

ForAVis - Explorative User Forum Analysis

Franz Wanner
University of Konstanz
78457 Konstanz, Germany
Franz.Wanner@uni-
konstanz.de

Thomas Ramm
University of Konstanz
78457 Konstanz, Germany
Thomas.Ramm@uni-
konstanz.de

Daniel A. Keim
University of Konstanz
78457 Konstanz, Germany
Daniel.Keim@uni-
konstanz.de

ABSTRACT

User generated textual content on the internet has become increasingly valuable during the past few years. Forums, blogs, twitter and other social media websites are accessible for a huge amount of people all over the world. Hence, methods and tools are needed to handle this vast bulk of textual data. In this paper we present an explorative forum analysis system helping various stakeholders to cope with the task analyzing user generated content in online forums. The used mobile communication forums picture an example of user generated content in online discussion forums.

Central to our system is a flexible visualization, which supports the analysis and exploration visually. Flexible means, that the ordering and the mapping of colors can be interactively changed by the analyst and the visualization is also capable to show the different structural levels of a user forum. The filter area offers beside well-known features many interesting features with respect to forum analysis, which we introduce in this paper. A detailed view of the particularly selected thread in the main visualization is presented in a third area. For a convenient manipulation and interaction we implemented intuitive mechanism. We describe the system and present various fictive user scenarios of different typical stakeholder tasks to illustrate the benefit of the system.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information filtering; H.4.3 [Information Systems]: Communications Applications—*bulletin boards*

General Terms

Application

Keywords

user generated content, social media analysis, forum analysis, forum visualization, visual analytics

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WIMS'11, May 25-27, 2011 Sogndal, Norway

Copyright © 2011 ACM 978-1-4503-0148-0/11/05... \$10.00

1. INTRODUCTION

The web is the largest information source in the world. Web 2.0 technology helps more and more people to actively contribute to this valuable information source by creating content in an easy way. There are many possibilities to take an active part in the web: forums, blogs, twitter, reviews and other ways to add comments.

A demonstrative example for the impact of social media was the death of Michael Jackson on June 25th, 2009. First the amount of twitter messages was doubled when the death became known to the masses. Afterwards the *Los Angeles Times* and *AP* authenticated the message. Another example is the video and the online rumor about picking Kryptonite bicycle locks. First there was a Youtube video and some discussions in bike and security forums. Kryptonite did not react to this contents, perhaps they did not even know anything about this issue. After short time customer complaints increased and Kryptonite had to admit that their locks were adversely affected by faulty design. After that, Kryptonite was forced to start a very expensive and expansive recall campaign.

As the examples show, one major aspect of the web is that enables people to meet others who share similar interests and exchange experiences. This happens amongst other social media in online forums. Exploration and searching in these forums is normally possible through keyword search or the author name (Figure 1). Options allow the user to narrow the results down and search for threads with more than a given number of replies or were posted within a specified time range. Other options allow the user to specify whether threads or posts are returned and the results list can usually be ordered in some way.

However, these search techniques are not particularly convenient for detecting relevant rumors and discussions. A relevant rumor is defined as a rumor, which could cause a negative effect on the reputation of a company. They do not allow searches on either a semantic level or based on some interesting features such as the sentiment of threads, post and authors. Effective and efficient monitoring of a forum is therefore not possible.

In this paper, we demonstrate a novel way of exploring and searching online forums using various features in combination with a visual representation. Section 2 describes related work and Section 3 shows the data, its structure, the

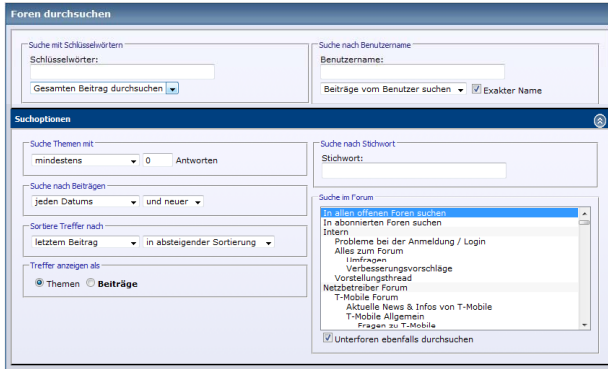


Figure 1: That is the extended search interface in a common online forum [6].

feature space and the user interface for forums analysis in detail. Section 4 illustrates the handling and advantages of our approach with use cases from both private and business domains.

2. RELATED WORK

Content Based Analysis

The analysis of user generated content has been done for many years, often using NLP and IR methods. Publications in this area are generally about summarizing discussions in forums [14], [38], [23] and to detect the conversation focus in threaded discussions [15]. An automatic scoring method which rates postings in online discussion forums, based on the value of their contribution has been demonstrated by [40]. Summarization has also been undertaken for blogs [48], microblogs [33] e.g. by sentence extraction [19].

A recommender system which includes relevance feedback based on brands, product categories and products discussed in shopping forums is shown in [12]. This work addresses the task of identifying product-related discussions in discussion forums. The result is a ranked list of relevant forums.

An example of detecting posts which do not belong to the topic (off topic detection) is [41]. They use discriminating terms to describe the topic of a thread. Novelty detection of content on sentence level was done in [47] and on document level in [27].

Assessing the quality of posts is very interesting in many tasks. A method on a textual base is showed in [8] and on non-textual features in [20]. [10] presents a domain specific comparison of frameworks measuring quality. [45] is an approach for predicting the quality of web forum posts domain independently. A framework for the credibility of posts is introduced in [44].

Another analysis task is hotspot detection and forecasting. Feature vectors used for this include the number of posts, the average number of answers in a discussion, the average sentiment polarity score of a post and the percentage of negative and positive posts [26]. An overview in the area of sentiment analysis in different domains is given in [30].

Analyzing the specific properties of emails, like for example thread-initial messages, to get improved archive overviews is shown in [29]. They also used visualizations developed for their purposes.

