Special Issue

Georg Fuchs*, Hendrik Stange, Ahmad Samiei, Gennady Andrienko, and Natalia Andrienko

A semi-supervised method for topic extraction from micro postings

Abstract: Social networking services have become a major channel for the digital society to share content, opinions, experiences on activities or events, as well as on products, services and brands. Evaluating digital feedback on the latter can be a valuable asset for companies seeking product and consumer insights. However, the analysis of short, noisy, fragmented, and often subjective textual data still remains a challenge. Typically, the human analyst needs to be actively involved during extraction and modeling to resolve ambiguities that will inevitable arise in such data and to put the model into context. This paper proposes a visual analytics approach that enables a first intuition and exploration of topics appearing in the text corpus, and facilitates the interactive-iterative refinement of the overall topic model describing the stream of tweets. A second contribution is the discussion of efficient graph community detection algorithms to extract initial topics as the starting point of interactive analysis that complement approaches such as LDA. The applicability and utility of the proposed approach is shown for a real-world use case: the analysis of product insights and topic-driven social networks analysis for a specific product line for an international hair styling and cosmetics company.

Keywords: Text analysis, visual analytics, topic models, social media.

ACM CCS: Information systems → Information retrieval → Document representation → Document topic models, Information systems → World Wide Web → Web applications → Social networks, Mathematics of computing → Discrete mathematics → Graph algorithms

DOI 10.1515/itit-2014-1078
Received August 13, 2014; revised November 3, 2014; accepted December 5, 2014

***Corresponding author: Georg Fuchs,** Fraunhofer IAIS, Sankt Augustin, e-mail: georg.fuchs@iais.fraunhofer.de
Hendrik Stange, Ahmad Samiei, Gennady Andrienko, Natalia Andrienko: Fraunhofer IAIS, Sankt Augustin

1 Introduction

An ever increasing number of people use social media such as Facebook and Twitter for varied purposes that include staying in touch with friends, sharing opinion or feedback about products, services and events of interest, maintaining situational awareness, receiving news, or following celebrities and politicians. As a result social media are now producing massive, real-time streams of largely unstructured textual content that reflect the zeitgeist of society which traditional media were not able to produce. This increased demand for effective and scalable means to extract meaningful patterns and relevant information from these data for a variety of applications, from consumer relations [24], market research [20], social sciences [18], to real-time distributed sensing e. g. in crisis management [5, 23]. A key challenge is the extraction of meaningful, possibly latent topics from these huge volumes of heterogeneous and unstructured textual data, ideally in (near) real-time. As a result, social media analysis has become a highly active research topic across different disciplines including information retrieval, text mining, machine learning and visual analytics.

Extracting topics from conventional text such as scientific articles and news has been a focus of research for a long time and there are already state of the art methods with quite acceptable accuracy. These methods are mainly based on co-occurrences of words in each topic. However, microblog texts such as Tweets are very short and typically contain many abbreviations, slang and misspellings. Thus unlike in regular text corpora (e. g., articles), there usually is not enough concurrency of thematically related words in each Tweet, thus capturing topics inherent in these data remains a challenge.

To address this challenge, the approach presented here combines algorithmic topic extraction with visual-interactive refinement of the topic model. It further links this topic model with a sociograph over the text corpus, i. e., a representation of what communities engage in which topics. Its main tenet is to leverage tacit domain knowledge of the human analyst to compensate the typically reduced quality of algorithmic topic extraction from

microblogs. Thus, it aims at reducing the relative impact the automated topic extraction part has on the analysis process as a whole, rather than attempting to provide an improved extraction algorithm in itself.

2 Related work

Due to the proliferation of, and wide variety of topics discussed via social media, micro- and social blogging have been investigated by researchers in computer science, social science, and other disciplines. Twitter in particular has been used as a partly public source for as diverse activities as event detection and tracking [3, 11], co-creation processes [18], to sentiment analysis [22]. The format of tweets however makes many natural language processing (NLP) tasks, such as part-of-speech (POS) tagging [9], named entity recognition (NER) [13], and topic detection [26] much more challenging.

