8-6-2011

Analyzing customer sentiments in microblogs – A topic-model-based approach for Twitter datasets

Stefan Sommer
T-Systems Multimedia Solutions GmbH, s.sommer@t-systems.com

Andreas Schieber
Dresden University of Technology, andreas.schieber@tu-dresden.de

Andreas Hilbert
Dresden University of Technology, andreas.hilbert@tu-dresden.de

Kai Heinrich
Dresden University of Technology, kai.heinrich@mailbox.tu-dresden.de

Analyzing customer sentiments in microblogs – A topic-model-based approach for Twitter datasets

Andreas Schieber
Dresden University of Technology
andreas.schieber@tu-dresden.de

Stefan Sommer
T-Systems Multimedia Solutions GmbH
s.sommer@t-systems.com

Andreas Hilbert
Dresden University of Technology
andreas.hilbert@tu-dresden.de

Kai Heinrich
Dresden University of Technology
kai.heinrich@mailbox.tu-dresden.de

ABSTRACT

In the Social Commerce customers evolve to an important information source for companies. The customers use communication platforms of the Web 2.0, for example Twitter, in order to express their opinions about products or discuss their experiences with them. These opinions can be very important for the development of products or the product range of a company. Our approach enables a company viewing opinions about its products which are published using the microblogging service Twitter. A first step in our research progress is detecting topics in a specific context. In a further step the entries corresponding to these topics has to be analyzed for opinions. For topic detection we use topic modeling with the Latent Dirichlet Allocation. In our paper we found event-based topics in the context of Sony's 3D TV sets. In future work we are able to implement Opinion Mining algorithms to determine sentiments in the entries corresponding to the detected topics.

Keywords

Social Commerce, Twitter, LDA, Topic Models, Knowledge Discovery, Opinion Mining

INTRODUCTION

"What am I doing?" best describes the basic idea of Twitter. Using the social network platform Twitter, people share news or opinions in short messages. Twitter is a so called microblog, which is a special kind of a blog that combines an ordinary blog with features of social networks. Twitter is the most popular microblogging application with more than 1.8 million net users in Germany in 2009 (Pettey and Stevens, 2009). Due to the positive developments of microblogs and Twitter in particular, these services become a valuable source for companies (Pak and Paroubek, 2010) (Barnes and Böhringer, 2009).

"What is the conversation about?" should be the question for companies. Today customers are considered as key communication partner, providing valuable feedback, requests and testimonials of a company's performance (Richter, Koch and Krisch, 2007). They share their sentiments with other customers through the communication platforms of the Web 2.0. By spreading their thoughts through these platforms, such as blogs, communities or social networks, customers influence the process of opinion making process of other customers (O'Connor et al., 2010). This phenomenon is defined by (Richter, Koch and Krisch, 2007) as the Social Commerce. They argue that Social Commerce is the evolved development of the electronic commerce. It changes the communication and interaction between companies and customers. In particular the relations and exchange of information between customers becomes more important. Companies of the Social Commerce need to know, how they perform in the customers perspective. They can use this information e.g. to optimize their product and service mix or involve customers in the product development as so called prosumers. Nevertheless the communications platforms of the Web 2.0 gain in importance, as the interaction between customers increases through this media (Stephen and Toubia, 2010).

We aim on mining customer opinions in microblogs. Because of the enormous number of entries it is very difficult to filter relevant content without using automated procedures. Opinion Mining offers automated analysis of text content and provides the classification of proved entries e.g. in positive, neutral, and negative ones (Liu, 2007). Before Opinion Mining a sub-ordinate target is to identify entries which contain expressions within a relevant context. For example, a product manager of Sony especially is interested in statements about Sony products (the example of Sony will be used continuously).

METHOLODGY AND RELATED WORK

We gain new insights by using the design science approach by (Hevner et al., 2004). The purpose of Hevner's approach is the development of an artifact which solves a specific problem. In this case the specific problem is the identification of microblog entries containing opinions or testimonials in a particular context. In our implementation we want to find entries about Sony products in the context of 3D technology as an example. During the paper we answer the following research questions:

1. What capabilities and challenges occur while analyzing the entries of microblogs because of the limited amount of characters?
2. How can we automatically identify the topics of the entries?