Visual Analysis

Narayan and Cheshire enrich the list with different visualizations [28]. They support the opportunity to visualize a forum with a modified histogram to see the activity over time or as a tree visualization to see the concatenation of posts in a forum. A third, based on work of Wattenberg and Millen [42] displays the discussion as sequence of rectangles. Important posts are highlighted with a color map to give the user feedback on positive or negative reader judgments.

[17] combine mining and interactive visualization techniques to analyse online discussions relating to consumer products. They tag messages with the topic they belong to, a relevance score, the polarity and more and apply analysis methods in an interactive way to these data.

Both, Turner et al. [39] and Engdahl et al. [13], illustrate forums using a treemap visualization. Engdahl's work here focused on visualization for a PDA device. [46] presents a possibility to structure posts within a thread using a treemap. [35] focuses on the social relations of a discussion. Different visualizations are used, such as a tree or a treemap.

For e-learning purposes, Giguët and Lucas [16] developed a system to support the tutor. This work aimed to analyse the posts concerning their point in time to see if there was collaborative work or not.

Lam and Donath [24] visualize the activity of discussions and users through moving objects. Discussions are rectangles which move along individually computed curves. The faster the movement the more active is the discussion, whereas the amplitude and frequency of the curve show the actuality.

DifVis [22] represents a thread as a square. The actuality and dimension is mapped on the position and the size, whereas a color coding shows activity, popularity and the duration of a thread.

Our visualization approach was inspired by Seesoft [11] and a visualization of Wikipedia edit sequences [43]. Seesoft is a tool for visualizing lines of code of large software projects; colours show when code was modified. The visualization for the Wikipedia edit sequences shows the history by means of a 'chromogram', a technique which Wattenberg et al. developed and is said to support finding patterns in the long sequences. The FilmFinder [9] has also a related search interface. The layout of ForAVis is inspired by [1]. An application for exploring the sentiment of blogs is shown in [25].

3. THE FORAVIS APPLICATION

3.1 Structure of Forums and Data

A conventional forum has a logical structure. On top it deals with a main topic, let's say mobile communication or automobiles. In most cases, beneath you will find some categories, also called sub-forums. Within mobile communi-

Netzbetreiber Forum	
	T-Mobile Forum (16 Betrachter) Themen die den deutschen Marktführer unter den Netzbetreibern betreffen T-Mobile Tarife T-Mobile News
	Vodafone Forum (19 Betrachter) Das Forum für den weltgrößten Mobilfunkbetreiber Vodafone Tarife Vodafone News
	E-Plus + Base Forum (16 Betrachter) Hier wird über Base und den Netzbetreiber E-Plus diskutiert E-Plus Tarife Base Tarife E-Plus News
	O2 Forum (37 Betrachter) Forum für den Netzbetreiber o2 O2 Tarife O2 News
	Provider der Netzbetreiber (12 Betrachter) Andere Provider wie Debitel- Mobilcom, The PhoneHouse, Sparhandy etc. Provider News Netzbetreiber Österreich & Schweiz
	Mobiles Internet / Mobile Datentarife (9 Betrachter) Fragen zu mobilen Datenflrates, Internet USB / Surfsticks Mobile Datentarife Datentarif News
	Prepaidkarten + Discount Anbieter (29 Betrachter) Alles rund um Prepaid Karten der Discounter Prepaid Tarife Prepaid News
	Mobilfunk/GSM (4 Betrachter) Alles, was in keines der anderen Foren zu GSM passt Mobilfunk News

Figure 2: Topics in the network operator sub-forum in a mobile communication forum [7].

cation forums sub-forums could be: one for internal issues regarding the forum, one for content which belongs to the network operators, one for topics about mobile phones, one for security themes and in most cases one sub-forum which deals with off-topic stuff. Please note that this enumeration is not complete at all. Every sub-forum has different topics as you can see in Figure 2. Each topic includes more or less threads and every thread consists of more or less posts. Posts are the smallest logical structure object in a forum and contain therefore the statement of an author.

We crawled nine different mobile communication forums (four are used in the screenshots later [2, 3, 4, 5]) to get data for ForAVis. Since we were only interested in discussions containing content about mobile communication providers, we filtered out all other discussions in a preprocessing step and stored the relevant discussions in an xml format. Thereby, we labeled quotations and used only the tagged descriptions of the emoticons. In a further step, these files were used to compute the sentiment scores as well as word and sentence counts. Finally, we stored the xml-data in a relational database having four tables: discussions, posts, authors and quotations. In the end, this resulted in about 5.000 discussions and approximately 40.000 posts.

3.2 Features

Table 1: Features for forum analysis

Author	Post	Thread
posts	length	posts
active days	word length	author diversity
duration	sentence length	starter activity
starter count	emoticons	hits
avg text length	shouting	last post
avg word length	thread count	duration
avg sentence length	question	active days
emoticons	response time	activity
shouting	sentiment	sentiment
sentiment		title
		provider

First a short description how we use the term *feature* in this work. A forum consists of many posts and threads. How often a thread was clicked by the users is considered as feature in this work. To capture the features of forums, we had to look on different structure levels. Hence, this is not enough for an author analysis, we additionally created a special feature set for authors which is accessible during the analysis on the post level. Table 3.2 shows an overview of our features. Through combination of different features during the analysis process, we empower the analyst in finding interesting posts, threads and authors. The user interface of the ForAVis-System is shown in 3.3.

3.2.1 Quantitative interestingness features

Author level

Posts shows the amount of posts from an author. It enables us to detect highly contributing authors in a forum with a lot of experience. *Starter count* shows how often an author starts a thread. This feature can be interesting if an analyst wants to extract people playing an active role and often initializing discussions in a forum. *Avg text length* is the average length of the postings of an author. We also considered *avg word length* and *sentence length*.

