COLING 2012

24th International Conference on Computational Linguistics

Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data

Workshop chairs:
Sriram Raghavan and Ganesh Ramakrishnan

9 December 2012
Mumbai, India

Diamond sponsors

Tata Consultancy Services
Linguistic Data Consortium for Indian Languages (LDC-IL)

Gold Sponsors

Microsoft Research
Beijing Baidu Netcon Science Technology Co. Ltd.

Silver sponsors

IBM, India Private Limited
Crimson Interactive Pvt. Ltd.
Yahoo
Easy Transcription & Software Pvt. Ltd.

Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data
Sriram Raghavan and Ganesh Ramakrishnan (eds.)
Revised preprint edition, 2012

Published by The COLING 2012 Organizing Committee
Indian Institute of Technology Bombay,
Powai,
Mumbai-400076
India
Phone: 91-22-25764729
Fax: 91-22-2572 0022
Email: pb@cse.iitb.ac.in

Also available online in the ACL Anthology at http://aclweb.org

Preface

The growth of online social media represents a fundamental shift in the generation, consumption, and sharing of digital information online. Social media data comes in many forms: from blogs (Blogger, LiveJournal) and micro-blogs (Twitter) to social networking (Facebook, LinkedIn, Google+), wikis, social bookmarking (Delicious), reviews (Yelp), media sharing (Youtube, Flickr), and many others. The information inherent in these online conversations is a veritable gold mine with the ability to influence every aspect of a modern enterprise – from marketing and brand management to product design and customer support. However, the task of drawing concrete, relevant, trustworthy, and actionable insights from the ever increasing volumes of social data presents a significant challenge to current day information management and business intelligence systems. As a result, there is growing interest and activity in the academic and industrial research communities towards various fundamental questions in this space:

- How do we collect, curate, and cleanse massive amounts of social media data?
- What new analytic techniques, models, and algorithms are required to deal with the unique characteristics of social media data?
- How does one combine information extracted from textual content with the structural information in the "network" (linking, sharing, friending, etc.)?
- What kind of platforms and infrastructure components are required to support all of these analytic activities at scale?

Currently, relevant work in this area is distributed across the individual conferences and workshops organized by different computer science research disciplines such as information retrieval, database systems, NLP, and machine learning. Furthermore, a lot of interesting innovations and hands-on experience from industrial practitioners is not publicly available. This workshop aims to bring together industrial and academic practitioners with a focus on an aspect of this problem of particular relevance to COLING – namely, robust and scalable techniques for information extraction and entity analytics on social media data.

We have put together an interesting program which consists of keynotes from two well known researchers, Marius Pasca (Google Research, Mountainview) and Dan Roth (University of Illinois at Urbana Champaign) which provides both an industry and academic perspective to the challenges of information extraction from social media data.

Marius Pasca's talk is on *Extracting Knowledge from the Invisible Social Graph*. The background, needs and interests of Web users influence the social relations they choose to have with other users, and the knowledge exchanged among users as part of those relations. The same factors influence the choice of queries submitted by users to Web search engines. Multiple users may refer to entities and relations of general interest in their posts and updates shared with others, just as they may refer to them in their search queries. In his talk, Marius discusses the types and uses of knowledge that can be extracted from collectively-submitted search queries, relative to extracting knowledge encoded in social media data.

Dan Roth's talk is on *Constraints Driven Information Extraction and Trustworthiness*. Computational approaches to problems in Natural Language Understanding and Information Extraction often involve assigning values to sets of interdependent variables. Examples of tasks of interest include semantic role labeling (analyzing natural language text at the level of "who did what to whom, when and where"), information extraction (identifying events, entities and relations), and textual entailment (determining whether one utterance is a likely consequence of another). However, while information extraction aims at telling us what a document says, we are also interested in knowing whether we can believe the claims made and the sources making them.Over the last few years, one of the most successful approaches to studying global decision problems in Natural Language Processing and Information Extraction involves Constrained Conditional Models (CCMs), an Integer Learning Programming formulation that augments probabilistic models with declarative constraints as a way to support such decisions. In this talk, Dan will present research within this framework, discussing old and new results pertaining to training these global models, inference issues, and the interaction between learning and inference. Most importantly, he will discuss extending these models to deal also with the information trustworthiness problem: which information sources we can trust and which assertions we can believe.

Tien Thanh Vu et. al. present an interesting application of predicting the stock prices by extracting and predicting the sentiments present in Twitter messages of these stocks. This is a very practical application of anticipating the fluctuations of stock markets using information gleaned from social media

Finding trends in social media is a very important activity and has a lot of practical applications, that include marketing. Nigel Dewdney et. al., present a study of the trends originating from blogs and contrasts it with trends from news on current affairs. Does information present in blogs reflect mainstream news or does it provide other trending topics which is significantly different from news: this is the matter of study in this work.

Twitter, a microblogging service, has been a very rich source of social media content and contain a large amount of short messages. Due to the short nature of messages, this type of an information extraction task can be quite challenging. The work by Kamel Nebhi discusses a rule-based system for identifying and disambiguating named-entities from tweets. Apoorv Agarwal et. al. present a system to perform end-to-end sentiment analysis of tweets. The system also explores the design challenges in building a sentiment analysis engine for twitter streams.

Organizing Team
Workshop on Information Extraction and Entity Analytics - 2012

Organizing Committee:

Sriram Raghavan (IBM Research - India)
Ganesh Ramakrishnan (IIT Bombay)
Ajay Nagesh (IIT Bombay)

Programme Committee:

Sunita Sarawagi (IIT Bombay)
Indrajit Bhattacharyya (Indian Institute of Science, Bangalore)
Rajasekar Krishnamurthy (IBM Research - Almaden)
L. V. Subramanian (IBM Research - India)
Sundararajan Sellamanickam (Yahoo! Labs, Bangalore)
Rahul Gupta (Google Inc, USA)
Anhai Doan (University of Wisconsin-Madison and WalmartLabs)
Kevin Chen-Chuan Chang (University of Illinois at Urbana-Champaign)
Parag Singla (IIT Delhi)
Mausam (University of Washington at Seattle)

Invited Speakers:

Marius Pasca (Google Research)
Dan Roth (University of Illinois at Urbana-Champaign)

Table of Contents

Workshop on Information Extraction and Entity Analytics on Social Media Data

Program

Saturday, 9 December 2012

09:15–09:30	Welcome and Introduction
09:30–10:30	**Keynote Talk** *Extracting Knowledge from the Invisible Social Graph* Marius Pasca, Google Research
	Paper presentations – Session 1
10:30–11:00	*Named Entity Trends Originating from Social Media* Nigel Dewdney
11:00–11:30	*Ontology-Based Information Extraction from Twitter* Kamel Nebhi
11:30–12:00	Tea break
12:00–13:00	**Keynote Talk:** *Constraints Driven Information Extraction and Trustworthiness* Dan Roth, University of Illinois at Urbana-Champaign
13:00–13:30	*– Buffer –*
13:30–14:30	Lunch
	Paper Presentations – Session 2
14:30–15:00	*An Experiment in Integrating Sentiment Features for Tech Stock Prediction in Twitter* Tien Thanh Vu, Shu Chang, Quang Thuy Ha and Nigel Collier
15:00–15:30	*End-to-End Sentiment Analysis of Twitter Data* Apoorv Agarwal and Jasneet Sabharwal
15:30–16:00	*Leveraging Latent Concepts for Retrieving Relevant Ads For Short Text* Ankit Patil, Kushal Dave and Vasudeva Varma
16:00–16:15	Closing Remarks