Introduction to the Special Section on Search and Mining User-Generated Content

JOSÉ CARLOS CORTIZO and FRANCISCO CARRERO, BrainSINS
IVÁN CANTADOR, Autonomous University of Madrid
JOSÉ ANTONIO TROYANO, University of Seville
PAOLO ROSSO, Technical University of Valencia

The primary goal of this special section of *ACM Transactions on Intelligent Systems and Technology* is to foster research in the interplay between Social Media, Data/Opinion Mining and Search, aiming to reflect the actual developments in technologies that exploit user-generated content.

Categories and Subject Descriptors: H.3 [**Information Systems**]: Information Storage and Retrieval; H.4 [**Information Systems**]: Information Systems Applications; I.2 [**Computing Methodologies**]: Artificial Intelligence; I.7 [**Computing Methodologies**]: Document and Text Processing

General Terms: Documentation, Experimentation, Algorithms

Additional Key Words and Phrases: Search, data mining, text mining, opinion mining, information retrieval, user-generated contents, social media

ACM Reference Format:
Cortizo, J. C., Carrero, F., Cantador, I., Troyano, J. A., and Rosso, P. 2012. Introduction to the special section on search and mining user-generated content. ACM Trans. Intell. Syst. Technol. 3, 4, Article 65 (September 2012), 3 pages.
DOI = 10.1145/2337542.2337550 http://doi.acm.org/10.1145/2337542.2337550

1. INTRODUCTION

Social Media have been able to shift the way information is generated and consumed. At first, information was generated by one person and consumed by many people, but now most information available on the Web is generated by users, which has changed the need of information access and management. Social Networks like Facebook or Twitter manage tens of PBs of information, with flows of hundreds of TBs per day, and hundreds of billions of relationships.

User-generated content provides an excellent scenario to apply the metaphor for mining any kind of information. In a social media context, users create a huge amount of data where we can look for valuable nuggets of knowledge by applying diverse search (information retrieval) and mining techniques (data mining, text mining, Web mining, opinion mining). In these kinds of data, we can find both structured information (ratings, tags, links) and unstructured information (text, audio, video), and we have to learn how to combine existing techniques in order to take advantage of the existing information heterogeneity while extracting useful knowledge.

The European Commission as part of the WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework partially supported the special issue on search and mining user-generated contents.

DOI 10.1145/2337542.2337550 http://doi.acm.org/10.1145/2337542.2337550

The primary goal of this special section of *ACM Transactions on Intelligent Systems and Technology* is to foster research in the interplay between Social Media, Data/Opinion Mining and Search, aiming to reflect the actual developments in technologies that exploit user-generated content.

2. ACCEPTED ARTICLES

For this special section, we received a total of 42 high-quality submissions, from which the following 9 were chosen for publication.

In "Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media", Paltoglou and Thelwall present a lexicon-based, unsupervised approach to sentiment analysis applicable to polarity and subjectivity detection in user-generated content typical of social media channels (datasets include Twitter, MySpace, and Digg). The authors support their argument by testing their method against traditional methods using three different real-world datasets extracted from three well-know social networks.

In "Leveraging Social Bookmarks from Partially Tagged Corpus for Improved Webpage Clustering" Trivedi et al. take a step further exploiting document and tag content to improve Web document clustering. The authors describe how multiview learning can be used for this problem, and they show a detailed comparative study on methods and features. One of the most important contributions of this article is a model for managing incomplete data that has great potential and possible applications in NLP-related tasks.

In "Information Retrieval in the Commentsphere", Potthast et al. review the literature for several information retrieval tasks for comments on Web items, such as blog posts. A wide-ranging literature survey organizes features and approaches within a framework. Moreover, indicative experiments are carried out on three tasks using two comment corpora.

In "On the Relationship between Novelty and Popularity of User-Generated Content", Carmel et al. study the effect of novelty on the popularity of user-contributed content in social media. The authors propose several measures of relative novelty with respect to different contexts of the author's previous posts, other users' comments on the latter, and all other users' posts. Novelty is measured as the normalized compression distance between a post and, respectively, the three aforementioned contexts.

Li et al. propose an entity-relationship query to search entities in the Wikipedia corpus in the article "Entity-Relationship Queries over Wikipedia". This approach supports multiple keyword-based predicates. The authors present a ranking framework and a Bounded Cumulative Model for accurate ranking of query answers.

In "EachWiki: Facilitating Wiki Authoring by Annotation Suggestion", Wang et al. propose a unified suggestion model and apply it to link, category, and semantic relation recommendation services, which aims at lightening the burden of both contributors and administrators in Wikipedia.

In "Nowcasting Events from the Social Web with Statistical Learning", Lampos and Christiani present a general methodology for inferring the occurrence and magnitude of an event or phenomenon by exploring the rich amount of unstructured textual information on the social part of the Web.

In "Ranking User Influence in Healthcare Social Media", Tang and Yang propose how to identify the influential users in a healthcare-related forum. Given a forum thread, a social network is constructed based on ask-reply/response relationships and influential users are identified as those with either the highest in-degree or the highest pagerank.

In “Evaluation of Folksonomy Induction Algorithms”, Strohmaier et al. report results from a broad comparative study of state-of-the-art folksonomy induction algorithms that have been applied and evaluated in the context of five social tagging systems.

We hope you enjoy reading these thought-provoking articles. They may very well form the basis for future research in this important field.