se o mesmo contiver detalhamento nesse nível, o que não é o usual [Machado et al. 2011].

3. LOCALIZAÇÃO DE EVENTOS DE TRÂNSITO

O método proposto neste artigo para detecção e localização de eventos relacionados ao trânsito está dividido em quatro etapas: (i) pré-processamento do texto do tweet, (ii) identificação do evento, (iii) localização por casamento exato, (iv) enriquecimento da localização por casamento aproximado.

O pré-processamento do texto do tweet inclui a remoção de acentos, *links* e citações a perfis do Twitter (e.g. @BHTrans). Nessa fase, postagens referentes a outras cidades também são identificadas e excluídas do restante do processo. Isso é necessário porque um dos perfis coletados (*WayTaxi*) aborda o trânsito em diversas capitais brasileiras, mas sempre fazendo referência explícita à capital em questão.

A segunda etapa identifica os eventos de interesse relacionados ao trânsito. Neste primeiro trabalho, especificamos os possíveis eventos e condições de trânsito manualmente. Até agora, utilizamos essa lista estática para assegurar o uso somente de tweets que dizem respeito ao trânsito. O conjunto de eventos utilizado atualmente é descrito na Tabela II. No futuro, propomos utilizar técnicas de aprendizado de máquina para detectar eventos automaticamente e de maneira dinâmica.

Dividimos as condições de trânsito em duas categorias principais: condição e evento. A condição refere-se à situação do trânsito em certo local em um dado momento (i.e., “lentidão”), enquanto eventos dizem respeito a acontecimentos que podem alterar o trânsito, afetando-o direta ou indiretamente (i.e., “acidente”). Tentamos abranger os eventos e estados tão bem quanto possível, mas temos ciência de que a lista não é exaustiva.

A terceira fase é o casamento exato, que utiliza um gazetteer para a geolocalização dos tweets, como descrito na Seção 3.1. Finalmente, a quarta etapa enriquece os dados da localização encontrada no passo anterior, como descrito na Seção 3.2.

3.1 Localização por casamento exato

Detectar referências a ruas e bairros no texto de um tweet é uma das principais tarefas do nosso sistema, como etapa necessária para localizar os eventos de trânsito. É necessário relacionar as condições de trânsito reportadas com os nomes oficiais de cada logradouro ou bairro para manter os relatórios consistentes para o usuário e possíveis análises futuras. Utilizamos um gazetteer que contém os nomes, geometria e geolocalização de logradouros, bairros, cruzamentos e trechos de vias de Belo Horizonte [Machado et al. 2011]. Esse gazetteer contém (1) o nome de 9.514 logradouros, incluindo a localização de cada segmento de logradouro entre cruzamentos, (2) a localização de 40.749 cruzamentos, associados aos nomes dos logradouros envolvidos, e (3) a posição do início e do fim de cada trecho de logradouro em cada bairro, estando disponíveis 47.211 trechos anotados.

Encontrar citações de locais no conteúdo de tweets, no entanto, é uma tarefa extremamente difícil. Como os tweets são limitados a 140 caracteres, os usuários tentam encurtar suas mensagens abreviando nomes muito utilizados. Além disso, algumas palavras são escritas incorretamente e alguns logradouros e bairros possuem nomes populares diferentes dos seus nomes oficiais.

Assim, além do gazeteer descrito acima, criamos um dicionário para nomes alternativos, em que cada nome oficial possui como sinônimos seus nomes populares. Criamos também um dicionário de formas abreviadas para os tipos de logradouros, como, por exemplo, “Av e Av.” para “Avenida” e “Vdt” para “Viaduto”. Esses dicionários foram usados em conjuntos com os nomes oficiais encontrados no gazeteer. O conjunto final de nomes será chamado de GEODIC.

Tendo como base o GEODIC, o método proposto procura inicialmente por nomes de logradouros e bairros utilizando casamento exato. Nessa etapa, os nomes contidos no GECODIC são procurados como *substrings* no texto do tweet. Para isso, existem duas estratégias: utilizar o nome do logradouro/bairro acompanhado da sua definição (i.e., “Bairro Prado”) ou não (i.e., “Prado”). As duas estratégias foram testadas e seus resultados comparados. Enquanto a primeira aumenta precisão, a segunda aumenta a revocação.

