Universidade Estadual do Norte Fluminense Darcy Ribeiro - UENF

Augusto Zangrandi

Detecção de Eventos com dados do Twitter

Campos dos Goytacazes/RJ

2014

Augusto Zangrandi

Detecção de Eventos com dados do Twitter

Monografia apresentada ao Curso de Graduação em Ciência da Computação da Universidade Estadual do Norte Fluminense Darcy Ribeiro como requisito para obtenção do título de Bacharel em Ciência da Computação, sob orientação do Profº. Luis Antônio Rivera Escriba.

Orientador: Luis Antônio Rivera Escriba.

Universidade Estadual do Norte Fluminense Darcy Ribeiro - UENF

Campos dos Goytacazes/RJ

2014

Augusto Zangrandi

Monografia apresentada junto ao Curso de Ciência da Computação, da Universidade Estadual do Norte Fluminense Darcy Ribeiro – Campos / RJ, como requisito para obtenção do título de Bacharel em Ciência da Computação. Orientador: Prof. Dr. Luis Antônio Rivera Escriba.

Aprovado em 29/09/2014.

COMISSÃO EXAMINADORA

Prof. Dr. Luis Antônio Rivera Escriba
Orientador - Universidade Estadual do Norte Fluminense Darcy Ribeiro

---

Prof. Dr. Ausberto S. Castro Vera
Universidade Estadual do Norte Fluminense Darcy Ribeiro

Prof. Dr. Fermin Alfredo Tang Montané
Universidade Estadual do Norte Fluminense Darcy Ribeiro

"Sua tarefa é descobrir o seu trabalho e, então, com todo o coração, dedicar-se a ele."

Buda

AGRADECIMENTOS

Primeiramente, gostaria de agradecer à minha família, que está comigo desde sempre, me apoiando em todas as fases, me dando forças para seguir e sendo a base com que cresci e aprendi a lidar com a vida.

Gostaria de agradecer à minha querida namorada, Ana Liz, que em um momento em que estava desanimado a escrever, oportunamente entrou novamente na minha vida, me dando suporte e energia, que investi nesse trabalho que tanto gostei do resultado.

Agradeço, também, ao meu orientador, que sempre se mostrou solícito ajudando em todas as etapas do trabalho, corrigindo erros, dando ideias e ensinando conceitos, o que ajudou para a confecção de um trabalho bom e coerente.

À Rodrigo Manhães, grande professor da UENF, que foge do padrão universitário rígido e apresenta aos alunos tecnologias recentes e animadoras, o que é de grande ajuda para seguirmos nesse caminho.

Ao meu primeiro sócio, Rafael Carvalho, que me ensinou a ter seriedade nas coisas que faço, e que o mundo está repleto de pessoas que se movimentam seguindo seu sonho de impactar e melhorar a vida das pessoas de maneira empreendedora.

Aos professores presentes na banca, que deram dicas valiosas para o trabalho.

À todos os amigos que fizeram parte da minha história na UENF, que dividiram comigo risadas e preocupações, reduzindo consideravelmente o esforço de passar por mais essa fase.

Obrigado.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Detecção de eventos consiste no processo de identificação de padrões de mudança relevantes em um sistema. O modelo implementa a detecção de eventos com dados das publicações criadas por usuários no serviço de rede social Twitter, serviço que tem recebido grande atenção acadêmica por sua alta popularidade e característica de troca de informações em tempo real. O modelo utiliza, como exemplo de evento, manifestações que ocorrem no território brasileiro durante o mês de agosto de 2014. Primeiramente, são obtidas as publicações criadas no serviço através da sua interface para desenvolvedores. Posteriormente, as publicações são convertidas para o seu modelo de espaço vetorial, para que seus vetores de características sejam extraídos e que os mesmos sejam utilizados para a fase de classificação com máquina de vetores de suporte. A classificação consiste na divisão das publicações obtidas em duas classes: positivas e negativas. Sendo positivas as que dizem respeito à uma manifestação real, com local e horário, e negativas as que somente mencionam manifestação mas não dizem respeito à um acontecimento real. Após classificadas, as publicações são agrupadas e exibidas em gráficos no formato de série-temporal e mapas de marcadores, aonde é possível detectar, através da análise visual, tanto o horário dos eventos, através de picos que surgem nos gráficos, quanto a sua localização, através do seu aglomeramento em determinadas regiões do mapa.