We use generative topic models which allow us automatically to find topics in a textual dataset. As data source we choose the microblogging service Twitter because of two reasons: First, Twitter is the most popular microblogging platform with a great amount of users as mentioned before. Second, most of the entries published with Twitter are free to read (Pak and Paroubek, 2010). After that we discuss the characteristics of microblogs and how we can use the process of Knowledge-Discovery-in-Databases (KDD) by (Fayyad, 1996) for our intention. The last section shows the evaluation of our approach by exemplarily determining topics in a Twitter datasets.

(Böhringer and Gluchowski, 2009) describe the microblogging service Twitter and how users can communicate with each other using this Web 2.0 platform. The entries in Twitter, called tweets, contain different content, for example opinions or testimonials. It can be interesting to analyze this content in order to get useful insights in customers' sentiments. (Oulasvirta et al., 2010) and (Tumasjan et al., 2010) show what insights this might be: (Oulasvirta et al., 2010) explain common findings such as the characteristics of users' self-disclosures; in contrast, (Tumasjan et al., 2010) use Twitter in a more specific case in order to reveal the policital sentiment of tweet authors.

For our purpose we use generative topic models. (Blei and Lafferty, 2009) give a fundamental explanation of topic models and their usage. We use the Latent-Dirichlet-Allocation (LDA) in our work which is first mentioned by (Blei, Ng and Jordan, 2003). Since its first publication other authors have successfully used the algorithm in order to identify topics. (Ramage, Dumais and Liebling, 2010) also used topic models to analyze tweets.

TOPIC MODELS IN MICROBLOGS

Capabilities and challenges of microblogs

For analyzing microblogs we have to face some special issues which characterize these short notes. (Böhringer and Gluchowski, 2009) introduce the microblogging service Twitter and its functionalities: First, Twitter users can communicate with each other by referencing the name of the communication partner with a prefixed "@". For example, userA writes an entry containing "@userB" in order to address userB. In addition, users can distribute an entry of another user by forwarding this entry with the prefixed characters "RT". For example, in order to forward the origin entry "tweet" of userA userB publishes the tweet with "RT @userA tweet". In this way, the range of an expression is increased, which ultimately will benefit the reputation of the origin author. Finally, there is a very important function of microblogs: the tagging of entries. Authors can tag their entries by adding keywords to the tweet. These keywords are called hashtags and can be recognized by the prefixed "#". In summary, the technical functionalities of Twitter provide several possibilities for analysis. Possible research areas are Social Network Analysis, Web Content Mining and Consumer Behavior Analysis.

The character limitation of 140 characters leads to the main challenge in analysis of microblogs. In order to write as much information as possible users tend to use abbreviations (for example, "4ever" is used as an abbreviation for "forever"). In addition, the informal way of speaking in microblogs and syntactic errors complicate the mining procedure. In contrast, (Bermingham and Smeaton, 2010) see the short length as strength of microblogs because the limited text can contain compact and explicit sentiments. In their paper they found classifying sentiments in microblogs easier than in blogs. Though, the brevity can also be an advantage.

As mentioned before several researchers have successfully analyzed entries of microblogs. Twitter is a popular data source because of the great number of text posts and the heterogeneous audience (Pak and Paroubek, 2010). In addition, Twitter users express a certain sentiment about different topics, so it is possible to analyze people's opinions on those topics. In this context (Pak and Paroubek, 2010) show examples of entries with expressed opinions in order to demonstrate the high potential of sentiment analysis in tweets. (Tumasjan et al., 2010) used 100.000 tweets in order to reveal the political sentiment in Germany. They found that the majority of the analyzed tweets reflects voter preferences and even come close to

traditional election polls. Users not only express their opinions but also discuss them with other users. The current research shows the interesting and useful insights by analyzing microblogs.

