Detecting Opinion Leaders and Trends in Online Social Networks

Freimut Bodendorf
University of Erlangen-Nuremberg
Lange Gasse 20, 90403 Nuremberg
Germany
0049-911-5302-450
bodendorf@wiso.uni-erlangen.de

Carolin Kaiser
University of Erlangen-Nuremberg
Lange Gasse 20, 90403 Nuremberg
Germany
0049-911-5302-295
carolin.kaiser@wiso.uni-erlangen.de

ABSTRACT
Today, online social networks in the World Wide Web become increasingly interactive and networked. Web 2.0 technologies provide a multitude of platforms, such as blogs, wikis, and forums where for example consumers can disseminate data about products and manufacturers. This data provides an abundance of information on personal experiences and opinions which are extremely relevant for companies and sales organizations. A new approach based on text mining and social network analysis is presented which allows detecting opinion leaders and opinion trends. This allows getting a better understanding of the opinion formation. The overall concept is presented and illustrated by an example.

Categories and Subject Descriptors
H.2.8 [**Database Management]** Database applications - *Data Mining*

General Terms
Algorithms

Keywords
Opinion Mining, Social Network Analysis, Leaders, Trends

1. MOTIVATION
The Internet is constantly developing from a static to a highly interactive medium. Users now have the possibility of not only obtaining information but also of actively generating contents. The transformation to Web 2.0 is favored by a number of technical, economical, and social factors. The increasing speed plus the decreasing costs of data transmission rates allow for development and copious usage of interactive Web applications. The new self-awareness of the information society has lead to the fact that more and more users connect online in social networks in order to exchange opinions. They interact with each other and influence their opinions reciprocally. Consumer feedback on a company's product is essential to recognize chances and risks and to implement appropriate marketing measures

SWSM'09, November 2, 2009, Hong Kong, China.

The theory of diffusion describes the spread of product opinions in social systems. According to this theory, not only the product characteristics (e.g. complexity) are an important factor in the formation of opinions but also the structure of the social networks [13]. The structure consists of various communication relationships among the users in social networks. Who communicates with whom thus influences the making of opinions. In order to assess opinion formation in online social networks, it is necessary to analyze specific users' opinions and their communication relationships. A manual analysis is labor intensive and only possible on a limited scale. Due to the myriad of communication platforms and users in Web 2.0 it is also a great challenge. Here, an approach is introduced which automatically extracts opinions and communication patterns in forums by text mining and identifies opinion leaders and trends using social network analysis.

2. RELATED WORK
Diffusion theory explains the spread of influence in social networks. Initially, it was applied in the fields of agriculture and medicine [13]. Recent research examines diffusion in Web 2.0. For example, Java et al. [7] show how to find influential bloggers for maximizing the spread of information in the blogosphere. Goyal et al. [6] identify leaders in social networks based on their actions. Song et al. [15] detect leading bloggers who dissiminate new ideas. However, opinion leaders and trends are not studied with respect to the polarity of opinions towards certain topics.

Text mining allows for detecting interesting parts in texts [8]. While prior research focused on the discovery of facts, current work centers on emotions and opinions [8]. There are already many relevant papers on text mining in connection with opinions in Web 2.0. For example, Dave et al. [4], Liu et al. [9] as well as Popescu and Etzioni [10] describe systems which analyze customer opinions and product reviews. Here, single opinions and reviews are determined and then aggregated. However, communication processes and relationships among customers who play significant roles in opinion forming have not been taken into consideration in previous work.

Social network analysis enables the examination of patterns in the relationships among interacting users [16]. It is used increasingly for Web 2.0. The papers of Welser et al. [18], Chang et al.[2] and Gomez et al. [5] for instance explore roles, positions, and patterns of communications in online conversations. However, the conversations' content is not considered. An integrated analysis of communication content and relationships is missing.

3. APPROACH

A new approach is presented which initially detects opinions and relationships among forum users by text mining. On this basis main influential factors for opinion forming in virtual communities are extracted. By social network analysis opinion leaders are identified and opinion evolvement is analyzed. The results reveal trends in the dynamics of online social networks. The analysis of online communities comprises four steps. In the first step the users' opinions on the product are extracted by text mining. In the second step, the communication relationships among users are identified by text based relationship mining methods. The extracted users, opinions, and relationships form a social network which is represented as graph. Nodes and edges of the graph can be characterized by attributes. The nodes represent the users of a forum and the edges their communication relationships. The nodes' attributes describe the users' opinions e.g. on a specific product and the edges' attributes show the frequency of communication. The resulting graph is analyzed by determining key figures for the position of single users and for the overall structure of the network. In this way opinion leaders and opinion trends can be identified. The approach is illustrated by looking at sample forums in which opinions on Apple's iPhone® are exchanged.

4. RECOGNITION OF OPINIONS

The goal of opinion recognition is to detect the attitude of a user towards a product based on his/her statements in forum postings. Attitudes are characterized by their polarity, which is modeled by the three classes 'positive', 'neutral', and 'negative'.

The automated recognition of positive, negative, and neutral opinions using text mining comprises two phases, i.e. the extraction of features from the text and the application of a learning algorithm to identify the polarity of the posting based on the extracted features. The extraction of features is carried out by linguistic and statistical analysis [17]. After removal of unimportant stop words (e.g. 'article') the remaining words are reduced to their stem and then their frequency is calculated. Those word stems which are especially distinctive for each of the three polarity classes are used as the main features of the postings.

Different methods such as Hidden Markov Models or Maximum Entropy can be used for classification [17]. Here, support vector machines [3] are employed because of their success in related projects [11]. Input consists of sample data records which in turn comprise various postings, their features, and the matching classes. By analyzing sample data support vector machines learn the parameters of the rule which classify the postings best. The rule allows for a clustering of the postings into two groups (binary classification). In case of three classes, three rules need to be learned: 'positive' versus 'not positive', 'negative' versus 'not negative', and finally 'neutral' versus 'not neutral'. A posting will be matched to the class which has the highest probability.