3.2 Enriquecimento da Localização

Essa etapa procura por nomes de logradouros e bairros relacionados aos locais identificados no passo anterior. Nesse caso, duas situações podem acontecer. Se o local previamente encontrado é um logradouro, buscamos um casamento aproximado com outras ruas com as quais ele possua um cruzamento em comum ou bairros por onde ele passa. Se o local é um bairro, tentamos encontrar um casamento aproximado com as ruas que passam pelo bairro. A motivação para essa estratégia vem do fato de o tweet ser limitado a 140 caracteres e os textos sobre trânsito serem, em geral, bastante concisos e citarem locais próximos.

O casamento utilizado nessa fase é um casamento *fuzzy* aproximado [Navarro 2001]. Enquanto a primeira técnica encontra nos tweets *substrings* idênticas aos nomes dos locais listados em GEODIC, o casamento *fuzzy* retorna um *score* que varia de 0 (*strings* completamente diferentes) a 100 (*strings* completamente idênticas), de acordo com a semelhança das *substrings* dentro do tweet com os nomes dos locais. Se o *score* for maior que determinado limiar, consideramos o casamento aceitável.

Note que, nessa fase, se duas ruas que se cruzam são citadas no mesmo tweet, a condição do trânsito é geocodificada correspondendo ao cruzamento. Se há mais de um cruzamento entre os mesmos logradouros, nenhuma geocodificação é feita. Note que, embora tenhamos outras informações sobre a latitude e longitude da rua, consideramos que só é seguro inferir sua localização precisa caso haja mais de uma referência.

Caso nenhum local tenha sido identificado, nada é feito. Optamos por essa estratégia para privilegiarmos a precisão na identificação de logradouros e bairros ao diminuirmos o escopo dos locais potencialmente citados. No futuro, trabalharemos formas de utilizar o casamento *fuzzy* aproximado de forma que a precisão não seja afetada.

4. BASE DE DADOS DO TWITTER

A base de dados criada para testar o método proposto foi extraída do Twitter utilizando um processo controlado. Foram coletados tweets de dez perfis cujo propósito principal é informar as condições do trânsito de Belo Horizonte e de outras cidades do país, tais como *@TransitoBH*, *@Transito98FM* e *@waytaxi*. Além disso, para fins de comparação, foram coletados os tweets postados pelo perfil *OficialBHTrans*, que é o perfil do órgão de trânsito de BH, e poderia ser considerado como um *ground truth*. No futuro, iremos contrastar essas informações com aquelas obtidas por coletas genéricas, mas que trazem outros problemas, tais como identificação de contexto, que teriam que ser considerados.

A coleta foi realizada por um período de três meses, entre abril e junho de 2012, e foram coletados 10.005 tweets dos perfis selecionados. Desses, 1.137 eram repostagens de mensagens de perfis dentro do conjunto. Repostagens entre perfis foram retiradas por conterem textos semelhantes, restando 8.868 tweets únicos.

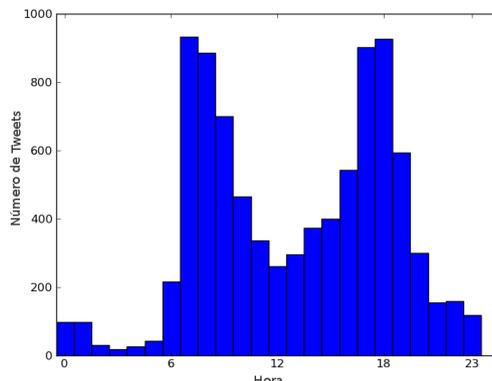


Fig. 1. Volume de tweets por hora

A Tabela I apresenta uma comparação entre os comportamentos do perfil oficial da BHTrans (*@OficialBHTrans*) e dos outros 10 perfis não oficiais que também reportam as condições de trânsito da cidade. São apresentados, para o perfil oficial e para os outros perfis, o número total de tweets, o número de dias no qual houve pelo menos um tweet, o número médio de tweets no tempo total analisado e o número médio de tweets entre os dias ativos (dias com pelo menos um tweet). No total, 91 dias foram analisados. É possível notar que, embora o perfil *@OficialBHTrans* possua um número médio de tweets por dia significativamente superior ao número médio de tweets produzidos por algum outro perfil individual, esse perfil atuou em somente 47 dos 91 dias analisados, enquanto os demais perfis cobriram 88 dos 91 dias analisados. Tal fato vai ao encontro do nosso argumento de que existe informação sobre o trânsito espalhada em diversos perfis, e que agrupá-la de forma coerente pode gerar informação de melhor qualidade.