Process of knowledge discovery in Twitter datasets

The process of Knowledge-Discovery-in-Databases (KDD) has been published by (Fayyad, 1996). This is the traditional way of analyzing data and it contains five steps. We adopted and modified this process for discovering knowledge in Twitter datasets as shown in Figure 1.

The first step is the selection of the target data. Here we use the Twitter search to select our target data out of the complete raw Twitter data. We access the Twitter search functionality by using the Twitter API. We send search strings containing our keywords (e.g. Sony 3D) to Twitter and get back corresponding tweets which were tweeted within the last two weeks. We decided to use the time-limited Twitter search stream at this stage because the amount of tweets is sufficient to show what insights can be gained by applying generative topic models. For our future work we plan to build a crawler which collects each tweet in order to gather a corpus with more entries.

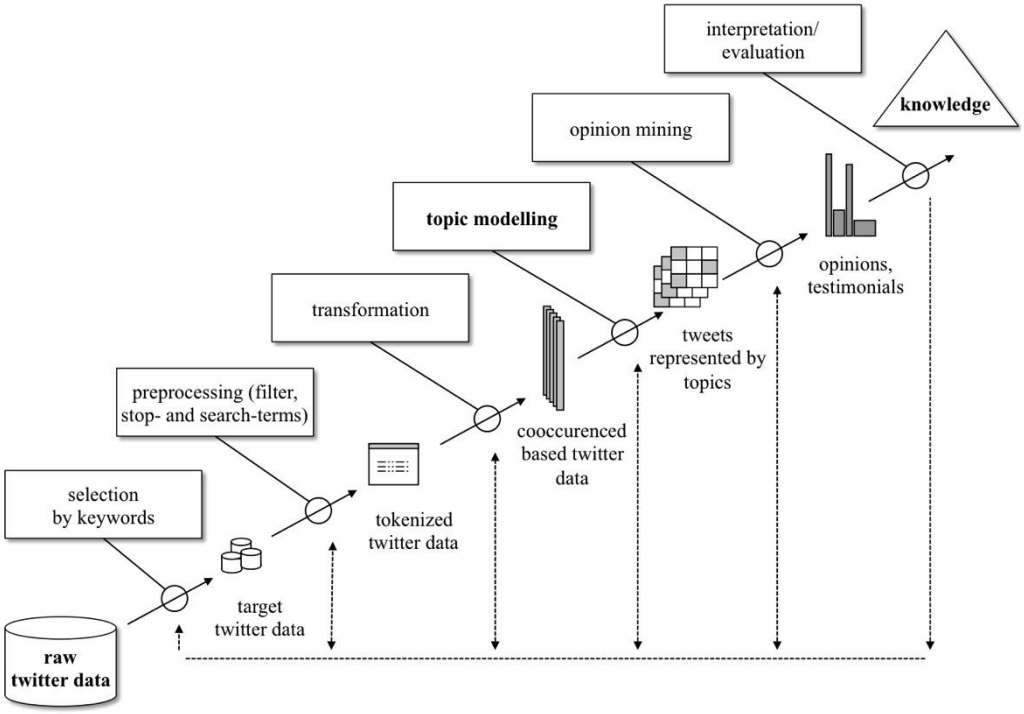

Figure 1: The process of knowledge discovery in Twitter datasets

After the selection of our target data we have to perform some preprocessing tasks. Our data source contains only the textual content from the Twitter entries, so we have one variable. In order to obtain useful results we remove some parts from each tweet. This includes stop words, the keywords of our search strings, single characters or numbers, and cross-references to other users (e.g. @userA). Afterwards we transform our corpus by lexicalizing it and yield co-occurrence based data in order to perform the Latent Dirichlet Allocation (LDA). The next step contains the topic modeling of the tweets. Therefore we implement the LDA algorithm by (Blei, Ng and Jordan, 2003) in order to identify topic clusters in our Twitter data.

