Structured Analytics in Social Media

Mahashweta Das
HP Labs Palo Alto
California, USA
mahashweta.das@hp.com

Gautam Das
University of Texas at Arlington
Texas, USA
gdas@uta.edu

ABSTRACT

The rise of social media has turned the Web into an online community where people connect, communicate, and collaborate with each other. Structured analytics in social media is the process of discovering the structure of the relationships emerging from this social media use. It focuses on identifying the users involved, the activities they undertake, the actions they perform, and the items (e.g., movies, restaurants, blogs, etc.) they create and interact with. There are two key challenges facing these tasks: how to organize and model social media content, which is often unstructured in its raw form, in order to employ structured analytics on it; and how to employ analytics algorithms to capture both explicit link-based relationships and implicit behavior-based relationships. In this tutorial, we systemize and summarize the research so far in analyzing social interactions between users and items in the Web from data mining and database perspectives. We start with a general overview of the topic, including discourse to various exciting and practical applications. Then, we discuss the state-of-art for modeling the data, formalizing the mining task, developing the algorithmic solutions, and evaluating on real datasets. We also emphasize open problems and challenges for future research in the area of structured analytics and social media.

1. INTRODUCTION

Motivation: The rise of social media has turned the Web into an online community where people connect, communicate, and collaborate with each other. This has led to the generation of huge amount of data, either in the form of user-generated content (e.g., Facebook posts, Yelp reviews, etc.), or in the form of user-item interactions (e.g., user X reviews restaurant A in Yelp) and user-user interactions (e.g., user X likes user Y's Facebook post). Recent years have witnessed a confluence of techniques from diverse fields like database, data mining, information retrieval, and machine learning to tap into the rich resource of social data and derive meaningful actionable insights. Structured analytics is one such powerful tool that focuses on discovering the structure of the relationships emerging from this social media use. It focuses on identifying the users involved, the activities they undertake, the actions they perform, and the items they create and interact with.

Broad Summary: In this tutorial, we present a general overview of the topic that includes discussion on taxonomic classification, technical challenges, and the various exciting applications enabled by it. Social media data is essentially unstructured, and thus is difficult to analyze. Thus, a core challenge in this form of analytics is figuring out how to organize and model the data in order to employ structured analytics on it. Given the scale of data available today and the wide variety of social analytics tasks, another research challenge is algorithm design. We systemize the technical development in this topic by identifying two broad objectives: analytic approaches to mine implicit behavior-based relationships that arises due to shared interests exemplified by users' online activities, and analytic approaches to mine explicit link-based relationships that arises due to direct connections between users. For each of the objectives, we review problem formalization, data model, algorithmic solutions, and experimental evaluation. Specifically, we cover four distinct structured analytics approaches in the context of social data mining. We also emphasize open problems and challenges for future research in this topic.

Depth and Coverage: This 1.5-hour tutorial aims at establishing a research checkpoint. The objective is to structure, review and summarize research so far in analyzing social interactions between users and items in the Web. The tutorial covers fundamental developments in the topic from data mining and database perspectives, with a strong focus on applications. It also emphasizes open problems and challenges for future research.

Intended Audience: Among the many new sources of big data that have emerged in the last decade, social media content is among the most potent and offers unprecedented challenges and opportunities to researchers in the communities of database, data mining, information retrieval, and machine learning. The tutorial would be of theoretical and practical interest to a large part of this community. It is aimed at active researchers from academia and industry working on related problems, as well as graduate students and young researchers seeking a new research topic.

Assumed Background: The tutorial will not require prior knowledge beyond the basic concepts covered in introductory database and data mining classes.

Proceedings of the VLDB Endowment, Vol. 8, No. 12

2. TUTORIAL OUTLINE

Our tutorial is divided into four parts.

Part I - Introduction (20 mins): Social media analytics is the science of developing models and algorithms in order to understand users and user-generated content in social media. It can be classified into: *text analytics* that analyzes user-generated content, *temporal analytics* that studies the evolution of user-behavior over time, *interactive analytics* that explores user actions and reaction in social media, and *structured analytics* that discovers the structure of relationships emerging from social media use. In this tutorial, we focus on structured analytics in social media, and present a series of real-world example applications. We identify the challenges, and systemize the research under this topic by identifying two broad objectives that focus on analyzing:
(I) implicit behavior-based relationships that arises due to shared interests exemplified by users' online activities
(II) explicit link-based relationships that arises due to direct connections between users.

For each objective, we consider two popular structured analytics approaches and review their technical developments in the order of problem formalization, data model, algorithmic solution(s), and experimental evaluation.