Table I. Perfis Especializados em Trânsito

Perfis	# de tweets	# de dias ativos	Tweets por dia (média)	Tweets por dia ativo (média)
OficialBHTRANS	1.543	47	16,9	32,8
Outros perfis	8.462	88	7,7 (por perfil)	8,0 (por perfil)
Total	10.005	91	109,9	-

Na Figura 1, são apresentados os volumes de tweets do período analisado, agrupados por hora. É possível observar que o volume de tweets a respeito do trânsito é maior no período da manhã, por volta das 9h, e à noite, por volta das 18h. Esse é um indicativo de correlação entre os tweets e os problemas de trânsito do mundo real, já que tais horários são conhecidos pela maior circulação de carros nas cidades.

A Tabela II mostra a frequência com que os eventos pré-definidos aparecem nos tweets. Note que certos eventos e condições aparecem com grande frequência. Observamos, em outra análise, não descrita aqui por limitação de espaço, que existe grande correlação entre certos termos.

Table II. Eventos e condições de trânsito mais frequentes no conjunto de Tweets

Evento/Condição	Número de Tweets	Evento/Condição	Número de Tweets
lento	2000	parado	209
acidente	582	liberado	198
retido	499	congestionado	100
normal	373	manifestação	86
intenso	305	interditado	48
atenção	277	complicado	31

5. EXPERIMENTOS E RESULTADOS

O método criado, associado aos tweets coletados, deu origem ao Observatório do Trânsito¹. Para avaliar sua eficácia na localização de eventos, foram anotados manualmente 505 tweets escolhidos aleatoriamente. As anotações incluem os nomes dos logradouros e bairros cuja situação do trânsito foi citada no texto do tweet. Aplicamos então o algoritmo sobre esse conjunto de dados e calculamos a taxa de concordância entre as ruas e bairros encontrados e as anotadas em cada tweet.

Dois conjuntos de experimentos foram realizados. O primeiro envolve todos os 505 tweets rotulados (e aparece nas tabelas como “Todos os Tweets”), inclusive aqueles para os quais o nosso método não encontrou nenhum local. O segundo (“Tweets Classificados”) considera uma amostra de 3% dos tweets sobre os quais o método utilizado extraiu pelo menos um local, e a precisão é calculada sobre esse número, de modo a medir a precisão das informações que serão efetivamente mostradas para o usuário.

Para cada um dos cenários, a precisão foi calculada de duas formas. Para “Acerto Completo”, um acerto só é contabilizado se o conjunto de locais encontrados é idêntico ao conjunto anotado pelos anotadores humanos, enquanto para “Acerto Parcial”, o acerto é contabilizado caso o conjunto de locais anotado contenha o conjunto de locais encontrados pelo nosso método. Na última abordagem, estamos interessados em saber se o nosso método consegue extrair, mesmo que de forma incompleta, informações do tweet. As avaliações são feitas considerando a identificação exclusiva de logradouros, a identificação exclusiva de bairros e as duas simultaneamente.

A revocação do método é dada pela razão entre o número de tweets que citam pelo menos uma localidade e dos quais o nosso método conseguiu extrair pelo menos um local, mesmo que incorreto, e o número de tweets que citam algum local (de acordo com os anotadores humanos).

Table III. Precisão Total (PT), Precisão por Rua (PR) e Precisão por Bairro (PB) pra cada método

	Baseline			Todas as Definições			Sem Def. de Bairro		
	PT	PR	PB	PT	PR	PB	PT	PR	PB
Acerto Completo, Todos os Tweets	0,29	0,32	0,73	0,57	0,79	0,73	0,74	0,79	0,90
Acerto Parcial, Todos Tweets	0,98	0,99	0,99	0,88	0,90	0,98	0,82	0,90	0,92
Acerto Completo, Tweets Classificados	0,50	0,52	0,93	0,50	0,80	0,66	0,69	0,75	0,87
Acerto Parcial, Tweets Classificados	0,83	0,87	0,96	0,81	0,84	0,87	0,76	0,87	0,89
Revocação	0,12			0,86			0,95		