We also want to ensure to not only capture one snapshot of the twitter blogosphere and therefore we iterate the process with 2 different Twitter data selections from different time periods. This will help us to gain an insight into the stability of clusters and the nature of topic transition in tweets. Using this whole analysis as a pre-selection mechanism for an opinion mining approach later on the iteration step is important for grabbing important conversation topics about products or technologies over time within a given context. Further on this helps us to understand the impact of the LDA approach as a method for representing microblog entries. It is known from (Blei and Lafferty, 2009) that the non-deterministic LDA probability model is very sensitive to the configuration of its initial parameter values, which could lead to very different results on the same datasets. As a result we may get very instable representations of blog conversations over time. This fact can only be taken into account with the application of the model at different points in time. However we expect of course a change in topics as

Twitter is of course influenced by daily news and recent topics around the world. Applying the LDA model over time may help us to further filter unwanted data noise like advertising.

A short overview about topic models and LDA is shown in the following section. The results of our analysis will be described in section after that.

Topic Models and the Latent Dirichlet Allocation

(Blei and Lafferty, 2009) introduced Topic Models as a powerful technique for finding useful structure in an otherwise unstructured collection. Topic Models are probabilistic models for unsupervised uncovering the underlying semantic structure of a document collection (e.g. Twitter). (Blei and Lafferty, 2009) used this technique for automatically managing the contents of the digital archive of the journal Science. The text documents are distilled into distributions of words that tend to co-occur in similar documents. These sets of words are summarized into topics. In order to generate the topic model of our corpus we used the LDA which is the most common topic model currently in use. Developed by (Blei, Ng and Jordan, 2003) it is a generalization of probabilistic latent semantic indexing (PLSI) allowing documents to have a mixture of topics. A disadvantage of the method is that it fails to model correlations between the occurrences of topics. However, in our case we do not assume that there is any correlation between the latent topics. This assumption however does not necessarily hold but for the purpose of exploration, which is our main goal here, we assume an absence of high correlation between topic clusters. The LDA model is therefore suitable for our approach on topic exploration.

EXPLORING TOPICS IN TWITTER DATASETS

Twitter datasets and the analytical approach

We apply our models on data that was collected by using the Twitter search API. We parameterize the search with three different queries corresponding to different levels of topic granularity. Referring to our Sony example we choose the terms "3D", "Sony 3D" and "Sony 3D KDL" to ensure we capture the context on the 3D technology in general as well as topics related to specific products. As mentioned above we conducted our analysis at two different times which doubles the sets of data we use for our modeling approach. The first corpus collection contains about 1500 Tweets which were gathered over two weeks from the 16th through the 30th of November 2010. The second corpus collection contains about 1200 Tweets which also were gathered over two weeks from the 8th through the 22th of January 2011. The data is tokenized in such a way that the special content elements of a Twitter posts like hashtags, user names and URLs are kept together. As another important step in preprocessing we removed the search terms and duplicate posts. In order to achieve our topic overview we implemented the LDA model using Gibbs-Sampling algorithm (Ramage, Dumais and Liebling, 2010). Since the number of posts we collected is relatively small, the sequential nature of Gibbs sampling poses no problem. Our LDA model contains 15 dimensions for each search stream. This decision is based on our main goal, which is exploring twitter datasets rather than vectorizing them for model operations, in which case a higher number of topics would be more appropriate to distinguish between posts.

Results

The results of our research contain distributions over topics as well as the mixture of topics occurring in single documents. Figure 2 shows the overall topic distribution for the "Sony 3D" corpus that was collected in 2010. The Top 8 Words which are most likely to characterize the most frequent topic (X7) are also presented in figure 2.