Post level

With regard to elaborateness the *post length* is important. Two quantitative linguistic features are the *word length* and the *sentence length*.

Discussion level

Beside post analysis it is also essential to make discussions measurable. *Posts* shows the amount of posts in a discussion. Another feature is *author diversity*. Here we calculate how many different authors contributed to a discussion compared to the total amount of posts in the thread. The more people participate in a discussion, the more general the topic of the thread seems to be. *Starter activity* reflects, if an author who starts the thread contributes in the course of the discussion. If there is high activity of the starter after starting the thread, he is interested in a solution for his problem. Another case could be, that the content of a thread is of high interest, then *hits* expresses how many users clicked on this discussion.

3.2.2 Time dependent interestingness features

Author level

Active days are days on which the author wrote at least one post. *Duration* is the time span since registration for the forum until today.

Post level

In an analysis it could be interesting to see how fast other users write a reply to the previous post. *Response time* measures the elapsed time, starting from the previous post. Following a controversial contribution, posts could be made within seconds.

Discussion level

To measure actuality we use *last post* to see, when the latest post was written. Another closely related time dependent property is the *duration* of a discussion. Thereby, we measure the time span from the first to the last post. *Active days* are days, where at least one post is written. *Activity*

contains, how active and lively a discussion is. A discussion is interesting in terms of activity if there are days in time when there was a high appearance of posts. It does not matter, when the discussion was interesting but only that there are days where the thread was extended by posts. Hence, we do not use the *duration* of a thread for this feature, but we set up the ratio of *active days* and the overall score of *posts* in the concerned discussion. The less the ratio, the more activity in terms of posts was concentrated on few activity days.

3.2.3 Affective features

Author level

Emoticons are an obvious source measuring emotional content in posts. Please note, that *emoticons* only show the sum of emoticons an author used globally in a forum. Here it does not matter, which sentiment the emoticons belong to. However, the polarity of an emoticon plays a role in *sentiment*, which contains a sentiment score. Here we take the taggings of the emoticons. For example, the tagging looks like "thumbs up" or "thumbs down". They have a higher impact on the sentiment score because they reflect the emotions quite clear and are meaningful even without context information in a post. Furthermore, we used positive and negative word lists to extract positive and negative terms. The lists we used, were developed especially for sentiment analysis in the area of news and blogs [31, 32]. Additionally, we adjusted the lists, e.g. we deleted general used terms and added special terms, to get better results in technical web forums. Furthermore, we paid attention to negations and ignored upper and lower cases, because in web forums people often ignore that. For the *sentiment* score we take the average of all sentiment scores on the post level written from the same author. Words and sentences where every character is capitalized are used for shouting in online discussion forums. Thus, *shouting* is a boolean value which tells the analyst if the author writes "loudly" in his postings or not. These posts can be also interesting in an analysis process.

Post level

The following three features are already introduced in the preceding paragraph. The difference here is, that the scores are features on the post level. So *emoticons* is the amount of emoticons in a post, *sentiment* reflects the sentiment score of a post and *shouting* if there are shouted words or not.

3.2.4 Content and other interestingness features

Post level

Thread count shows the position of a post in its corresponding thread. Interesting in this respect could be the intention for writing a post e.g. as a reply to a previous post. Hence, in *question* we store a boolean value if a posting contains a question or not.

Discussion level

Not only the popularity or chronology are worth being measured. Just the content of a thread provides meaningful information. Therefore, we measure the *sentiment* of a thread using a sentiment dictionary. Furthermore, we store the *provider* a user is talking about in the thread, because in many tasks that could be of high impact. Also important: the *title* of a thread.

3.3 The ForAVis User Interface

ForAVis was implemented in Java. For functionality and interactivity purposes we used the *Prefuse* framework [18]. In Figure 3 you see the complete ForAVis user interface. Our system follows the principles of visual data exploration: "overview, zoom and filter, then details on demand" [34]. The anticipated exploration behaviour has three steps: first we want to give the user an overview of the data set. Afterwards, we enable the user to filter the data depending on the task. Since we use *Linking & Brushing* [21] the user will see immediately the result in the visualization. Now the user is able to see interesting discussions and this can be used as starting point for a further explorative analysis. By clicking on a discussion we show a post frequency visualization, the whole thread segmented in its posts and further details (features and content) in a text field.

3.3.1 Main window

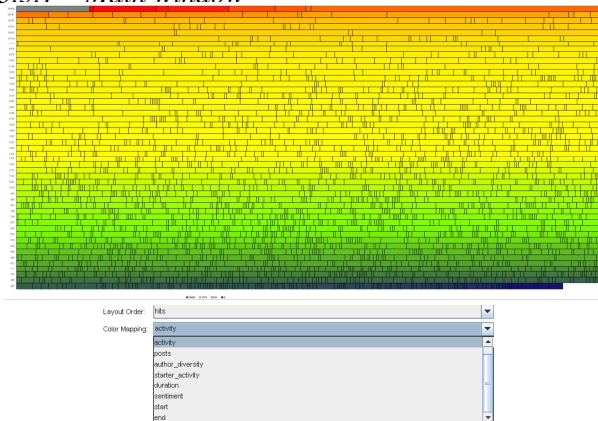


Figure 4: The main visualization of ForAVis. Two drop down menus enable the user to interact with the visualization. Available features for the layout ordering and the color mapping are: hits, activity, posts, author diversity, starter activity, duration, sentiment, start and end.

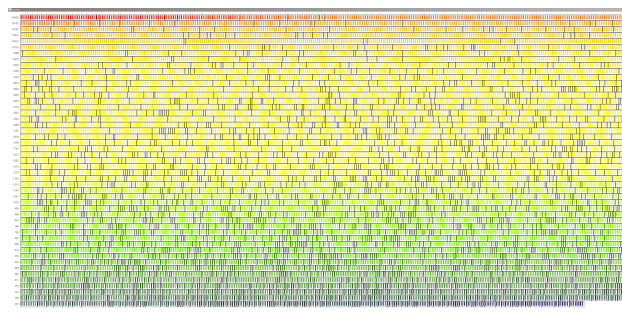


Figure 5: The complete ForAVis user interface in the post mode. Each post of a thread is visible within the horizontal thread bars. The perception of the colours is different compared to the picture above. This is a result of the white spaces between each post item in the visualization.