Several works utilize microblogs as distributed "social sensors", i. e. for the detection, localization, and tracking in space and time of abrupt (disaster) events such as earthquakes [5, 23]. Common to these approaches is the notion of "bursty" event-related and/or trending topics emerging rapidly from a continuous stream of posts, changing its overall topic distribution [12, 14]. This makes them quite accurate for detection of specific events in space and time (e. g., for disaster monitoring and mitigation) but much less suitable for in-depth analysis of semantic topics, user communities and information distribution channels related to more general concepts (e. g., product reception).

The Latent Dirichlet Allocation (LDA) topic modeling [1] family of approaches is probably the most widely used algorithm currently. And while base LDA performance on substantial texts is very good, being a generative Bayesian model the result quality for short and very short text declines significantly. A number of LDA variants have been proposed to improve the performance for microblogs and similar sources in particular.

(Semi-)supervised LDA variants try to address the sparsity problem by user-guided labeling of the corpus, such as sLDA [15], L-LDA [19] with Tweet-level supervision on labeling, and Tweet-LDA [26] designed specifically to handle language peculiarities encountered in Tweets. AT-LDA [21] associates each document with just a single topic (rather than a topic distribution) but also an author, making it more aligned with how Tweets and other microblog posts are usually structured.

Then there are approaches that not only rely on user-guided labeling of training data but even more significant involvement of the analyst by means of visual analytics. Topic Streams [7] is a web-based interactive visualization system that allows to follow and explore Twitter conversations on large-scale events. Vox Civitas [6] is aimed at helping journalists extract news from social media in reaction to broadcast events. Both, however, are limited to keyword-based topic definitions. Lead Lines [8] extracts time series of latent topics from CNN news articles and Twitter in parallel using standard LDA instead. Based on the I-SI architecture [25], it combines methods for temporal trending, event detection, and user-directed hierarchical topic (re)structuring in an interactive visual tool. Our approach similarly focuses on automatic processing for initial results and user-guided refinement. By comparison, we aim for the middle ground between simple keyword-based and the more cohesive, but also less stable (in terms of iterative refinement) topic models output by LDA.

3 Initial topic extraction based on graph community detection

For many real-world applications the analysis of loosely coupled micro postings such as Tweets requires contextual structuring, topic extraction, and social embedding. The task of this three-dimensional semantic structuring is rarely covered adequately in all aspects by current approaches which focus mainly on one or two dimensions.

This work was motivated by one of the authors' (unpublished) Master thesis in which several LDA variants were applied to a fixed set of three training corpora of 200 000, 1 000 000, and 3 000 000 tweets respectively from the UK. We could make the following observations:

- Despite using a stopword list specifically produced for this corpus [16] and reducing tweets to only verbs and nouns – one time each as solitary measure and once with both in conjunction – still irrelevant words were assigned to topics.
- Many tweets were assigned to semantically unrelated topics, mainly due to the lack of contextual content in the original tweet, which is further amplified by the necessary reduction during NLP pre-processing.
- LDA requires to specify the number of latent topics in the corpus beforehand as an input parameter [1]. We found topic cluster to be highly unstable, i. e., even minor changes to the topic count result in large changes to the overall topic-term associations.

Although the results gained from this study may partly be influenced by the test data (a sample of the Twitter stream) and we do not claim representativeness, we found the first

two issues in particular to be due to the following factors, which is also in line with what has been reported elsewhere [9, 26]:

- **Social Noise**: Users can send messages about virtually any topic, e. g. what they eat for lunch.
- **Synonyms and homonyms**: Words may have equal or different meanings which varies by context.
- **Non-standard language**: Usage of short message services has led to new linguistic diversification, e. g. abbreviations, neologisms, or creative hashtags.
- **Grammatical looseness**: Tweets (partly) ignore common syntax, punctuation or sequence of words.

To cope with these problems and limitations we propose a visual-interactive approach for guided topic detection and modeling that combines algorithmic topic extraction and non-parametric social network analysis with a human-centric topic refinement process. This facilitates a visual analytics workflow (see Figure 1) for topic extraction whereby the human analyst can impart feedback on intermediate results and thus guide the topic learning process to compensate for lack of contextual content in Twitter and other microblogging data.

