Overview of the 2nd International Workshop on Search and Mining User-generated Contents

Jose Carlos Cortizo
BrainSins
Avda. M-40 N°15 1°10
Alcorcon, Madrid (Spain)
j.cortizo@brainsins.com

Francisco Carrero
BrainSins
Avda. M-40 N°15 1°10
Alcorcon, Madrid (Spain)
f.carrero@brainsins.com

Ivan Cantador
U. Autonoma de Madrid
C/Frco. Tomas y Valiente 11
Madrid, Spain
ivan.cantador@uam.org

Jose Antonio Troyano
Universidad de Sevilla
Avda. Reina Mercedes s/n
Sevilla, Spain
troyano@us.es

Paolo Rosso
U. Politecnica de Valencia
Camino de Vera s/n
Valencia, Spain
prosso@dsic.upv.es

ABSTRACT

This overview introduces the aim of the SMUC 2010 workshop, as well as the list of papers presented in the workshop.

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

Keywords

search, data mining, text mining, opinion mining, information retrieval, user-generated contents, social media

1. AIM OF THE WORKSHOP

The 2nd International Workshop on Search and Mining User-generated Contents (SMUC) was be held in Toronto, Canada, as a workshop of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), on October 30, 2010.

SMUC 2010 aims to become a forum for researchers from several Information and Knowledge Management areas like data/text mining, information retrieval, semantics, etc. that apply their work into the fields of Social Media and Opinion/Sentiment Analysis where the main goal is to process user generated contents.

User generated content provides an excellent scenario to apply the metaphor of 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 several search techniques (information retrieval) or mining techniques (data mining, text mining, web mining, opinion mining, etc.). In this kind of data we can find both structured information (ratings, tags, links, etc.) and unstructured information (text, audio, video, etc.), and we must learn to combine existing techniques in order to take advantage of this heterogeneity while extracting useful knowledge.

Copyright is held by the author/owner(s).
CIKM'10, October 26–29, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0099-5/10/10.

2. SMUC STATISTICS

SMUC workshop received a total of 25 submissions, 10 from Asia, 9 from North-America, 5 from Europe and 1 from Africa. From all the received submissions, 8 (32%) were accepted as full paper submissions and 7 papers (28%) were accepted as posters.

If we focus on the main topics of the workshop, the distribution of accepted papers (full papers and posters) is the following one. Seven papers are focused on Data/Text Mining topics, 4 papers are related to Opinion Mining and the remaining 4 papers focuses on Search, Spam filtering and Tagging.

Another distribution of papers is obtained considering the information sources used for experimentation: 4 accepted papers used Twitter as main information source, another 4 papers used Product Reviews (mainly Amazon) for their experimentation, 3 papers used Webpages, and the remaining 4 papers used Wikipedia, Programable Web, Delicious and Blogs.

3. FULL PAPERS

"A Knowledge-Rich Approach to Feature-Based Opinion Extraction from Product Reviews" describes a domain-specific resource-based system for opinion extraction, focusing on the description and generation of those resources.

"Exploiting Tag and Word Correlations for Improved Webpage Clustering" considers page-text and tags as two separate views of the data, and learn a shared subspace that maximizes the correlation between the two views, in order to improve webpage clustering and categorization.

"Entity-Relationship Queries over Wikipedia" proposes a structured query mechanism, entity-relationship query, for searching entities in Wikipedia corpus by their properties and inter-relationships.

"A Formal Study of Classification Techniques on Entity Discovery and their Application to Opinion Mining" focuses on examining the effectiveness of various classification techniques on entity discovery and their application to the opinion mining task.

"Classifying Latent User Attributes in Twitter" includes a novel investigation of stacked-SVM-based classification algorithms over a rich set of original features, applied to classifying four latent user attributes (gender, age, regional origin, and political orientation).

"Spam Detection with a Content-based Random-walk Algorithm" presents a novel approach to web spam detection based on a random-walk algorithm that obtains a ranking of pages according to their relevance and their spam likelihood.

"Exploiting Web Reviews for Generating Customer Service Surveys" proposes a method of automatically generating service surveys through mining Web reviews to address the issue of the subjectivity present on traditional surveys.

"Characterization of the Twitter @replies Network: are User Ties Social or Topical?" presents an exhaustive characterization of a dataset from a popular micro-blogging service, Twitter. A quantitative analysis of the users' interactions in the implicit network derived from tweet replies, is also presented.

4. POSTERS

"Mining Social Tags to Predict Mashup Patterns" proposes a tag-based approach for predicting mashup patterns, thus deriving inspiration for potential new mashups from the community's consensus.

"A Weighted Tag Similarity Measure Based on a Collaborative Weight Model" proposes a weight model for tagging systems that considers the user dimension unlike existing measures based on tag frequency. This work also proposes weighted similarity model that is conceptually different from the contemporary frequency based similarity measures.

"How to Interpret the Helpfulness of Online Product Reviews: Bridging the Needs Between Customers and Designers" focuses on how to automatically build the connection between online customer's voting and designer's rating and predict the customer reviews' helpfulness based on the review content.

"Web-based Statistical Fact Checking of Textual Documents" aims to make use of the web-content to calculate a statistical support score for textual documents.

"On the Difficulty of Clustering Company Tweets" presents and compare a number of different approaches based on clustering that determine whether a given tweet refers to a particular company or not.

"Cross-media Impact on Twitter in Japan" reports the characteristics of Twitter users in Japan, and the impact of media such as publications, and TV programs on Twitter community.

"Extracting Emotion Topics from Blog Sentences - Use of Voting from Multi-Engine Supervised Classifiers" presents a supervised multi-engine classifier approach followed by voting to identify emotion topic(s) from English blog sentences.

5. KEYNOTE SPEAKER

SMUC's keynote speaker will be Bing Liu (University of Illinois at Chicago), a reputed expert on Web and Text Mining and Sentiment Analysis. Bing Liu is a Professor of Computer Science at the University of Illinois at Chicago (UIC). He obtained his PhD in Artificial Intelligence from the University of Edinburgh. Before joining UIC in 2002, he was with the National University of Singapore. He has published extensively in the fields of data mining, Web mining and opinion mining in leading conferences and journals. His research has been focused on classification based on associations, interestingness in data mining, learning from positive and unlabeled examples, Web data/information extraction, and opinion mining and sentiment analysis. He has also written a textbook titled "Web Data Mining: Exploring Hyperlinks, Contents and Usage Data". On professional services, Liu has served as associate editors of IEEE Transactions on Knowledge and Data Engineering, and SIGKDD Explorations, and is in the editorial boards of several other journals. He also served or serves as program chairs of IEEE International Conference on Data Mining (ICDM-2010), ACM Conference on Web Search and Data Mining (WSDM-2010), ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2008), SIAM Conference on Data Mining (SDM-2007), ACM Conference on Information and Knowledge Management (CIKM-2006), and Pacific Asia Conference on Data Mining (PAKDD-2002), and as area chairs of International World Wide Web Conference (WWW-2005, WWW-2010) in charge of the data mining track. In addition, he has served extensively as program committee members, and senior program committee members of leading conferences in data mining, Web technologies and natural language progressing.

6. INDUSTRY PANEL

An industry panel will be held as the conclusion of SMUC workshop. In this panel, experts on search and mining user-generated contents from several companies will discuss about the actual and future needs on technologies related to SMUC topics.

7. ACKNOWLEDGMENTS

We would like to thank the ACM CIKM 2010 organizing committee for their support in making this workshop possible. We also appreciate very much the dedication of the reviewers and the cooperation of our fellow organizing committee members.