Part II - Analytics for Implicit Behavior-Based Relationships (30 mins): This part of the tutorial summarizes research that considers structured analytics techniques to identify meaningful patterns in social media content that are behavioral relationships arising due to similarities (or, dissimilarities) in user interaction with items, and with each other. Each of these problems involves a task that leverages the rich metadata associated with the objects in the data (e.g., users, items) in order to conduct the simultaneous analysis of how the object attributes (e.g, user demographics, item descriptions, etc.) influence user activities in the web. We cover two specific work in details that capture a reasonable set of mining tasks and applications.

(i) Data Cube and Lattice: The roots of structured analytics can be traced to discovery driven exploration of OLAP data cubes. We discuss how the authors in [1][2] model content from collaborative content sites as data cube lattice, and mine interesting patterns that reveal how ratings and tags are assigned by certain users to certain items.

(i) Relational DBMS and SQL: Recent times have witnessed the popularity of databases for storing, managing, and analyzing social data. We study how a full-fledged database system can be made to provide an effective solution to the popular machine learning problem of providing recommendations to users by mining social data [3].

Part III - Analytics for Explicit Link-Based Relationships(30 mins): This part of the tutorial summarizes research that considers structured analytics techniques to identify meaningful patterns in social web content that are social ties arising due to direct connections between users. Each of these problems involves a task that leverages the topological properties of a social information network, and identify connected sub-structures involving the data objects (e.g., users, items) in order to analyze how the network connections influence user activities. Once again, we cover two analytics approaches in details that capture a reasonable set of social analytics tasks and applications.

(i) Meta Structure and Network Schema: Most real-world social data today form heterogeneous information networks consisting of multiple typed objects and multiple typed links. We discuss how the authors in [4][5] conduct similarity search in such networks by systematically leveraging the description of the meta structure of the network.

(ii) OLAP and Aggregation: We also study structured analytics technique for mining general graph data that commonly represents user activities, intents, and interactions in the web. We illustrate how an an aggregate analytics based approach summarizes graph data better than popular statistical approaches [6].

Part IV - Conclusions and Future Directions(10 mins): We discuss future work and our research perspective at the end of each of the individual work presented. In addition, we conclude our tutorial by highlighting several open problems.

3. BIOGRAPHICAL SKETCHES

Mahashweta Das is a Research Scientist at HP Labs, Palo Alto where she works in the Analytics group. She graduated with a PhD in Computer Science from the University of Texas at Arlington in 2013. Her research interests include databases, data mining, machine learning, algorithms and social computing. She has published several refereed articles in top-tier conferences such as SIGKDD, VLDB, and SIGMOD. Her VLDB paper was selected for Best Papers of 2012 VLDB. Her PhD dissertation "Exploratory Mining of Collaborative Social Content" received Honorable Mention at ACM SIGKDD 2014 Doctoral Dissertation Award.

Gautam Das is a Full Professor in the Computer Science and Engineering Department of the University of Texas at Arlington. Prior to UTA, Dr. Das has held positions at Microsoft Research, Compaq Corporation and the University of Memphis. He graduated with a BTech in Computer Science from IIT Kanpur, India, and with a PhD in computer science from the University of Wisconsin−Madison. Dr. Das's research interests span social computing, data mining, information retrieval, databases, graph and network algorithms, and computational geometry. His research has resulted in many papers at premier conferences and journals. His work has received several awards, including the IEEE ICDE 2012 Influential Paper Award.

4. SELECTED REFERENCES

[1] M. Das, S. Amer-Yahia, G. Das, and C. Yu. MRI: meaningful interpretations of collaborative ratings. *PVLDB*, 4(11):1063–1074, 2011.

[2] M. Das, S. Thirumuruganathan, S. Amer-Yahia, G. Das, and C. Yu. An expressive framework and efficient algorithms for the analysis of collaborative tagging. *VLDB Journal*, 23(2):201–226, 2014.

[3] M. Sarwat, J. Avery, and M. F. Mokbel. A recdb in action: Recommendation made easy in relational databases. *PVLDB*, 6(12):1242–1245, 2013.

[4] Y. Sun and J. Han. Mining heterogeneous information networks: a structural analysis approach. *SIGKDD Explorations*, 14(2):20–28, 2012.

[5] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003, 2011.

[6] Y. Tian, R. A. Hankins, and J. M. Patel. Efficient aggregation for graph summarization. In *Proceedings of ACM SIGMOD*, pages 567–580, 2008.