Living Analytics Methods for the Web Observatory

Ernesto Diaz-Aviles
L3S Research Center
University of Hannover, Germany
diaz@L3S.de

ABSTRACT

The collective effervescence of social media production has been enjoying a great deal of success in recent years. The hundred of millions of users who are actively participating in the *Social Web* are exposed to ever-growing amounts of sites, relationships, and information.

In this paper, we report part of the efforts towards the realization of a Web Observatory at the L3S Research Center (`www.L3S.de`). In particular, we present our approach based on *Living Analytics* methods, whose main goal is to capture people interactions in real-time and to analyze multidimensional relationships, metadata, and other data becoming ubiquitous in the social web, in order to discover the most relevant and attractive information to support observation, understanding and analysis of the Web. We center the discussion on two areas: (i) *Recommender Systems for Big Fast Data* and (ii) *Collective Intelligence*, both key components towards an *analytics toolbox* for our Web Observatory.

Categories and Subject Descriptors

H3.3 [**Information Search and Retrieval**]: Information filtering; K.4 [**Computer and Society**]

General Terms

Design, Human Factors, Algorithms

Keywords

Web Observatory; Big Fast Data; Recommender Systems; Collective Intelligence

1. INTRODUCTION

The Web of people is highly dynamic and the life experiences between our on-line and "real-world" interactions are increasingly interconnected. The massive amounts of information from people's daily living interactions require new and innovative analytic models to observe, understand and predict the dynamics of the actors and components of the Web.

The objective of such *Living Analytics* models is to help people live better, for example, by assisting them with the overwhelming amount of choices they face as they consume goods, services, social media and leisure time [10]. *Living Analytics* aims at capturing people interactions in real-time, and at analyzing and processing them in order to produce valuable output that is fed back to the Web of people for the benefit of its members.

Consider, for example, a system envisioned to support its users to conduct reliable assessments of dynamics topics on the Web, such as: views on political developments, economic events and crises, as well as pandemics or natural catastrophes. Moreover, Social Media technologies have given rise to citizen journalism, so the public at large may actively contribute to generating new content in these areas. The goal of such system is to go beyond individual resources to an aggregated overview, by automatically collecting the relevant sources, extracting the required data, aggregating the results, and finally enabling in-depth investigation with tools designed to support visual analysis. The need for overviews of high volume and user-generated content is crucial for many stakeholders, for example, journalists, opinion analysts, social scientists, public safety institutions, product designers and marketers, and well as the general public.

On-line applications that tackle the information deluge problem exist, at least for selected topics. For example, detecting the global trends based on the volume of Twitter[1] messages (*tweets*) or generating visualizations from publicly available statistics, e.g., Google Public Data Explorer[2]. While these tools are very useful, the respective services tend to be limited to a very small fraction of interesting information. In addition, users are presented with "one-size-fits-all" solutions that do not necessarily reflect their individual interests.

Having this scenario in mind, several interesting research questions arise in the scope of *Living Analytics*. For example:

- **Understanding and Predicting Behavior in Real-Time Context.** What theoretical, methodological, and empirical extensions are needed to observe and analyze the behavior of users and groups in the network in near or real-time, as it is occurring? and how is it possible to follow the evolution of network behavior over extended periods of time?

- **Collective Intelligence**. What content do people create and share? How can the collective behavior of people be harnessed to solve complex tasks? How does

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

[1] `twitter.com` .

[2] `www.google.com/publicdata/`.

collective intelligence evolve over time? How can the trends observed in social media be employed to improve discovery performance?

- **Real-Time Modeling and Experimentation**. How can the ability to update predictive models, as well as, execute and interpret experiments in these real-time networked settings of users, and their group interactions, be realized and improved?

In this paper, we discuss our approach to some of these challenges and present a set of tools to capture people's interactions from a highly dynamic stream of data, which help to understand user preferences and predict user actions. In particular, we center our discussion around recommender systems for big fast data and collective intelligence.