The overall approach comprises five stages: (1) data pre-processing, (2) corpus term co-occurence graph creation, (3) algorithmic graph community detection (initial topic model **TG**), (4) creation of the Sociograph **SG** from observable conversation patterns, and (5) Visual-interactive exploration of the integrated **TG**–**SG** and iterative refinement of the topic model.

As a note on step (3), when designing our approach we initially used LDA for creation of the initial topic model. However, the instability of LDA in terms of what effects adjusting the predetermined topic count, as well as sensitivity to slight corpus changes (e. g., removing even a single tweet from a person identified as outlier in the **SG**), have on the returned topic clusters meant iterative parameter refinement would have hard to predict, and often difficult to understand, consequences. This was particularly true if the analysts were domain experts, not text analysis experts, as in our use case (see Section 5). For this reason, we opted to use a graph community detection (GCD) algorithm as an iterative, semi-supervised topic clustering method. GCD strikes a balance between very simple (keyword lists) and more comprehensive but far less stable topic clusters (LDA). It should be noted that our approach does not preclude the use of LDA variants – rather, GCD complements the set of available automatic topic extraction methods.

Data pre-processing entails cleansing message data from non-Latin symbols (e. g., Arabic) and text elements not directly relevant to topic extraction, such as URLs. Hashtags and @-tags (user names) are stripped of their prefix character (# and @, respectively) but are otherwise treated as regular words since they often are important sentence constituents in tweets. Next, we apply stopword removal as well as stemming, POS tagging, tokenization, and lemmatization.

Term co-occurrence graph creation takes every unique term in any Tweet as a graph node. A graph edge encodes the co-occurrence of two terms in at least one tweet. Node weights represent the relative frequency of term occurrences over the entire text corpus (all tweets). Likewise, edge weights represent the co-occurrence frequency of term pairs over all tweets.

Graph community detection means finding sub graphs (communities) which have densely connected nodes inside and are sparsely connected to other sub graphs [2, 10]. The modularity measure of a graph *partitioned* into communities is a value between $[-1, 1]$ indicating the ratio of link densities of (i. e., node degree with respect to) intra-community to inter-community links.

Sociographs SG encode message sender–recipient relationships. Formally, it is a weighted directed graph where each node represents a sending and/or receiving

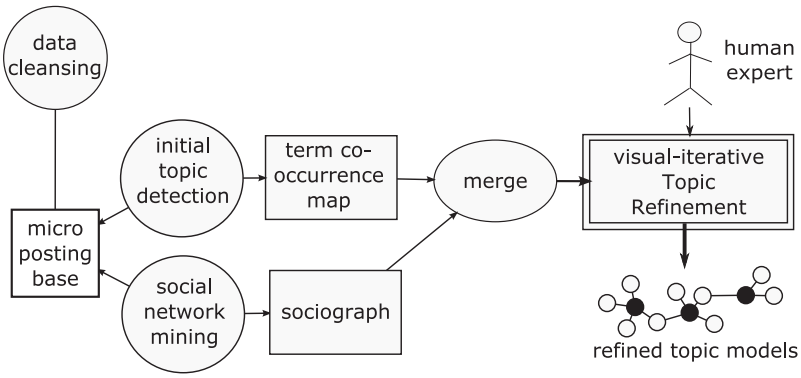

Figure 1: Visually guided topic extraction process flow.

user, and each edge indicates a direct communication link between nodes e. g., conversation, re-tweets etc. Edge weights (i. e., strength of user relationships) are calculated as $\sqrt{C_{SR}}$, C_{SR} the number of messages from sender S to recipient user R. The sociograph of a corpus thus represents user community structure information based on observable communication patterns, complementing the corpus' topic graph.

4 Visual-interactive topic model refinement

The Visual Data Exploration component initially displays the core communities found by the algorithm using several linked views (compare Figure 2):

- Primary (A) and secondary (B) tabular lists of detected topics (core communities) for topic pair selection.
- Detail view for terms (nodes) in the currently selected topic pair comprising a tabular ranked view of relative term frequencies (C) and a tag cloud of topic terms with frequency-weighted font (D).
- Digest view (E) displaying a sample of tweets associated with either or both selected topics.
- Graphical representation of the co-occurrence graph (see Figure 3, middle).

Contribution of each of the topics is consistently color-coded in all views – blue for primary, magenta for secondary topic. These views are designed to give the analyst fast insights regarding overall topic composition, microblog associations, and their pairwise relation with respect to common terms.

