

# Knowledge Extraction and Consolidation from Social Media (KECSM 2012)

## Preface

In this new information age, where information, thoughts and opinions are shared so prolifically through online social networks, tools that can make sense of the content of these networks are paramount. In order to make best use of this information, we need to be able to distinguish what is important and interesting, and how this relates to what is already known. Social web analysis is all about the users who are actively engaged and generate content. This content is dynamic, rapidly changing to reflect the societal and sentimental fluctuations of the authors as well as the ever-changing use of language. While tools are available for information extraction from more formal text such as news reports, social media affords particular challenges to knowledge acquisition, such as multilinguality not only across but within documents, varying quality of the text itself (e.g. poor grammar, spelling, capitalisation, use of colloquialisms etc), and greater heterogeneity of data. The analysis of non-textual multimedia information such as images and video offers its own set of challenges, not least because of its sheer volume and diversity. The structuring of this information requires the normalization of this variability by e.g. the adoption of canonical forms for the representation of entities, and a certain amount of linguistic categorization of their alternative forms.

Due to the reasons described above, data and knowledge extracted from social media often suffers from varying, non-optimal quality, noise, inaccuracies, redundancies as well as inconsistencies. In addition, it usually lacks sufficient descriptiveness, usually consisting of labelled and, at most, classified entities, which leads to ambiguities.

This calls for a range of specific strategies and techniques to consolidate, enrich, disambiguate and interlink extracted data. This in particular benefits from taking advantage of existing knowledge, such as Linked Open Data, to compensate for and remedy degraded information. A range of techniques are exploited in this area, for instance, the use of linguistic and similarity-based clustering techniques or the exploitation of reference datasets. Both domain-specific and cross-domain datasets such as DBpedia or Freebase can be used to enrich, interlink and disambiguate data. However, case- and content-specific evaluations of quality and performance of such approaches are missing, hindering the wider deployment. This is of particular concern, since data consolidation techniques involve a range of partially disparate scientific topics (e.g. graph analysis, data

## II

mining and interlinking, clustering, machine learning), but need to be applied as part of coherent workflows to deliver satisfactory results.

The KECSM 2012 workshop aims to gather innovative approaches for knowledge extraction and consolidation from unstructured social media, in particular from degraded user-generated content (text, images, video) such as tweets, blog posts, forums and user-generated visual media. KECSM has gathered novel works from the fields of data analysis and knowledge extraction, and data enrichment, interlinking and consolidation. Equally, consideration has been given to the application perspective, such as the innovative use of extracted knowledge to navigate, explore or visualise previously unstructured and disparate Web content.

KECSM 2012 had a number of high-quality submissions. From these, the 8 best papers were chosen for the two paper sessions of the programme. To initiate the workshop, a keynote on perspectives of social media mining from an industry viewpoint was given by Seth Grimes.

We sincerely thank the many people who helped make KECSM 2012 such a success: the Program Committee, the paper contributors, and all the participants present at the workshop. In addition, we would like to add a special note of appreciation for our keynote speaker, Seth Grimes, and the ARCOMEM project (<http://www.arcomem.eu>) for funding the best paper prize.

Diana Maynard  
Stefan Dietze  
Wim Peters  
Jonathon Hare

# Organisation

## Organising Committee

Diana Maynard, University of Sheffield, United Kingdom  
Stefan Dietze, L3S Research Centre, Leibniz University Hannover, Germany  
Wim Peters, University of Sheffield, United Kingdom  
Jonathon Hare, University of Southampton, United Kingdom

## Program Committee

Harith Alani, The Open University, United Kingdom  
Sören Auer, University of Leipzig, Germany  
Uldis Bojar, University of Latvia, Latvia  
John Breslin, NUIG, Ireland  
Mathieu D'Aquin, The Open University, United Kingdom  
Anita de Waard, Elsevier, The Netherlands  
Adam Funk, University of Sheffield, United Kingdom  
Daniela Giordano, University of Catania, Italy  
Alejandro Jaimes, Yahoo! Research Barcelona, Spain  
Paul Lewis, University of Southampton, United Kingdom  
Véronique Malaisé, Elsevier, The Netherlands  
Pavel Mihaylov, Ontotext, Bulgaria  
Wolfgang Nejdl, L3S Research Centre, Leibniz University Hannover, Germany  
Thomas Risse, L3S Research Centre, Leibniz University Hannover, Germany  
Matthew Rowe, The Open University, United Kingdom  
Milan Stankovic, Hypios & Universit Paris-Sorbonne, France  
Thomas Steiner, Google Germany, Germany  
Nina Tahmasebi, L3S Research Centre, Leibniz University Hannover, Germany  
Raphael Troncy, Eurecom, France  
Claudia Wagner, Joanneum Research, Austria

## Keynote Speaker

Seth Grimes, Alta Plana Corporation, USA

## Sponsors

The best paper award was kindly sponsored by the European project AR-COMEM (<http://arcomem.eu>).