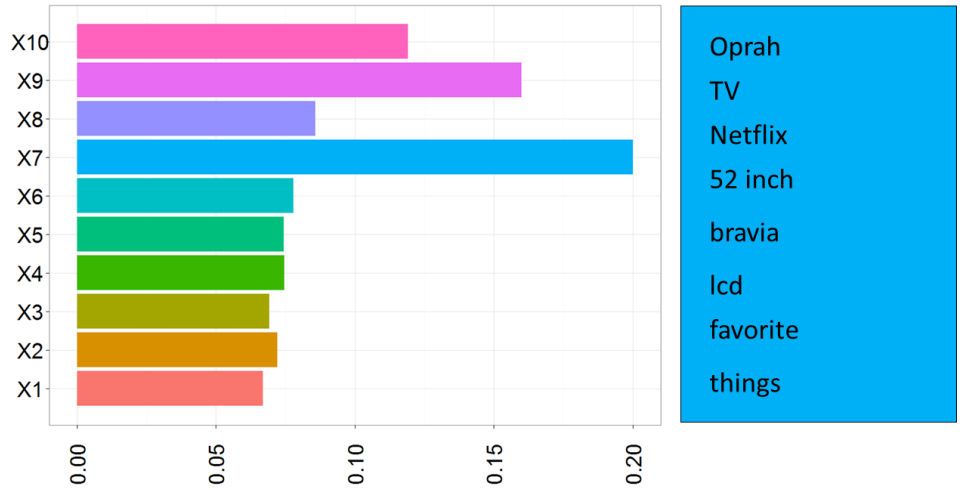

Figure 2: Topic distribution and Top-Word characterization over Twitter posts containing the term "Sony 3D" in the 2010 corpus collection

By looking at the results we state that the topic distribution over all documents becomes more specific and probability mass is likely to be distributed to single topics instead of being equally as level of granularity in the search context changes. We also observe a near uniform distribution in the "3D" case whereas the "Sony 3D" and the "Sony 3D KDL" datasets show highly skewed distributions which indicates that in those cases the LDA model was heavily weighting certain topics with a high proportion. This result is partly explained by the short-term nature of our datasets. In a certain time interval it is more likely for a few topics like recent news or events to be commented on in social networks like Twitter. In our "Sony 3D" example we can clearly distinguish the event based posts from other tweets like advertisement posts. An example for this is Oprah naming a Sony 3D TV one of her favorite things (which is also the number one topic in our model). The following examples in figure 3 contain different opinions that could be useful to achieve our future goal of analyzing customer opinions in the Social Commerce.

Lauxa_VFcgcrl: @rhtniyk alikifoley Lol. She needs to STOP!!- RT LeonThomasIII: 52 inch Sony Bravia 3D tv Oprah is the shyt http://ftmd.dhm.ro/TdMJr

2OneQuestions: You better call oprah. RT @JamieFoxxy: I need that new Sony 3d52' tv.

freeestufff: iPad tops Netflix, Sony 3D for Oprah's 'Favorite Things' | How iLiving: Describing it as her "number one favorit... http://bit.ly/fWx4tV (expand)

kamzou08: iPad tops Netflix, Sony 3D for Oprah's 'Favorite Things' | How iLiving: iPad tops Netflix, Sony 3D for Oprah's '... http://bit.ly/gAG6tP (expand)

GossipToday98: #NateBerkus, Did Oprah Hype 3D TV to Help Sony, Discovery? -http://ow.ly/1rKKLx

Figure 3: Sample Twitter documents identified to be highly connected to the Oprah Topic from the "Sony 3D"Data.

We also recognize that due to the limitation of words in a Twitter post the topic mixture in a single document is more likely to only hold one topic with a high proportion or even with a proportion of 1. Figure 4 shows such a distribution, representing five random documents from the "Sony 3D" corpus. The figure illustrates that in every case a high proportion is assigned to one of the ten possible topics. This fact is important because using a model for Twitter datasets which is not likely to identify tweets as one-topic documents, or at least having one topic dominate the others in the matter of proportion, would not be suitable for analyzing Twitter datasets. Having inspected that snapshot of the topic landscape from 2010 we were also interested in cluster behavior over time. To analyze this behavior we use the second corpus collection from early 2011 and then compare the two results.