2. TOWARDS OUR WEB OBSERVATORY

Undoubtedly, analytics methods for the social web are becoming increasingly necessary to make sense out of the huge amount of user generated content and reduce the complexity for human user understanding. Current research greatly benefits from cross-disciplines, including machine learning, recommender systems, and computational social science. The integration of interdisciplinary evidence also represents an important ingredient of our efforts towards the Web Observatory at the L3S Research Center.

Our approach based on *Living Analytics* combines the key technologies of statistical machine learning, large scale data mining, and computational tools for the analysis of dynamic social networks with analytics focused on user behavior and social media. For example, matrix factorization in collaborative filtering systems [9], social stream mining [14], social tagging systems [8], epidemic intelligence [6], computational social science [7] and sentiment analysis [11].

Our current main objective is to provide a set of tools to capture people's interactions from a highly dynamic stream data, automatically annotate resources on the Web, and to understand and predict user actions and preferences. The objective of such *Living Analytics* models is not only to observe and analyze the web, but also to assist people with the overwhelming amount of choices they face as they consume goods, services, social media and leisure time [13].

Recommender Systems for Big Fast Data

The main problem is not the actual access to the content (e.g., *Twitter* or *YouTube*), rather the problem is to transform this huge mass of data into useful insight. Effective recommendation systems and personalized ranking are key elements in this scenario. In particular, instead of creating one global ranking, we tackle the problem of personalized ranking, that is, the rankings should reflect the individual taste of the users.

Based on these ideas, we have introduced methods that in the presence of highly dynamic social media streams, create in real-time user-specific rankings based on individual preferences that are inferred from users' past system interactions. Our online ranking approaches for collaborative filtering are based on matrix factorization. At their core stochastic gradient descent is used for optimization, which makes the algorithms easy to implement and efficiently scalable to large-scale datasets.

We have demonstrated the usefulness of our methods for the task of recommending personalized topics to users. Having Twitter as test bed, we showed that our online approaches largely outperform highly competitive state-of-the-art matrix factorization techniques for collaborative filtering, not only in terms of recommendation quality, but also in terms of time and space savings [1, 2].

Collective Intelligence

In the scope of Collective Intelligence [12], we explored and demonstrated the potential of monitor social media streams (e.g., Twitter) for early warning of disease outbreaks. Furthermore, for outbreak analysis and control, many studies have been made for systems that return documents in response to a query. Little effort has been devoted to exploiting learning to rank in a personalized setting, specially in the domain of epidemic intelligence. In [5, 4], we presented an innovative personalized ranking approach that offers decision makers the most relevant and attractive tweets for risk assessment, by exploiting latent topics and social hashtagging behavior in Twitter.

For Computational Social Science, we conducted an empirical study that shows how the real-time nature of social media streams, in particular, Twitter, can be leveraged to take the pulse of political emotions in emerging regions of the world, namely: Latin America. We performed a sentiment analysis of tweets and brief blog posts over a period of six months. This work presents, not only the extracted emotions and polarity, but also goes a step forward and quantifies which combination of emotions explains better the public's opinion [3].

The aforementioned studies constitute key components towards an *analytics toolbox* for our Web Observatory. The lessons learned through these works also open exciting directions for future research, which are outlined in the next section.

3. FUTURE DIRECTIONS

There are several potential future directions we want to explore, particularly related to information filtering in the presence of highly dynamic data.

- **Better Learning Algorithms for Online Collaborative Filtering.** The success of collaborative filtering heavily relies upon the ability to translate the observed behavior to a meaningful cost function. We strongly believe that The Top-N recommendation task needs to be treated as a ranking problem as discussed in [1, 2]. The exploration of directly optimizing information retrieval metrics for personalized ranking has started and may significantly improve recommendation performance.

- **Prediction of Individual and Collective Behavior in Real-Time Context.** Modeling complex nonlinear dynamics and high-dimensional data, such as social media streams, is an active area of research in machine learning and recommender systems. Many of the existing models, such as matrix factorization and neighborhood based algorithms have been widely used in practice. However, these models are limited in the types of structure they can model. What other methods could potentially capture nonlinear dynamics and also make multimodal predictions handling missing inputs?