Abstract

Event detection consists in the identification of relevant patterns of change in a system. The model implements event detection with data from the publications created by the users in the social networking service Twitter, one service that has received great scholarly attention due to its high popularity and characteristic of real time information exchange. The model uses, as an example of event, manifestations that occur in Brazilian territory, during the month of August 2014. First, the model obtain the publications through the Twitter interface for developers. Then, the publications are converted to their vector space model, for extracting their feature vectors so that they can be classified by the support vector machine. The classification consists on the division of the publications into two groups: positives and negatives. The positives means that the publication referes to a real manifestation, with time and location, and the negatives means that the publication only mentions the word, but not a real occurrence. Once classified, the publications are grouped and displayed in time-series graphs and marker maps, where it's possible to detect, by visual analyse, both the time of the event, through peaks that appear in the graph, and their location, through aglomeration in certain regions of the map.

Referências

24765:2010(E), I. *Systems and software engineering — Vocabulary*. Geneva, Switzerland, 2010.

ABOU-ZLEIKHA, M. et al. Non-linguistic vocal event detection using online random forest. In: *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*. [S.l.: s.n.], 2014. p. 1326–1330.

ADEDOYIN-OLOWE, M.; GABER, M. M.; STAHL, F. A survey of data mining techniques for social media analysis. *CoRR*, abs/1312.4617, 2013.

AGGARWAL, C. C.; ZHAI, C. A survey of text classification algorithms. Springer US, p. 162 – 222, 2012.

AIELLO, L. M. et al. Sensing trending topics in twitter. IEEE, 2013.

ALI, S.; SMITH, K. A. Kernal width selection for svm classification: A meta-learning approach. *International Journal of Data Warehousing and Mining*, IGI Publishing, v. 1, 2005.

ALLEN, J. F.; FERGUSON, G. Action and events in temporal logic. *Journal of Logic and Computation*, Oxford Journals, v. 4, p. 531 – 579, 1994.

BECKER, H.; NAAMAN, M.; GRAVANO, L. Beyond trending topics: Real-world event identification on twitter. *Fifth International AAAI Conference on Weblogs and Social Media*, Association for the Advancement of Artificial Intelligence, 2011.

BEGOLI, E.; HOREY, J. Design principles for effective knowledge discovery from big data. In: *Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on*. [S.l.: s.n.], 2012. p. 215–218.

BENEVENUTO, F. et al. Detecting spammers on twitter. CEAS, 2010.

CHAPELLE, O. et al. Choosing multiple parameters for support vector machines. *Machine Learning*, Kluwer Academic Publishers, v. 46, n. 1-3, p. 131–159, 2002. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A%3A1012450327387>.

COOK, K. A.; THOMAS, J. J. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. [S.l.]: Pacific Northwest National Laboratory, 2005.

COX, J.; PLALE, B. Improving automatic weather observations with the public twitter stream. *IU School of Informatics and Computing*, Indiana University, 2011.

DONG, X. et al. Multiscale event detection in social media. *CoRR*, abs/1404.7048, 2014.

FIENBERG, S. E.; SHMUELI, G. Statistical issues and challenges associated with rapid detection of bio-terrorist attacks. John Wiley and Sons, 2005.

GAO, D. et al. Sequential summarization: A full view of twitter trending topics. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, v. 22, n. 2, p. 293–302, Fevereiro 2014. ISSN 2329-9290.

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. *CS224N Project Report*, Stanford University, 2009.

GRIGORIK, I. *Support Vector Machines (SVM) in Ruby*. 2008. Disponível em: <https://www.igvita.com/2008/01/07/support-vector-machines-svm-in-ruby/>. Acesso em: 30/06/2014.

GUPCHUP, J. et al. Model-based event detection in wireless sensor networks. *CoRR*, abs/0901.3923, 2009.

HONG, J.; LIU, C.-C.; GOVINDARASU, M. Detection of cyber intrusions using network-based multicast messages for substation automation. In: *Innovative Smart Grid Technologies Conference (ISGT), 2014 IEEE PES*. [S.l.: s.n.], 2014. p. 1–5.

