Information Extraction from Digital Social Trace Data with Applications to Social Media and Scholarly Communication Data

Shubhanshu Mishra
School of Information Sciences.
University of Illinois at Urbana-Champaign
mishra@shubhanshu.com

Abstract

Information extraction (IE) aims at extracting structured data from unstructured or semi-structured data. The thesis starts by identifying social media data and scholarly communication data as a special case of digital social trace data (DSTD). This identification allows us to utilize the graph structure of the data (e.g., user connected to a tweet, author connected to a paper, author connected to authors, etc.) for developing new information extraction tasks. The thesis focuses on information extraction from DSTD, first, using only the text data from tweets and scholarly paper abstracts, and then using the full graph structure of Twitter and scholarly communications datasets. This thesis makes three major contributions.

First, new IE tasks based on DSTD representation of the data are introduced. For scholarly communication data, methods are developed to identify article and author level novelty [Mishra and Torvik, 2016] and expertise. Furthermore, interfaces for examining the extracted information are introduced. A social communication temporal graph (SCTG) is introduced for comparing different communication data like tweets tagged with sentiment, tweets about a search query, and Facebook group posts. For social media, new text classification categories are introduced, with the aim of identifying enthusiastic and supportive users, via their tweets. Additionally, the correlation between sentiment classes and Twitter meta-data in public corpora is analyzed, leading to the development of a better model for sentiment classification [Mishra and Diesner, 2018].

Second, methods are introduced for extracting information from social media and scholarly data. For scholarly data, a semi-automatic method is introduced for the construction of a large-scale taxonomy of computer science concepts. The method relies on the Wikipedia category tree. The constructed taxonomy is used for identifying key computer science phrases in scholarly papers, and tracking their evolution over time. Similarly, for social media data, machine learning models based on human-in-the-loop learning [Mishra et al., 2015], semi-supervised learning [Mishra and Diesner, 2016], and multi-task learning [Mishra, 2019] are introduced for identifying sentiment, named entities, part of speech tags, phrase chunks, and super-sense tags. The machine learning

models are developed with a focus on leveraging all available data. The multi-task models presented here result in competitive performance against other methods, for most of the tasks, while reducing inference time computational costs.

Finally, this thesis has resulted in the creation of multiple open source tools and public data sets (see URL below), which can be utilized by the research community. The thesis aims to act as a bridge between research questions and techniques used in DSTD from different domains. The methods and tools presented here can help advance work in the areas of social media and scholarly data analysis.

Awarded by: University of Illinois at Urbana-Champaign.
Supervised by: Jana Diesner
Available at: https://shubhanshu.com/phd_thesis/

Selected Publications

Shubhanshu Mishra and Vetle I. Torvik. Quantifying Conceptual Novelty in the Biomedical Literature. *D-Lib Magazine*, 22(9/10), 9 2016. ISSN 1082-9873. doi: 10.1045/september2016-mishra. URL http://www.dlib.org/dlib/september16/mishra/09mishra.html.

Shubhanshu Mishra and Jana Diesner. Detecting the Correlation between Sentiment and User-level as well as Text-Level Meta-data from Benchmark Corpora. In *Proceedings of the 29th on Hypertext and Social Media - HT '18*, pages 2–10, New York, New York, USA, 2018. ACM Press. ISBN 9781450354271. doi: 10.1145/3209542.3209562. URL http://dl.acm.org/citation.cfm?doid=3209542.3209562.

Shubhanshu Mishra, Jana Diesner, Jason Byrne, and Elizabeth Surbeck. Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, pages 323–325, New York, New York, USA, 2015. ACM Press. ISBN 9781450333955. doi: 10.1145/2700171.2791022. URL http://doi.acm.org/10.1145/2700171.2791022http://dl.acm.org/citation.cfm?doid=2700171.2791022.

Shubhanshu Mishra and Jana Diesner. Semi-supervised Named Entity Recognition in noisy-text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 203–212, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. URL http://aclanthology.info/papers/semi-supervised-named-entity-recognition-in-noisy-texthttp://aclweb.org/anthology/W16-3927.

Shubhanshu Mishra. Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media - HT '19*, pages 283–284, New York, New York, USA, 2019. ACM Press. ISBN 9781450368858. doi: 10.1145/3342220.3344929. URL http://dl.acm.org/citation.cfm?doid=3342220.3344929.