Closer examination of topics is supported via a set of filters, Figure 2 (F). These include filtering by specific key terms, minimum topic weight (relative frequency), and non-zero modularity. The latter effectively filters out marginal topics that are not clearly demarcated from neighboring graph communities, thus focusing the information display on well-structured topics. Likewise, the detail views (C–E) can be filtered to display statistics and content associated with the primary, secondary, or both topics combined. This lets the analyst investigate what partial and/or overlapping aspects are covered by a pair of potentially related topics.

For the iterative-interactive refinement of detected topics, it is possible to (i) merge two correlated topics, (ii) remove marginal or nonsensical topics entirely (e. g., comprised of non-English terms), (iii) remove a single irrelevant or nonsensical term from a given topic, and (iv) remove a specific microblog as non-related to a given topic (e. g., with particularly "creative" content). Figure 2 shows the structural effects of these operations for a hypothetical topic community graph.

Every interaction immediately updates the topic graph. This includes updates to the absolute and rela-

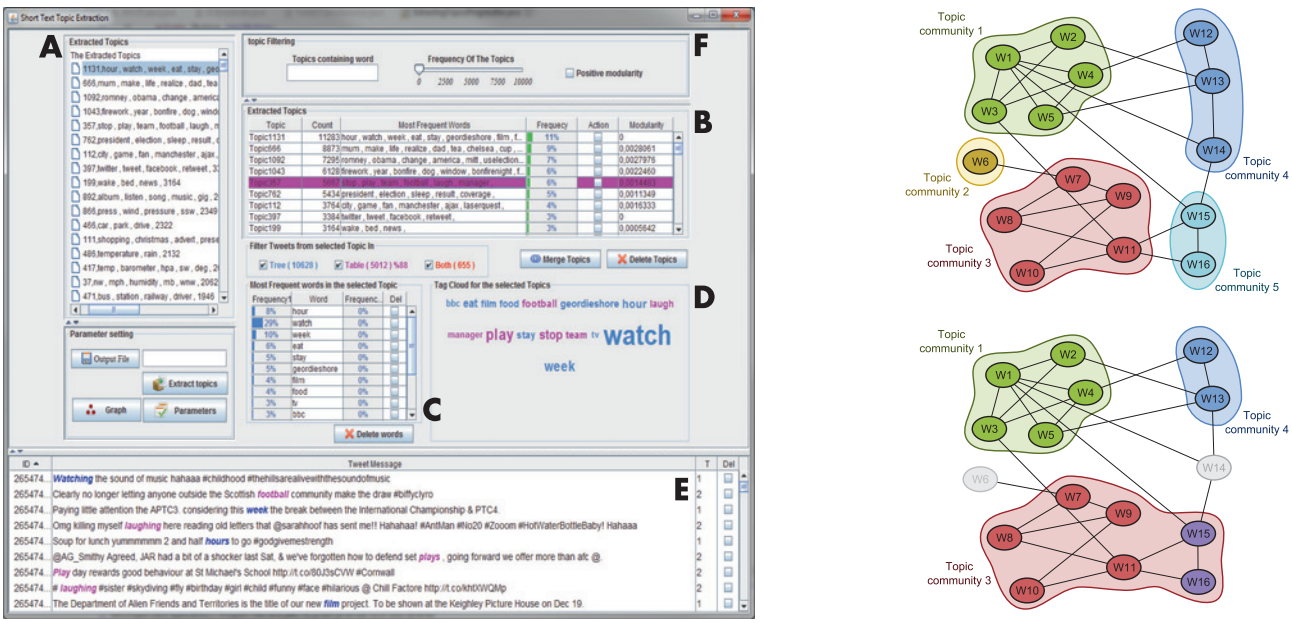

Figure 2: The visual-interactive topic graph refinement tool (left); effects of principal topic community graph refinement operations term removal (W14), topic removal (topic 2 with single word W6), and topic merging (topics 3 + 5) (right).

Figure 3: Example: Sociograph for the hair styling market (left); full topic graph (middle); manually-refined topic graph on product related tweets (right).

tive term frequencies, associated community modularity – which may transitively affect connected topics – and of course the list of topics itself. Using these operations, the human analyst can iteratively refine the model until it retains all topics with their relevant terms for the given application domain and/or the analysis task at hand.