In thread mode the main visualization (Figures 4 and 3) show each thread in a horizontal bar in a chromogram visualization with heatmap coloring from red to yellow to green

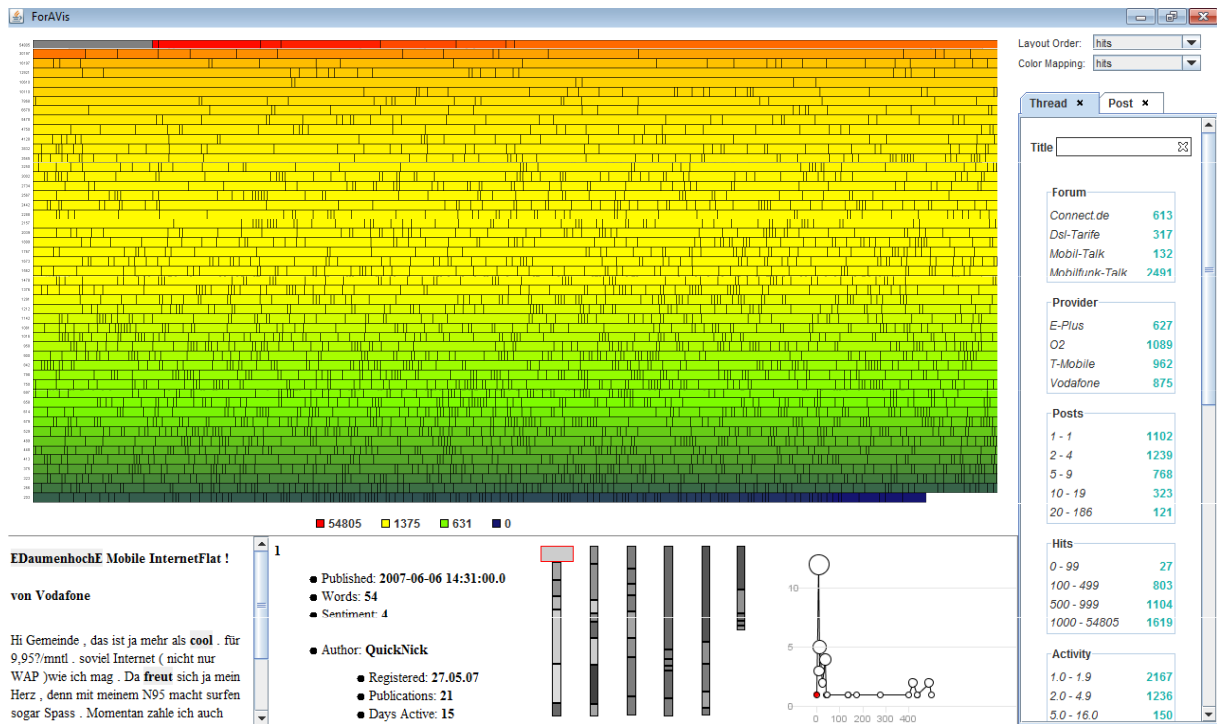


Figure 3: The complete ForAVis user interface. The main visualization and its menus to change the ordering and coloring in the upper right corner. At the right margin is the filter area and at the bottom of the screen, exactly below the visualization, you can see the detail area with a text field, a thread visualization and a frequency graph.

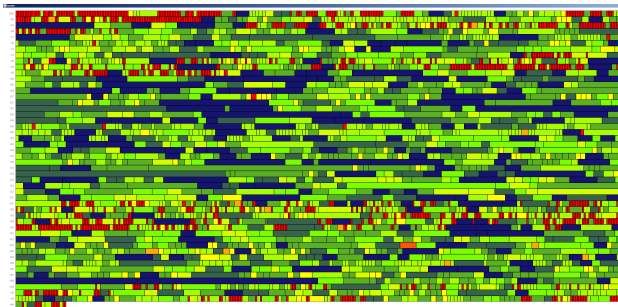


Figure 6: ForAVis with changed layout ordering (sentiment) and changed colour mapping (starter activity) of the thread applied. Threads with the worst sentiment score are in the lower right corner or in this case also just in the last line. Red threads reflect a high starter activity, whereas the blue ones show a low starter activity. Yellow and green are amongst red and blue.

to blue. The length of a horizontal bar is proportional to the amount of posts in it. In the post mode the visualization looks like Figure 5. Here each post is visible separately. Due to perception and analysis reasons the coloring of the whole thread remains in the post mode (if a thread is marked, post level information is shown in the detail area). We used a heuristic to align the thread bars to get a filled visualization window. The color mapping is done using three equi-depth bins of the data to align the colors. A legend helps the user to get a feeling for the values behind the colors but we also tried to map the colors preattentively (Choosing *hits* leads amongst other to red colored threads, which indicate often clicked respectively "hot" discussed topics of high interest in a thread). A tooltip shows the facts of a thread during moving on it with the cursor. The same happens for each post in the post view. The default layout and coloring is set to *hits*, but two menus enable the user to interact with the system and to change the layout and the color mapping. Possible other features beside hits are activity, posts, author diversity, starter activity, duration sentiment, start and end (Figure 4). The layout ordering always starts in the upper left corner from left to right, like reading text. Hence, the maximum value of the ordering feature is always the first bar in the upper left corner. Accordingly, the smaller values are in the lower right corner or the smallest ones are definitely at the end of the last line in the main visualization. At the beginning of each row in order to support and to alleviate the orientation in the visualization, we show the value of the first bar in this row. In Figure 6 you can see the visualization in thread mode with changed layout ordering (sentiment) and

changed color mapping (starter activity) applied.