Na Tabela III, são apresentados os resultados de precisão total (logradouros e bairros) e por logradouro e bairro para cada método utilizado, além da revocação obtida. O método utilizado como *baseline* consiste em procurar por casamentos perfeitos nos textos com os nomes extraídos dos gazetteers, sem o processamento a partir do qual criamos GEODIC. Para o nosso método, utilizamos as duas estratégias citadas na Seção 3: utilizando a definição do logradouro/bairro (i.e., “Avenida Afonso Pena”) e não utilizando (i.e., “Afonso Pena”). Os resultados foram inferiores quando não utilizamos a definição dos logradouros e resolvemos não mostrá-la por falta de espaço. Na Tabela III, são mostrados os resultados do nosso método utilizando a definição de todos os logradouros e bairros (Todas as Definições) e utilizando a definição dos logradouros mas não dos bairros (Sem Def. de Bairro).

Os resultados do *baseline* para acertos parciais apresentaram alta precisão, especialmente ao considerarmos bairros e logradouros individualmente, indicando alta confiança na identificação de uma localidade a partir do casamento exato. A precisão no caso dos acertos completos, no entanto, é baixa, especialmente para os logradouros e considerando-se o conjunto completo de tweets analisados.

¹<http://inweb-dev.speed.dcc.ufmg.br/transitobh/>

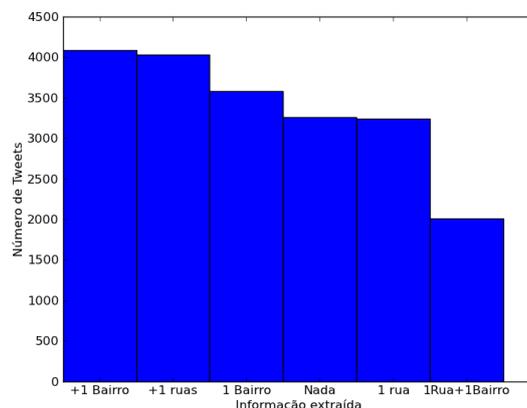


Fig. 2. Informações extraídas

Além disso, a revocação do *baseline* é muito pequena, indicando a inviabilidade da identificação das localidades simplesmente a partir do casamento exato.

O crescimento da revocação é notável ao aplicarmos o método proposto na Seção 3. Enquanto a porcentagem de tweets para os quais foi possível extrair alguma localização cresceu de 12% para 86%, a maior perda de precisão no acerto parcial, para toda a base analisada, foi de somente 10%. Além disso, o ganho de precisão no acerto completo para ruas foi de 47% para a base completa. Um impacto negativo, no entanto, foi a redução da precisão para o acerto completo dos bairros nos Tweets classificados. O principal motivo desse efeito foi a identificação errônea de alguns bairros pelo casamento *fuzzy*, que ocorreu devido a muitos deles terem o nome de palavras relativamente comuns (“centro”, “serra”, “aeroporto”). Finalmente, o relaxamento na exigência da definição para identificação dos bairros proporcionou um aumento de 9% na revocação sobre o método original, com um ganho de precisão de 27% no acerto completo para os bairros. Isso se deve ao fato de muitas referências a bairros serem feitas sem a definição “bairro” antecedendo o nome correspondente. A precisão para o acerto completo para ruas e bairros passou de 57% para 74%.

Utilizando o método sem a definição de bairros, analisamos todos os tweets da nossa base, visando quantificar os tipos de conteúdo dos tweets quanto aos tipos dos locais citados. Os resultados são mostrados na Figura 2, em que cada barra representa um “grupo” de tweets. O primeiro grupo apresenta um ou mais bairros, o segundo um ou mais logradouros, o terceiro exatamente um bairro, o quarto nenhuma localidade, o quinto exatamente um logradouro e o sexto exatamente um bairro e um logradouro. Note o significativo número de tweets que citam mais do que apenas uma rua, o que favorece a identificação de cruzamentos ou trechos e a consequente geolocalização da mensagem.