IHLER, A.; HUTCHINS, J.; SMYTH, P. Adaptive event detection with time-varying poisson processes. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2006. (KDD '06), p. 207–216. ISBN 1-59593-339-5.

IWAKURA, T.; OKAMOTO, S. A fast boosting-based learner for feature-rich tagging and chunking. *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, p. 17 – 24, 2008.

JANSEN, B. J. et al. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, ASIS and T, 2009.

JAVA, A. et al. Why we twitter: Understanding microblogging usage and communities. *9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ACM, 2007.

JIANG, W.; BLUMBERG, A. F.; BUTTREY, S. E. Event detection challenges, methods, and applications in natural and artificial systems. *14th ICCRTS: "C2 and Agility"*, Lockheed Martin MS2, 2009.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, Springer Berlin Heidelberg, p. 137 – 142, 1998.

KELLY, R. *Twitter Study Reveals Interesting Results About Usage – 40is Pointless Babble*. 2009. Disponível em: <http://www.pearanalytics.com/blog/2009/twitter-study-reveals-interesting-results-40-percent-pointless-babble/>. Acesso em: 25/03/2014.

KIM, E. *Everything You Wanted to Know about the Kernel Trick*. Setembro 2013. Disponível em: <http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html>. Acesso em: 05/08/2014.

LIU, Z.; JIANG, B.; HEER, J. immens: Real-time visual querying of big data. *Computer Graphics Forum*, Blackwell Publishing Ltd., v. 32, n. 3pt4, p. 421–430, 2013.

MAI, E.; HRANAC, R. Twitter interactions as a data source for transportation incidents. *TRB 2013 Annual Meeting*, TRB, 2012.

MATUSZKA, T.; VINCELLER, Z.; LAKI, S. On a keyword-lifecycle model for real-time event detection in social network data. In: *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*. [S.l.: s.n.], 2013. p. 453–458.

MEGRI, A. C.; NAQA, I. E. Prediction of the thermal comfort indices using improved support vector machine classifiers and nonlinear kernel functions. *Indoor and Built Environment*, 2014. Disponível em: <http://ibe.sagepub.com/content/early/2014/07-/11/1420326X14539693.abstract>.

OZER, S.; CHEN, C. H.; CIRPAN, H. A. A set of new chebyshev kernel functions for support vector machine pattern classification. *Pattern Recognition*, v. 44, n. 7, p. 1435 – 1447, 2011. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science-/article/pii/S0031320311000021>.

PAK, A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. LREC, 2010.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. 2010.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. Communications of the ACM, v. 18, p. 613 – 620, 1975.

SANTOS, L. M. et al. Twitter, análise de sentimento e desenvolvimento de produtos: Quanto os usuários estão expressando suas opiniões? Prisma.com, n. 13, 2010.

SOULE, A.; SALAMANTIAN, K.; TAFT, N. Combining filtering and statistical methods for anomaly detection. *Internet Measurement Conference*, USENIX Association, p. 331 – 344, 2005.

STANKOVIC, M.; ROWE, M.; LAUBLET, P. Mapping tweets to conference talks: A goldmine for semantics. *Social Data on the Web Workshop at the International Semantic Web Conference*, Shanghai, China, 2010.

TAKAHASHI, T.; ABE, S.; IGATA, N. Can twitter be an alternative of real-world sensors? *Human-Computer Interaction, Part III*, Springer-Verlag Berlin Heidelberg, p. 240 – 249, 2011.

TUMASJAN, A. et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. Association for the Advancement of Artificial Intelligence, Munich, Germany, 2010.

TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, AI Access Foundation and National Research Council Canada, p. 141–188, 2010.

WANG, X. et al. Real time event detection in twitter. *14th International Conference, WAIM*, Springer-Verlag Berlin Heidelberg, p. 502 – 513, 2013.

WASSON, C. S. *System Analysis, Design, and Development*. [S.l.]: John Wiley and Sons, Inc., 2006.

WENG, J.; LEE, B.-S. Event detection in twitter. *Fifth International AAAI Conference on Weblogs and Social Media*, Association for the Advancement of Artificial Intelligence, 2011.

YANG, Y.; PIERCE, T.; CARBONELL, J. A study on retrospective and on-line event. *ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, p. 28 – 36, 1998.