3.3.2 Filter area

On the right side of the ForAVis screen you can see the filter area in Figure 3. The available filter options depend on the structure level (Figure 3.2). We give the user the choice to explore the forums on the post or on the thread level. An extract of both filter areas is shown in 7. The filters for authors were integrated in the post filter menu. The option to search for authors is practically an extensive possibility we give to the user to search and explore posts. Additionally, we offer a button, to get a list including the ten most actively contributing authors in the forums. In all cases the implemented technique bases upon *Elastic Lists* which were introduced by Stefaner et al. [36, 37]. This helps the user to gain insight in the data and its structure. Our intention was to give the user the ability and flexibility to to change his search behaviour in an easy way. The combination of filters on thread or post respectively author level can be chosen freely by the user. All filter options show the user how many data objects will be visualized, if a particular feature is selected. All applied filters immediately update the filters and the main visualization and thereby, all data objects which do not contain any selected feature are faded out. Inversely, you can highlight a region in the main visualization and just this region is visualized in the main window. All the labels of the visualization are adjusted and also the list entries are adapted. In general, the list options which would lead to no result are faded out automatically, too.

All the introduced techniques guide fast to exploration results. During the exploration process the fast visual feedback supports the user in finding patterns and interesting discussions or articles.

To make the search process more effective we also give the possibility to search for terms in a free text field. On discussion level the search is applied to the titles of discussions, on post level to the content of articles. The result also affects the visualization and the lists in order to give the user fast feedback.

3.3.3 Detail area

After clicking an object in the main window the whole thread is displayed in the detail area. In Figure 3 at the lower boundary and in Figure 9 you can see the components of the detail view. It consists of a text field for showing posts and other interesting features belonging to this post. In the text of the post the sentiment words are highlighted. Additionally, you can see the author, when he registered for the forum, how many publications he has done, how many active days he has, his average sentiment. Furthermore, the word count per post and word count per sentence of his postings. Belonging the chosen post the following features are displayed: when it was published, how long the post is and its sentiment.

Also in the detail view we provide a visualization which is closely related to *conversation thumbnails* [42]. Each rectangle represents a posting coloured in greyscales. The darker the color, the more negative the sentiment of the posting. Since the colormap of the main visualization is not static associated with the sentiment score, we decided to use stati-

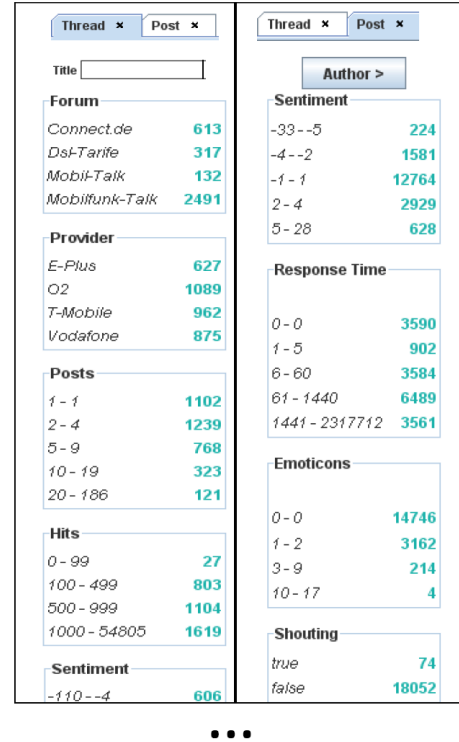


Figure 7: Incomplete filter area of ForAVis: on the left side you can see filters for the thread view, on the right side for post view.

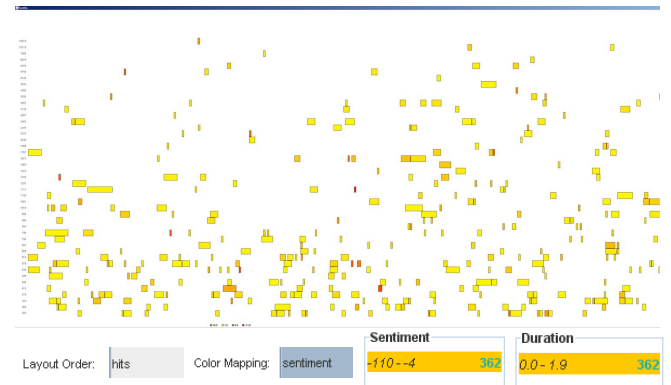


Figure 8: Applied Filters in action: 362 threads are shown with bad sentiment and a very short duration (the background colouring of the list entries has nothing in common with the colours in the visualization, but designates the applied filters). The layout ordering shows *hits*, the color mapping the *sentiment* of the threads.

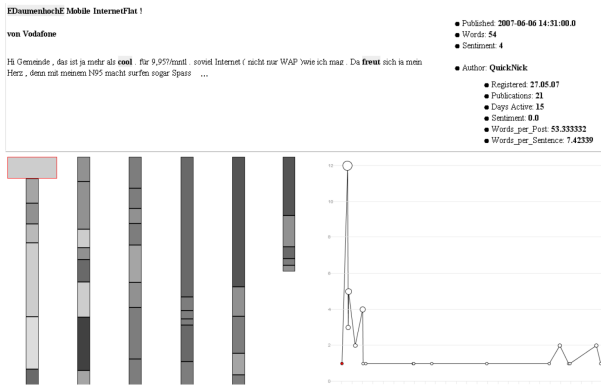


Figure 9: The components of the detail view. It consists of a text field, a discussion visualization and line graph to show the activity of the thread measured in postings per day.

cally greyscales for the sentiment in detail view. The height of the rectangle is proportional to the number of words of the posting. Unlike the representation in the main window, the discussion is visualized vertically in the detail view. Two reasons led us to this design: on the one hand postings in internet forums are displayed in vertical lists, separated on consecutive pages, on the other the user has the possibility to scroll through the entire thread, what improves the interaction with the tool and the usability. A marked posting is drawn to a larger scale.