Na Tabela IV, são apresentadas as precisões para acertos totais e parciais de cada um dos grupos (barras) que aparecem no gráfico 2. Foram anotados manualmente 3% de cada um dos grupos. Note que o método proposto tem maior acerto nos grupos em que são encontrados um logradouro e um bairro. Quando não consegue extrair nada, o método esteve certo em 61% dos casos anotados.

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

É possível observar, na Tabela IV, que pelo menos uma rua pôde ser identificada em 81% dos tweets analisados, e pelo menos um bairro em 72%. Pode-se notar, também, o ganho de precisão quando mais de uma localidade é identificada, pelo menos para o acerto parcial. Os melhores resultados foram aqueles em que um trecho de uma rua foi identificado, ou seja, em que foram encontrados uma rua e um bairro seccionado por ela.

6. CONCLUSÃO E TRABALHOS FUTUROS

Os resultados encontrados apontam para um caminho promissor na identificação de eventos e condições no trânsito a partir de tweets, e ainda existem várias maneiras de melhorá-los. Neste primeiro trabalho, não procuramos identificar casos em que dois ou mais tweets se referem ao mesmo evento ou condição do trânsito, mas de formas diferentes. Um trabalho de caracterização de eventos e condições de trânsito deve ser feito visando agrupar eventos e condições semelhantes, além de permitir a identificação de causalidades.

Os eventos e condições de trânsito foram analisados somente de forma pontual, sem levar em conta os padrões temporais que podem ser inferidos a partir dos dados coletados. No futuro, pretendemos estudar e implementar técnicas que possibilitem a identificação de padrões e *outliers* nos dados.

Embora seja possível geocodificar várias ocorrências e condições de trânsito, muitas ficam desconhecidas por não sabermos exatamente em qual posição de determinada rua ou bairro um dado evento aconteceu. Uma possível forma de aumentar o número de condições de trânsito geocodificadas é utilizar referências a lugares conhecidos feitas nos tweets, tais como shoppings, bares, praças etc.

Por fim, um estudo sobre as diferenças de informação obtidas considerando tweets postados por usuários em geral e aqueles postados por usuários especializados em trânsito deve ser feito. O quanto ganhamos tornando a coleta mais complexa? A resposta para essa pergunta virá em trabalhos futuros.

REFERENCES

- AMITAY, E., HAR'EL, N., SIVAN, R., AND SOFFER, A. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '04. ACM, New York, NY, USA, pp. 273–280, 2004.
- CARDOSO, N., SILVA, M. J., AND SANTOS, D. Handling implicit geographic evidence for geographic ir. In *Proceedings of the 17th ACM conference on Information and knowledge management*. CIKM '08. ACM, New York, NY, USA, pp. 1383–1384, 2008.
- CHENG, Z., CAVERLEE, J., AND LEE, K. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. CIKM '10. ACM, New York, NY, USA, pp. 759–768, 2010.
- DANIEL, B. *Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena*. Number vol. 1. Igi Global, 2010.
- DELBONI, T. M., BORGES, K. A. V., AND LAENDER, A. H. F. Geographic web search based on positioning expressions. In *Proceedings of the 2005 workshop on Geographic information retrieval*. GIR '05. ACM, New York, NY, USA, pp. 61–64, 2005.
- GOLDBERG, D., WILSON, J., AND KNOBLOCK, C. From text to geographic coordinates: the current state of geocoding. *URISA Journal* 19 (1): 33–47, 2007.
- MACHADO, I. M., DE ALENCAR, R. O., DE OLIVEIRA CAMPOS JUNIOR, R., AND DAVIS, C. A. An ontological gazetteer and its application for place name disambiguation in text. *J. Braz. Comp. Soc.* 17 (4): 267–279, 2011.
- NAVARRO, G. A guided tour to approximate string matching. *ACM Comput. Surv.* 33 (1): 31–88, Mar., 2001.
- PAPPA, G., MEIRA JR., W., ALMEIDA, V., VELOSO, A., AND DA SILVA, A. Observatório da web: Uma plataforma de monitoração, síntese e visualização de eventos massivos em tempo real. *Anais do XXXVII Seminário Integrado de Software and Hardware*, 2010.
- TWAROCH, F. A., SMART, P. D., AND JONES, C. B. Mining the web to detect place names. In *Proceedings of the 2nd international workshop on Geographic information retrieval*. GIR '08. ACM, New York, NY, USA, pp. 43–44, 2008.