Social Media Analytics: Tracking, Modeling and Predicting the Flow of Information through Networks

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

Online social media represent a fundamental shift of how information is being produced, transferred and consumed. User generated content in the form of blog posts, comments, and tweets establishes a connection between the producers and the consumers of information. Tracking the pulse of the social media outlets, enables companies to gain feedback and insight in how to improve and market products better. For consumers, the abundance of information and opinions from diverse sources helps them tap into the wisdom of crowds, to aid in making more informed decisions.

The present tutorial investigates techniques for social media modeling, analytics and optimization. First we present methods for collecting large scale social media data and then discuss techniques for coping with and correcting for the effects arising from missing and incomplete data. We proceed by discussing methods for extracting and tracking information as it spreads among the users. Then we examine methods for extracting temporal patterns by which information popularity grows and fades over time. We show how to quantify and maximize the influence of media outlets on the popularity and attention given to particular piece of content, and how to build predictive models of information diffusion and adoption. As the information often spreads through implicit social and information networks we present methods for inferring networks of influence and diffusion. Last, we discuss methods for tracking the flow of sentiment through networks and emergence of polarization.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database applications—*Data mining*

General Terms: Algorithms; Experimentation.

Keywords: Social media analytics, Social networks, Information diffusion, Information cascades, Influence maximization

1. INTRODUCTION

The emergence of the *Web* and *Online Social Media* represents a fundamental shift as it has added important new dimensions to the production and dissemination of news and information. Social Media allows for social interaction, using highly accessible and scalable publishing techniques. Users can generate content, access information, and potentially reach large audiences. Social Media also replaces the traditional one-way mass-media to consumer communication channel with an interactive dialogue, which allows for the creation and exchange of user-generated content. This opens rich venues for mining and analyzing social media data. Companies analyze social media data to perform analytics, sentiment analysis or find influencers. Users browse information and opinions from diverse sources that helps them tap into the wisdom of crowds, to aid in making more informed decisions. However, this also opens a question of how do we overcome the information overload and provide a rich and coherent user experience?

Social Media provides a connection between our social networks, personal information channels and the mass media. Social Media data in the form of user-generated content on blogs, microblogs like Twitter, discussion forums, product review and multimedia sharing websites presents many new opportunities and challenges to both producers and consumers of information. Although there is a vast quantity of data available, the consequent challenge is to be able to analyze the large volumes of user-generated content and often implicit links between users, in order to gain meaningful insights.

The goal of this tutorial is to address methods, metrics and predictive tasks, as well as actionable explanatory analysis of social media data. The tutorial will survey recent methods and algorithms for large scale social media analytics and address the following questions:

- How do we collect massive amounts of social media data and what techniques can be used for correcting for the effects and biases arising from incomplete and missing data?
- What methods can be used to extract and track the flow of interesting pieces of information that spread and diffuse among the users? How can we identify the subset of content that is discussing not only a specific entity, but higher level concepts?
- Having identified the subset of relevant content, how do we identify the most authoritative or influential authors? How do we quantify the influence of users on the adoption and spread of different topics? How do we maximize the overall influence?
- How do we tease apart emerging topics of discussion from the constant chatter in the blogosphere and other social media? How do we extract and model the temporal patterns by which information grows and fades over time?
- How do we predict popularity of memes and other pieces of information that spread through the social media networks?
- The information spreads via implicit networks. How do we identify and infer such networks of influence and diffusion? How do we discover implicit links between users?

Tutorial video, slides and all other materials are available at `http://snap.stanford.edu/proj/socmedia-www/`

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

- How does sentiment flow through networks and how does polarization occur?
- How do we overcome the information overload and provide users with rich and coherent experience?
- How to deal with unreliable and often conflicting information? What notions of trust are appropriate?

Tutorial presentation, slides and other materials are available at `http://snap.stanford.edu/proj/socmedia-www/`

2. REFERENCES

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

[2] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 665–674, New York, NY, USA, 2008. ACM.

[3] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In *Web Intelligence*, pages 207–214, 2005.

[4] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004.

[5] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high quality content in social media, with an application to community-based question answering. In *WSDM '08: ACM International Conference on Web Search and Data Minig*, pages 183–194, 2008.

[6] M. De Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher. How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media? In *ICWSM '10: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 2010.

[7] D. Fisher, M. Smith, and H. T. Welser. You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In *HICSS '06: Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, volume 3, page 59b, 2006.

[8] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 211–220, New York, NY, USA, 2009. ACM.

[9] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *KDD '05: Proceeding of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 419–428, 2005.

[10] M. Goetz, J. Leskovec, M. Mcglohon, and C. Faloutsos. Modeling blog dynamics. In *International Conference on Weblogs and Social Media*, May 2009.

[11] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010.

[12] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD '05: Proceeding of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87, 2005.

[13] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501, 2004.

[14] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 403–412, 2004.

[15] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[16] D. Kempe, J. M. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

[17] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 568–576, 2003.

[18] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *WWW'10: Proceedings of the 19th International World Wide Web Conference*, April 2010.

[19] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506, 2009.

[20] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, 2010.

[21] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD '07: Proceeding of the 13th ACM SIGKDD international conference on Knowledge discovery in data mining*, 2007.

[22] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SDM '07: Proceedings of the SIAM Conference on Data Mining*, 2007.

[23] J. Leskovec, A. Singh, and J. M. Kleinberg. Patterns of influence in a recommendation network. In *PAKDD '06: Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 380–389, 2006.

[24] S. Myers and J. Leskovec. On the convexity of latent social network inference. In *NIPS '10: Advances in Neural Information Processing Systems*, 2010.

[25] S. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina. Correcting for missing data in information cascades. In *WSDM '11: ACM International Conference on Web Search and Data Minig*, 2011.

[26] J. Yang and Leskovec. Modeling information diffusion in implicit networks. In *ICDM '10: IEEE International Conference On Data Mining*, 2010.

[27] J. Yang and Leskovec. Patterns of temporal variation in online media. In *WSDM '11: ACM International Conference on Web Search and Data Minig*, 2011.