The second visualization in the detail view shows a line graph. It indicates the frequency of posts in time since the start of the thread. On the x-axis we display the duration of the discussion and on the y-axis the number of posts to illustrate the activity. In addition, the circle size on the graph is also proportional to the sum of postings at this time. *Linking & Brushing* [21] helps the user to explore the thread in an intuitive way. If the user selects a post in the detailed thread visualization, the corresponding circle on the frequency graph is colored in red. Supplementary, all posts belonging to this day get a red frame. These changes also happens, if the user choose a post item in the main visualization. The content of the accordant post appears in the textfield, where the title is shown on top.

4. SCENARIOS IN MOBILE COMMUNICATION FORUMS

In this section we provide three concrete example scenarios to demonstrate how our application, ForAVis can be used for different analysis tasks. Our dataset consists of the public content of nine mobile communication forums. Since we have real data, we present stakeholder scenarios in these forums. Here we mention users, companies and forum operators respectively forum moderators. Furthermore, we identify another application areas where ForAVis could be used. Although the scenarios are based on real data, they are completely fictitious.

4.1 Users Perspective

Tom wants to have a new mobile phone contract

Tom is searching for a new contract with a mobile communication provider. He is a student and therefore he is able to get special student conditions. Tom started to search for a contract on the various websites of the providers. But it is not that easy to find a suitable contract due to multiple contract options. On some websites he did not even find the possibility for a student contract. This led him to use ForAVis for further exploration of user experiences and for discussions concerning student contracts. First he typed in the phrase 'student' in the search text field on the thread level in the filter area. The result are 28 discussions with respect to 'student' in their title. In the filter area he chose one provider after the other. First he began with the provider E-Plus and sorted the 4 threads according to their *sentiment* and mapped *hits* on color. He found quickly one discussion with bad sentiment and many hits which told him, that E-Plus suspended the possibility for students to get a cheaper contract and additional bonuses.

He continued and selected another provider O2, where he found 16 threads having 'student' in the title. Clicking through the discussions quickly he recognized, that O2 obviously still have contracts for students. Taking a look at the lower end of the main visualization (the layout ordering is still *sentiment*) he found out, that the discussion with the worst sentiment was only about not finding a student option on the website. That does not matter for him, because through ForAVis it was quickly clear for him that O2 offers this product and the discussion did not have poor sentiment regarding O2 or the contract conditions. In addition, he also applied the filter for *hits*. He only wanted to see discussions with more than 1000 hits. The result set consisted of seven threads. Through this he discovered, that one thread, started on Nov. 10th 2007, discussed a voucher to get 150 SMS instead of 100 SMS, which was obviously good to know.

4.2 Companies Perspective

Do we have trouble...

which we do not know yet? This could be the first question and a starting point for a company using ForAVis. For example, T-Mobile, a German mobile communication provider, could search for threads in the T-Mobile sub-forum. There are 962 threads. The layout ordering and color mapping stay in the default alignment *hits*. The company is keen to find out new issues for T-Mobile posts and hence the filters are set to include the posts with bad *sentiment* dedicated (another option can be applying *emoticons*). Only threads with one author are selected. In the end 45 discussions are found to have been started with a really negative posting. Finally, the *hits* filter selects only discussions having more than 1000 hits. The outcome of this filtering step is 12 discussions. To use the space more efficiently for the visualization the analyst could arrange the discussions in a sequence filling the whole visualization area. The results could help the company to discover rumors early on and also issues they are not aware of.

Which product is missing in our portfolio?

What do people like from other providers. What is popular with the community? These are interesting questions

Table 2: 3 discussions with bad sentiment for T-Mobile. These discussions can be also seen at the end of the visualization.

Title	Sentiment	#Hits	#Authors
"Withdrawal = mobil more expensive"	-11	1183	1
"iPhone: 4000 Euro invoice despite Complete-M-Rate"	-7	4611	1
"T-mobile discontinue providing t-email"	-5	3390	1

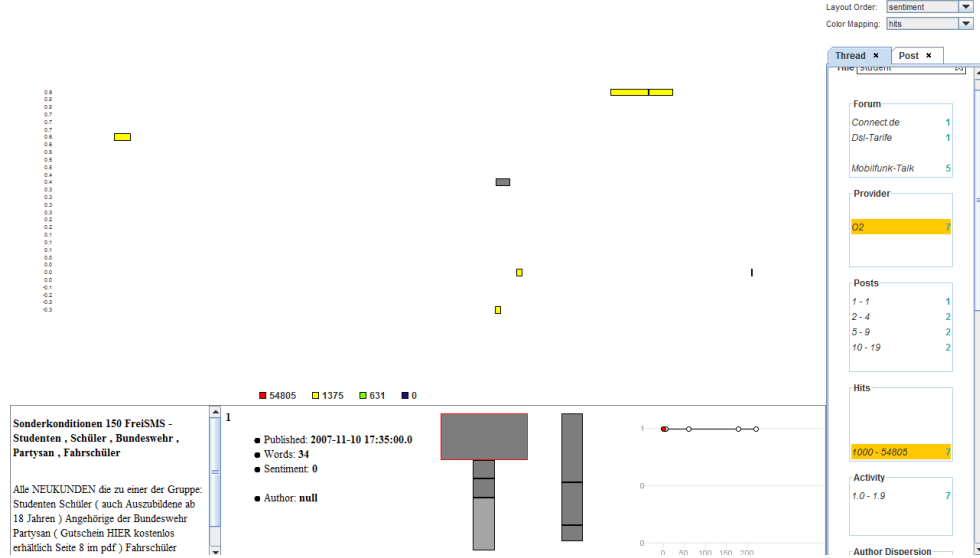


Figure 10: Tom is searching for a new contract: the result set of his analysis consists of seven threads.

for further analysis. Assume that T-Mobile wants to know why people like Vodafone. They apply "Vodafone" in the *Provider* filter, the best sentiment category in the sentiment filter and discussions having more than 1000 *hits*. The results are 233 discussions. Both, the layout ordering and color mapping is still in the default alignment *hits*. The top three threads out of four have 41, 31 and 17 postings. To explore these discussions the analyst uses the detail view, to get an impression of the content: all these threads deal with Vodafone products (Figure 11).