The GCD algorithm depends on a few parameters influencing the type and accuracy of the results. This concerns the minimum and maximum size (node count) and frequency (number of associated tweets) of detected topic communities controling the balance between many marginal and few generic "super topics", as well as the minimum node and edge weights (prune low-frequency terms and spurious term co-occurrences, respectively).

In addition, our approach uses the notion of *core communities* representing the most promising topics on which visual inspection and refinement should concentrate. Such communities are denoted by an above-average absolute size (number of terms) and frequency (number of associated tweets). Adjusting this parameter allows the analyst to make a tradeoff between accuracy of the extracted initial topics vs. the flexibility afforded by (but also time required for) the interactive exploration and refinement of the topic graph.

As is often the case, finding appropriate parameter settings needs to be done in a heuristic way. For the corpus of the use case described in the following section we determined the following settings to work well. Minimum node, edge, and topic frequencies should be set in proportion to the corpus size; a value between 0.1% and 1% of the total tweet count yielded a good balance between overall topic graph size and topic detection sensitivity. The minimum size of core communities is not strongly related to the size of the corpus. We achieved good results with size thresholds between 5 and 10.

5 Application example: Product perception on twitter

Together with technology and business experts in the hair and beauty industry we engaged in a project seeking information on brand recognition, perceived product qualities, and market response to products and technologies in order to improve business processes, individualize services, and ultimately gain competitive advantages. The project's analytical tasks were threefold:

1. Who is talking with whom and what roles exist?
2. What are the target groups and influencers talking about topics, experiences, opinions?
3. How is the social response linked to products, brands or markets?

Initially tackling these points using LDA and deep neural networks [17] we found it difficult to seamlessly include the business domain exerts in the analysis process due to the issue discussed in Section 3. In particular, the "black box" nature of LDA makes topic models hardly traceable by the non-text mining expert, with correspondingly low confidence in them. Likewise, the sensitivity to what appeared to the business experts as minor changes caused confusion and made the model hard to interpret for them – they tended to build topic-related stories around the terms identified by LDA and would be thrown off by the sometimes significant changes in topic clusters.

Therefore, we applied the proposed interactive approach of semi-supervised topic extraction not only for analysis of the final corpus, but already in an iterative data acquisition process: (i) Set up and iteratively calibrate the "social listeners" (data gathering processes) to the specifics of the targeted groups and content to expand and refine the text corpus. This would take into account

insights gained from the previous analysis step. (ii) Automated structuring and information retrieval from the corpus data collected so far using GCD and/or LDA, as chosen by the analyst. (iii) Interpretation and refinement of the initial model using the interactive tool, derive adjustments to the data acquisition process.

We collected more than 35 million tweets in May and June 2014 using the public Twitter API with a continuously refined set of query terms like "hair", "styling", user and product names, as well as other named entities such as key community members.

In building the term co-occurrence graph **TG** we included user names and hashtags treated as compound nouns. Additionally, we tagged known brand and product names of interest as named entities. The Sociograph **SG** was created based on direct communication links implied by @-tags designating user names, including official brand accounts (see Figure 3, left). We integrated **TG** and **SG** by exploiting co-occurrence of hashtags and terms in a given user's messages, thus linking users, terms, entities, and hashtags. We then applied GCD to this integrated graph to extract an initial set of topics based on content (opinions, experiences, ...), hashtags, and involved users (hairstylists, media, testimonials, ...), see Figure 3, middle.

Using topic filtering to eliminate social noise and topics of less interest we reduced the set of 35 million to ~ 247 000 relevant tweets. This reduced set was analyzed for one particular dyeing product. First we looked at topics and terms co-occurring with that product, i. e., linked by edges to the product name node in **TG**.

We discovered that the product of interest frequently co-occurs with topics related to professional training. Some tweets mentioned experiences from taking a class, learning success, or anticipation of class participation. For product development and market building this topic context was regarded as highly relevant.