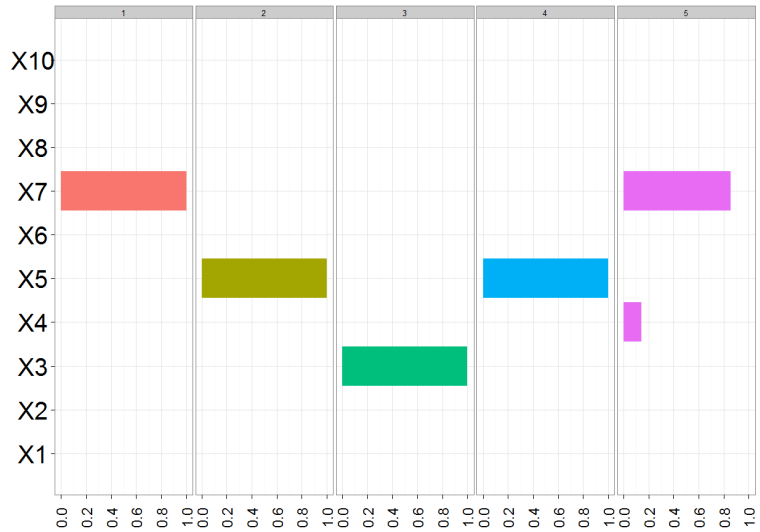

Figure 4: Topic distribution of five random Twitter posts containing the term "Sony 3D"

Figure 5 shows the topic distribution for the "Sony 3D" keywords in the 2011 corpus in the same fashion we presented it in figure 2. The most frequent cluster that appears is X9. It is described by the keywords which are most likely to represent that topic cluster.

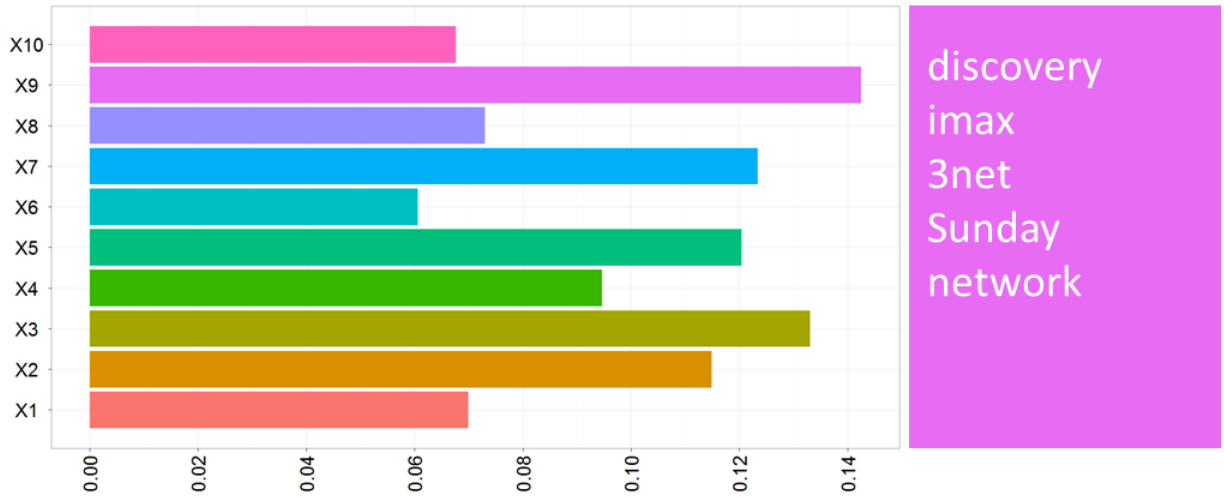

Figure 5: Topic distribution and Top-Word characterization over Twitter posts containing the term "Sony 3D" in the 2011 corpus collection

The findings about short-term nature and the big influence of reason events that reflect on our keyword topic, like the launch of the 3D network 3net by Sony, Discovery and IMAX, can be confirmed within the second corpus collection.