- **Real-Time Experimentation.** How to conduct experimental evaluations at large scale in real-time networked settings involving users and their group interactions? A/B testing is a common practice in the industry to evaluate new project features and to support decision making processes, but such evaluations are expensive and time consuming. The exploration of new approaches that align long-term goals with the objective functions optimized by the learning models is an interesting research direction.

- **From Datasets to Modelsets.** Given the large scale of the data to be observed, it is challenging to efficiently capture it all, and share it across Web Observatories. Furthermore, the intellectual property rights and terms of service are issues that cannot be ignored. We believe that besides the efforts of sharing the data, we can also focus on building the mechanisms for sharing the models built on the data observed over time. One major advantage of these *modelsets* is that they are usually of a much smaller footprint than the datasets used to build them. The modelsets can then be shared across the Web Observatories, and applied to analyze the current stream or to predict new events. Such models can also be designed to be updated online, as discussed in Section 2.

4. CONCLUSION

Our research on collaborative filtering in social media streams and collective intelligence, expands focus into new directions, namely that of the emerging science of the Web.

The methods presented in this paper constitute a set of tools to understand and analyze the social web following a *Living Analytics* approach towards our Web Observatory at the L3S Research Center. Note that the analytics *toolbox* presented is by no means complete, giving interesting opportunities for future research.

We have outlined several potential research directions. However, research on *Living Analytics* as unified field is very new, and there are many broad open questions to consider as outlined in the Introduction. We believe that answering many of those questions will allow us to build more intelligent Web Observatories for the benefit of society.

5. REFERENCES

[1] E. Diaz-Aviles, L. Drumond, Z. Gantner, L. Schmidt-Thieme, and W. Nejdl. What Is Happening Right Now ... That Interests Me? Online Topic Discovery And Recommendation In Twitter. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, 2012.

[2] E. Diaz-Aviles, L. Drumond, L. Schmidt-Thieme, and W. Nejdl. Real-time Top-N Recommendation In Social Streams. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 59–66, New York, NY, USA, 2012. ACM.

[3] E. Diaz-Aviles, C. Orellana-Rodriguez, and W. Nejdl. Taking The Pulse of Political Emotions In Latin America Based on Social Web Streams. In *LA-WEB '12: Proceedings of the 2012 Latin American Web Conference*. IEEE Computer Society, 2012.

[4] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Epidemic Intelligence For The Crowd By The Crowd. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*, 2012.

[5] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Towards Personalized Learning To Rank For Epidemic Intelligence Based on Social Media Streams. In *Procedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 495–496, New York, NY, USA, 2012. ACM.

[6] M. Fisichella, A. Stewart, A. Cuzzocrea, and K. Denecke. Detecting Health Events on The Social Web To Enable Epidemic Intelligence. In *String Processing and Information Retrieval*, volume 7024 of *Lecture Notes in Computer Science. SPIRE'11*, pages 87–103. Springer Berlin Heidelberg, 2011.

[7] J. Giles. Computational Social Science: Making The Links. *Nature*, 488(7412):448–450, Aug. 2012.

[8] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag Recommendations In Social Bookmarking Systems. *AI Communications*, pages 231–247, 2008.

[9] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques For Recommender Systems. *Computer*, August 2009.

[10] LARC. Living Analytics Research Center. http://www.larc.smu.edu.sg/, 2013.

[11] B. Liu. *Sentiment Analysis And Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[12] MIT. Handbook of Collective Intelligence (wiki), 2013.

[13] B. Schwartz. *The Paradox of Choice : Why More Is Less*. Harper Perennial, 2004.

[14] A. Zubiaga. Real-time Analysis And Mining of Social Streams. In *International AAAI Conference on Weblogs and Social Media*, ICWSM'12, 2012.