We then used visual-interactive topic model refinement to join corresponding term communities to form a *super topic* (topic cluster) on *product education*. Matching these findings with the Sociograph we observed several key users involved in professional training linking to various aspects of this super topic. A domain expert provided us with background knowledge on trainers and training locations (entities). This allowed us to establish links and identify further topics and user communities that were only indirectly linked to training but did not directly mention key phrases like "training", "education" etc. in their message.

This resulted in the final topic map of Figure 3, right comprising ~ 10 000 tweets describing many facets of product-related training. In particular, this focused topic map facilitated novel insights on training-related product perception, market establishment, as well as identification of dissemination opportunities for said training curriculum.

6 Conclusion

Social media have become an almost ubiquitous and important source of information on a wide variety of topics. A sizable body of work deals with automated text and graph mining approaches on social data, however analysis of very short text such as micro postings remains very challenging due high noise and missing context. The integration of application domain experts into the analysis and interpretation phase of the topic detection process is a promising way to cope with data gaps (e. g. missing links), and to include contextual and implicit knowledge. This paper proposed an approach that combines graph community detection-based extraction of initial topics with visual-interactive refinement of the topic model. It also allows integration of insight that could further be gained from the sociograph underlying the topics and communications apparent from the input corpus. Its applicability has been tested in a number of projects from different domains including hair and beauty, consumer electronics and disaster management.

Future work will include review of other GCD algorithms, in particular, non-disjunctive graph partitioning such as [10] to more efficiently handle homonyms (i. e., terms or hashtags appearing in several topics with different semantic meaning); as well as means to capture and analyze temporal dynamics of topics and topic evolution in social media streams (e. g., [11]). A planned extension to the interactive refinement strategy is the option to manually define negative edge weights i. e., topic anti-terms that must *not* appear in a micro posting if associated with a given topic. Finally, a train of work will concentrate on the integration of more advanced NLP methods e. g., calculation of *embeddings* defining semantic phrases [4].

Funding: This work has been partially supported by funding from European research project FP7-ICT-2011-8 INSIGHT, http://www.insight-ict.eu/.

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.
2. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):10008–10019, 2008.
3. J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. Ebert, and T. Ertl. Spatiotemporal Social Media Analytics for Abnormal Event Detection using Seasonal-Trend Decomposition. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, 2012.
4. Y. Chen, B. Perozzi, R. Al-Rfou, and S. Skiena. The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226*, 2013.
5. A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski. #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, 17:124–147, 2013.
6. N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 115–122, 2010.
7. M. Dörk, D. Grün, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(4):1129–1138, 2010.
8. W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. Zhou. LeadLine: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 93–102, 2012.
9. K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of 49th Annual Meeting of the ACL: Human Language Technologies (HLT'11)*, pages 42–47, 2011.
10. S. Gregory. A Fast Algorithm to Find Overlapping Communities in Networks. In *Machine Learning and Knowledge Discovery in Databases*, LNCS 5211, pages 408–423, 2008.
11. T. Kraft, D. X. Wang, J. Delawder, W. Dou, L. Yu, and W. Ribarsky. Less after-the-fact: Investigative visual analysis of events from streaming twitter. In *Large-Scale Data Analysis and Visualization (LDAV), 2013 IEEE Symposium on*, pages 95–103. IEEE, 2013.
12. X. Liang, W. Chen, and J. Bu. Bursty Feature Based Topic Detection and Summarization. In *Proceedings of 2nd International Conference on Computer Engineering and Technology (ICCET)*, volume 6, 2010.
13. X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of 49th Annual Meeting of the ACL: Human Language Technologies (HLT'11)*, pages 359–367, 2011.
14. M. Mathioudakis and N. Koudas. TwitterMonitor: Trend Detection over the Twitter Stream. In *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, pages 1155–1158, 2010.
15. J. D. McAuliffe and D. M. Blei. Supervised Topic Models. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, number 21, pages 121–128, 2008.
16. A. K. McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.
17. G. Paass and B. Pratap. Semantic Role Labeling Using Deep Neural Networks. In *The International Workshop on Representation Learning* (RL2014), 2014.
18. F. Piller, A. Vossen, and C. Ihl. From social media to social product development: the impact of social media on co-creation of innovation. *Unternehmung*, 66(1), 2012.
19. D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP '09), pages 248–256, 2009.
20. W. Ribarsky, X. Wang, and W. Dou. Social Media Analytics for Competitive Advantage. Invited paper, *Computers & Graphics*, Volume 38C (Special Issue on EuroVA 2013), pages 328–331, 2013.
21. M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. Learning Author-topic Models from Text Corpora. *ACM Transactions on Information Systems*, 28(1):1–38, Jan 2010.
22. H. Saif, Y. He, and H. Alani. Alleviating Data Sparsity for Twitter Sentiment Analysis. In *Proceedings of Making Sense of Microposts (MSM2012)*, 2012.
23. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web*, pages 851–860. ACM, 2010.
24. M. Timonen, P. Silvonen, and M. Kasari. Classification of short documents to categorize consumer opinions. In *Online proceedings ADMA'11*, 2011.
25. X. Wang, W. Dou, Z. Ma, J. Villalobos, Y. Chen, T. Kraft, and W. Ribarsky. I-SI: Scalable Architecture for Analyzing Latent Topical-Level Information from Social Media Data. In *Computer Graphics Forum*, Volume 31, pages 1275–1284, 2012.
26. W. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *33rd European conf. on Advances in Information Retrieval*, pages 338–349, 2011.