4.3 Forum Operators Perspective

Searching for new moderator candidates

The task of a moderator is to guarantee for the code of conduct in a forum and to pay attention that the community members are on familiar terms in communicating with each other. So the post level is the relevant one here for this task. The analyst is therefore searching for a frequent author with positive sentiment scores in his posts which could indicate friendliness. Also a moderator should not shout, hence *shouting* has to be false. An interesting observation outside the visualization can be done: the author list in the filter area shows the ten most contributing authors. Within these authors we found a candidate named "Matzezetel" with the amount of 17 contributions. But the webmaster is leader of the list: 71 posts remain after filtering on the highest sentiment level.

Detractive posts and authors

When the new moderator is found he can do his job using ForAVis. Ultimately, the task for him is to find detractive

posts and their authors. Trying the combination of *shouting* and *sentiment* led to no results. In order to make a deeper analysis, the moderator would like to have more features to solve this task.

5. CONCLUSIONS

We presented an interactive visual analysis tool for exploring user generated discussion boards. All the data we used is publicly available. The visualization and the possibility to combine almost all features on different levels give great flexibility to the analyst in different search and exploration tasks. Our fictive scenarios showed application examples for ForAVis. The presented tool could be also helpful in other monitoring tasks, for example with respect to security issues and in other communication applications such as chatrooms. For another applications and tasks and for a deeper analysis of forums e.g. out of topic detection, more advanced features are needed, which take for example linguistic features in account. But also obvious features should be implemented like e.g. the distribution of sentiment on the post level in the main visualization. We have also not considered another structures such as dialog conversations which often occur in forums.

In the future we want to integrate more metrics and data mining methods in the analysis. A user study for further evaluation purposes is considered.

6. REFERENCES

- [1] Bar code in the light spectrum of a star. *Geo Magazin* – *Is there anyone else?*, (6):84–102, June 2010.

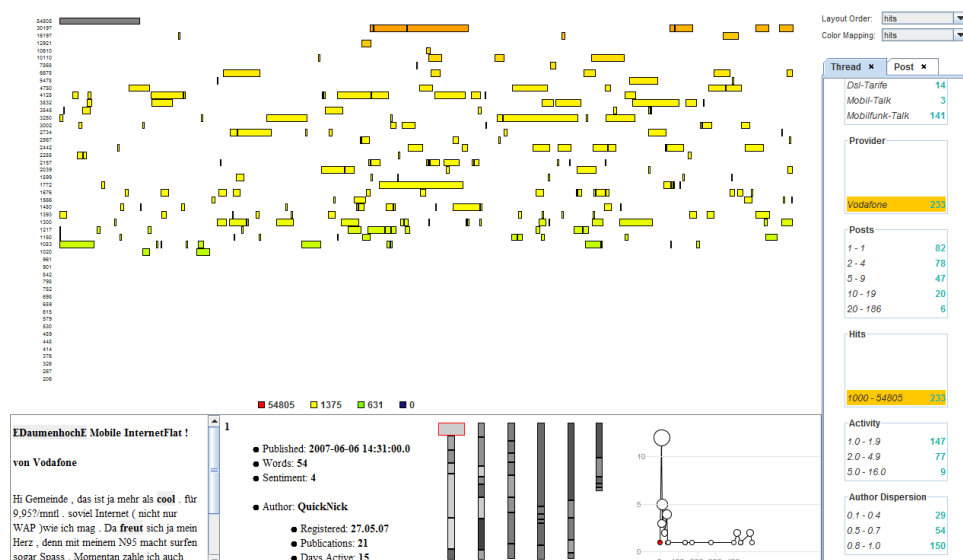


Figure 11: Why people like Vodafone: applied filter 'Vodafone', the best sentiment category and discussions having more than 1000 hits. The results are 233 discussions. Both, the layout ordering and color mapping still display the default alignment hits. The first post of the first thread is shown in the detailed view.