CONCLUSION

The customer communication via Web 2.0 technologies is an important evolutionary step in the process of opinion making. In particular, microblogs offer capabilities which allow powerful analysis in the field of Opinion Mining. Knowing and using the detected opinions is the key to understand the customer and his insights on events and other topics which evolve around a certain product or technology. This knowledge can be used e.g. to improve the products or product ranges of a company.

With our approach we are able to identify posts corresponding to such important topics, that is, topics with a high probability of occurrence. Using LDA we can distinguish between posts which are useful for exploring customer opinions and those

which hold less useful information. Although our results suggest that there is still a massive amount of work to be done in the area of content recognition in microblogs to ensure a distinction between useful and noisy data. It is also necessary to detect long term effects which includes a large-scale evaluation with data streams that are collected over an appropriate time period. We gained new views on social network which is a first step in order to know what the conversation really is about.

For our future work we are able to filter tweets that are likely to be relevant in our context. In the example the Sony product manager is able to view entries with interesting topics that reflect on Sony 3D TV sets. The next step is to expand the analysis by implementing suitable algorithms for Opinion Mining in order to enable the product manager to analyze the underlying sentiments of the pre-selected posts. In addition a Twitter crawler has to be developed in order to automatically collect and preselect tweets over a broader period to analyze opinions in the long run.

REFERENCES

1. Barnes, S.J. and Böhringer, M. (2009) 'Continuance Usage Intention in Microblogging Services: The Case of Twitter', Proceedings of the 17th European Conference on Information Systems, 1-13.
2. Bermingham, A. and Smeaton, A. (2010) 'Classifying Sentiment in Microblogs - Is Brevity an Advantage?', Proceedings of the 19th ACM international conference on Information and knowledge management, 1833-1836.
3. Blei, D. and Lafferty, J. (2009) Topic Models, [Online], Available: http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf [30 Nov 2010].
4. Blei, D., Ng, A. and Jordan, M. (2003) 'Latent Dirichlet Allocation', Journal of Machine Learning Research, pp. 933-1022.
5. Böhringer, M. and Gluchowski, P. (2009) 'Microblogging', Informatik-Spektrum, pp. 505-510.
6. Fayyad, U. (1996) Advances in Knowledge Discovery and Data Mining, Menlo Park: AAAI Press.
7. Hevner, A., March, S., Park, J. and Ram, S. (2004) 'Design Science in Information Systems', MIS Quarterly, 28, pp. 75-105.
8. Liu, B. (2007) Web Data Mining, Berlin: Springer.
9. O'Connor, B., Balasubramanyan, R., Routledge, B. and Smith, N. (2010) 'From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series', Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 122-129.
10. Oulasvirta, A., Lehtonen, E., Kurvinen, E. and Raento, M. (2010) 'Making the ordinary visible in microblogs', Personal and ubiquitous computing, Vol. 14 (3), pp. 237-249.
11. Pak, A. and Paroubek, P. (2010) 'Twitter as a Corpus for Sentiment Analysis and Opinion Mining', Proceedings of the International Conference on Language Resources and Evaluation, 1320-1326.
12. Pettey, C. and Stevens, H. (2009) Gartner's Hype Cycle Special Report for 2009, [Online], Available: http://www.gartner.com/it/page.jsp?id=1124212 [7 Dec 2010].
13. Ramage, D., Dumais, S. and Liebling, D. (2010) 'Characterizing Microblogs with Topic Models', Fourth International AAAI Conference on Weblogs and Social Media.
14. Richter, A., Koch, M. and Krisch, J. (2007) 'Social Commerce - Eine Analyse des Wandels im E-Commerce', Bericht 2007/03, Fakultät Informaitk, Universität der Bundeswehr München.
15. Stephen, A.T. and Toubia, O. (2010) 'Deriving Value from Social Commerce Networks', Journal of Marketing Research, Nr. 2 Vol. 67, pp. 215-228.
16. Tumasjan, A., Sprenger, T., Sandner, P. and Welpe, I. (2010) 'Predicting Elections with Twitter - What 140 Characters Reveal about Political Sentiment', Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 178-185.