Bionotes

Dr. Georg Fuchs
Fraunhofer IAIS, D-53757 Sankt Augustin
georg.fuchs@iais.fraunhofer.de

Georg Fuchs is a senior research scientist and project manager at Fraunhofer IAIS working in the field of visual analytics with a strong emphasis on spatio-temporal data analysis. His research interests include information visualization in general and visualization of spatio-temporal data in particular, visual analytics methodologies, task-driven adaptation of visual representations and Smart

Visual Interfaces, as well as computer graphics and rendering. Georg Fuchs has co-authored 38+ peer-reviewed research papers and journal articles, and received a best short paper award at Smart Graphics 2008.

Hendrik Stange
Fraunhofer IAIS, D-53757 Sankt Augustin
hendrik.stange@iais.fraunhofer.de

Hendrik Stange is a research fellow at the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS and a project manager of many research and industrial big data projects with international consortia. He has a background in data mining and spatial business intelligence. Hendrik studied at the Otto-von-Guericke University Magdeburg and received his degree in Business Informatics in 2007. Since, he specialized in mobile analytics for outdoor advertising and telecommunications. Current research focuses on learning on streams of heterogeneous poly-structured data and visual analytics for big data applications.

Ahamd Samiei
Fraunhofer IAIS, D-53757 Sankt Augustin
asppagh@gmail.com

Ahamd Samiei is PhD student at Fraunhofer IAIS. He recently obtained is MSc. in Computer Sciences on the topic of semi-supervised topic extraction from Twitter. His research interests include natural language processing, text mining, linked data and data mining in general.

Dr. Gennady Andrienko
Fraunhofer IAIS, D-53757 Sankt Augustin
gennady.andrienko@iais.fraunhofer.de

Gennady Andrienko is a lead scientist responsible for the visual analytics research at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) and professor (part-time) at City University London. He co-authored the monographs “Exploratory Analysis of Spatial and Temporal Data” (Springer, 2006) and “Visual Analytics of Movement” (2013), 60+ peer-reviewed journal papers, 20+ book chapters and more than 100 conference papers. Since 2007, Gennady Andrienko is chairing the ICA Commission on GeoVisualization. He co-organized scientific events on visual analytics, geovisualization and visual data mining, and co-edited 10 special issues of journals.

Dr. Natalia Andrienko
Fraunhofer IAIS, D-53757 Sankt Augustin
natalia.andrienko@iais.fraunhofer.de

Natalia Andrienko has been working at GMD, now Fraunhofer IAIS, since 1997. Since 2007, she is a lead scientist responsible for the visual analytics research. Since 2013 she is professor (part-time) at City University London. She co-authored the monographs “Exploratory Analysis of Spatial and Temporal Data” (Springer, 2006) and “Visual Analytics of Movement” (2013), over 60 peer-reviewed journal papers, over 20 book chapters and more than 100 conference papers. She received best paper awards at AGILE 2006 and IEEE VAST 2011 and 2012 conferences, best poster awards at AGILE 2007 and ACM GIS 2011, and VAST challenge award 2008.