- [2] Mobile communication forum on www.connect.de, 2010. <http://community.magnus.de/forum/f353/> as retr. in Jan., 2010.
- [3] Mobile communication forum on www.dsl-tarife.de, 2010. <http://www.dsl-tarife.de/> as retr. in Jan., 2010.
- [4] Mobile communication forum on www.mobil-talk.de, 2010. <http://www.mobil-talk.de/> as retr. in Jan., 2010.
- [5] Mobile communication forum on www.mobilfunk-talk.de, 2010. <http://www.mobilfunk-talk.de/> as retr. in Jan., 2010.
- [6] mobilfunk-talk, 2010. <http://www.mobilfunk-talk.de/search.php> as retr. on Nov. 24, 2010.
- [7] mobilfunk-talk, 2010. <http://www.mobilfunk-talk.de> as retrieved on Nov. 24, 2010.
- [8] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, pages 183–194. ACM, 2008.
- [9] C. Ahlberg and B. Shneiderman. Visual information seeking using the filmfinder. In *Conference companion on Human factors in computing systems*, CHI '94, pages 433–434, New York, NY, USA, 1994. ACM.
- [10] K. Chai, V. Potdar, and T. Dillon. Content Quality Assessment Related Frameworks for Social Media. *Computational Science and Its Applications-ICCSA 2009*, pages 791–805, 2009.
- [11] S. Eick, J. Steffen, and E. Eric Jr. S. 1992. Seesoft – a tool for visualizing line oriented software statistics. *IEEE Transactions on Software Engineering*, 18(11):957–968.
- [12] J. Elsas and N. Glance. Shopping for Top Forums: Discovering Online Discussion for Product Research. 2010.
- [13] B. Engdahl, M. Köksal, and G. Marsden. Using treemaps to visualize threaded discussion forums on PDAs. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1355–1358. ACM, 2005.
- [14] R. Farrell. Summarizing electronic discourse. *Intelligent Systems in Accounting, Finance & Management*, 11(1):23–38, 2002.
- [15] D. Feng, E. Shaw, J. Kim, and E. Hovy. Learning to detect conversation focus of threaded discussions. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 208–215. Association for Computational Linguistics, 2006.
- [16] E. Giguet and N. Lucas. Creating discussion threads graphs with Anagora. In *Proceedings of the 9th international conference on Computer supported collaborative learning-Volume 1*, pages 616–620. International Society of the Learning Sciences, 2009.
- [17] N. Glance et al. Deriving marketing intelligence from online discussion. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 419–428. ACM, 2005.
- [18] J. Heer, S. Card, and J. Landay. prefuse: a toolkit for interactive information visualization. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430. ACM, 2005.
- [19] M. Hu, A. Sun, and E. Lim. Comments-oriented blog summarization by sentence extraction. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 901–904. ACM, 2007.
- [20] J. Jeon, W. Croft, J. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th annual*

international ACM SIGIR conference on Research and development in information retrieval, pages 228–235. ACM, 2006.

- [21] D. Keim. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.
- [22] B. Kim and P. Johnson. Graphical interface for visual exploration of online discussion forums. In *SYSTEMICS, CYBERNETICS AND INFORMATICS*, volume 4, pages 43–47, 2006.
- [23] M. Klaas. Toward indicative discussion fora summarization. Technical report, University of British Columbia, 2005.
- [24] F. Lam and J. Donath. Seascape and volcano: visualizing online discussions using timeless motion. In *CHI’05 extended abstracts on Human factors in computing systems*, pages 1585–1588. ACM, 2005.
- [25] H. Lee, P. Ferguson, N. O’Hare, C. Gurrin, and A. F. Smeaton. Integrating interactivity into visualising sentiment analysis of blogs. 2010.
- [26] N. Li and D. Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2):354–368, 2010.
- [27] X. Li and W. Croft. An information-pattern-based approach to novelty detection. *Information Processing & Management*, 44(3):1159–1188, 2008.
- [28] S. Narayan and C. Cheshire. Not Too Long to Read: The tldr Interface for Exploring and Navigating Large-Scale Discussion Spaces. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.
- [29] P. Newman. Exploring discussion lists: steps and directions. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 126–134. ACM, 2002.
- [30] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
- [31] R. Remus, G. Heyer, and K. Ahmad. Sentiment in German language news and blogs, and the DAX. *Text Mining Services 2009*, pages 149–158, 2009.
- [32] R. Remus, U. Quasthoff, and G. Heyer. SentiWS-a publicly available German-language Resource for Sentiment Analysis, 2010.
- [33] B. Sharifi, M. Hutton, and J. Kalita. Experiments in microblog summarization. In *IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust*, 2010.
- [34] B. Shneiderman. The eye have it: A task by data type taxonomy of information visualizations. *Visual Languages*, 2006.
- [35] M. Smith and A. Fiore. Visualization components for persistent conversations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 136–143. ACM, 2001.
- [36] M. Stefaner and B. Muller. Elastic lists for facet browsers. In *Database and Expert Systems Applications, 2007. DEXA’07. 18th International Conference on*, pages 217–221. IEEE, 2007.
- [37] M. Stefaner, T. Urban, and M. Seefelder. Elastic lists for facet browsing and resource analysis in the enterprise. In *19th International Conference on Database and Expert Systems Application*, pages 397–401. IEEE, 2008.
- [38] A. Tigelaar, R. OP DEN AKKER, and D. Hiemstra. Automatic summarisation of discussion fora. *Natural Language Engineering*, 16(02):161–192, 2010.
- [39] T. Turner, M. Smith, D. Fisher, and H. Welsler. Picturing Usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication*, 10(4):7, 2005.
- [40] N. Wanas, M. El-Saban, H. Ashour, and W. Ammar. Automatic scoring of online discussion posts. In *Proceeding of the 2nd ACM workshop on Information credibility on the web*, pages 19–26. ACM, 2008.
- [41] N. Wanas, A. Magdy, and H. Ashour. Using automatic keyword extraction to detect off-topic posts in online discussion boards, 2009.
- [42] M. Wattenberg and D. Millen. Conversation thumbnails for large-scale discussions. In *CHI’03 extended abstracts on Human factors in computing systems*, pages 742–743. ACM, 2003.
- [43] M. Wattenberg, F. Viégas, and K. Hollenbach. Visualizing activity on wikipedia with chromograms. *Human-Computer Interaction-INTERACT 2007*, pages 272–287, 2010.
- [44] W. Weerkamp and M. De Rijke. Credibility improves topical blog post retrieval. *ACL-08: HLT*, pages 923–931, 2008.
- [45] M. Weimer and I. Gurevych. Predicting the perceived quality of web forum posts. In *Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing, RANLP*, volume 2007, 2007.
- [46] K. Yee and M. Hearst. Content-Centered Discussion Mapping. *Online Deliberation 2005/DIAC-2005*.
- [47] Y. Zhang and F. Tsai. Combining named entities and tags for novel sentence detection. In *Proceedings of the WSDM’09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 30–34. ACM, 2009.
- [48] L. Zhou and E. Hovy. On the summarization of dynamically introduced information: Online discussions and blogs. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, Stanford, CA*, 2006.