After classification, the opinions are assigned to the forum users. If a user has diverging opinions he is linked to the class reflecting his strongest opinion.

In order to validate the opinion classification a total of 929 forum postings about the iPhone were manually allocated to the three classes. A stratified five-fold-cross-validation yielded precision and recall shown in Table 1.

Table 1: Validation Results

Opinion	Precision	Recall
Positive	0.68	0.61
Negative	0.69	0.67
Neutral	0.69	0.73

5. IDENTIFICATION OF RELATIONSHIPS

A relationship between two forum users exists if one user (sender) refers to another user (recipient) in at least one posting. A communication relationship is characterized by the sender, the recipient, and the communication frequency.

In the field of text mining there are two approaches which deal with the recognition of relationships among entities in texts: coreference resolution and relationship extraction [17]. The aim of coreference resolution is to recognize if two words in one text refer to the same object. A typical application is to determine whether a specific pronoun refers to a previously mentioned noun. Relationship extraction deals with the identification of different relationships between two components of a sentence. For example, the task could be to find the relationship 'is producer of'. Looking abstractly at this challenge, the goal is to solve a binary classification problem. Each pair of possible text entities will be divided into two classes: 'has defined relation' and 'has no defined relation'. The classification is based on statistical-linguistic attributes of the one or the other text element and their common attributes.

When classifying communication relationships of users in a forum, each posting of a user is analyzed for relationships with previous postings of different users. Important signs indicating the existence of a relationship between two users are: mentioning of user names, quotations, distances of the postings and appearance of specific words. These criteria are checked by string matching and used as attributes for classification. The classification is done by support vector machines. After classification, the relation frequency of all pairs is calculated by summation.

For validation purposes, a human annotator marked 1002 relationships within 17 threads about the iPhone. A stratified five-fold cross-validation yielded the results depicted in Table 2.

Table 2: Validation Results

Precision	Recall
0.81	0.78

6. DETECTION OF OPINION LEADERS

The extracted users with their opinions and relationships form a social network which is analyzed first on user level. The goal is to identify opinion leaders. Opinion leaders are persons who play a crucial role in forming opinions. Their special position and their communication habits within the network allow them to influence the opinion of many other users. They offer an orientation for users who are searching information. The identification of opinion leaders provides great potential for marketing. The orientation function of opinion leaders with positive opinions can be used to deliberately spread specific information. Opinion leaders with negative views can be counteracted in a targeted way.

There are many key figures in the field of social network analysis which describe the position and the communication habits of us-

ers. When observing opinion leaders, the key indicator "centrality" is of special importance. It marks on a scale from 0 to 1 how central a user is within the network [14]. Three different figures are common: degree centrality, closeness centrality, and betweenness centrality. Degree centrality measures how frequently a user communicates directly with others [16]. Closeness centrality measures the closeness of one user to all other users [16]. Betweeness centrality shows how often a specific user can be found on the shortest connecting path of all pairs of users [16].

Degree centrality is an indicator for local opinion leaders. Forum users with high degree centrality communicate very frequently directly with other users and have the possibility of influencing them. Their influence, however, is limited to their direct environment. Closeness and betweenness centrality not only look at the direct communication relationships but also at indirect ones and thus are indicators for global opinion leaders. Forum users who have the shortest distance to all other users are most likely to be noticed by others. They are very independent from the information transferred by other users. Those users found on communication paths between other users have the highest chance of controlling the information flow between users.

Figure 1 shows an examplary thread from the forum computerbase.de. User 28 has the highest degree centrality with a value of 0.13. However, the user's positive opinion on the iPhone is only shared by his closer environment. User 34 with a closeness and betweenness centrality of 0.3 is closest to all other users and controls the information flow best. He has a negative opinion which is also with 49% the most frequent opinion class in the forum (only 34% positive opinions and 17% neutral opinions).

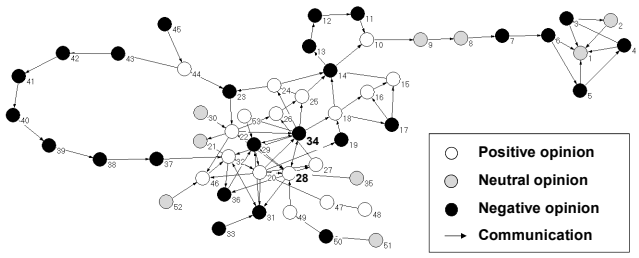

Figure 1: Thread from the forum computerbase.de

For the first validation, in this case 17 threads from forums regarding the iPhone are chosen. The focus is put on opinions of the users with the highest value of closeness and betweenness centrality as well as on the overall opinion within the forum in terms of the most frequent opinion class. The validation is carried out via Spearmansch correlation coefficient [1]. The Spearmansch correlation coefficient of 0.56 shows an important relation between the opinion of the user with the highest closeness centrality and the overall opinion in the forum. The positive correlation between the opinion of the user with the highest betweenness centrality and the overall opinion was at a value of 0.49 and significant as well.

7. DETECTION OF OPINION TRENDS

The aim of the analysis on network level is to explain and forecast the formation of opinions in a forum. Two aspects of opinion formation are of special interest: first the characteristics of a network which lead to a particular opinion trend and second the characteristics of a network in which the trend follows the opinion of the opinion leader. An early identification of positive and negative trends allows to plan and support effective reinforcing and counteracting measures.

Social network analysis provides a number of key figures which describe the structure of the entire network. For analyzing opinions, the key figures density, randic connectivity, and closeness centralization are especially relevant. Density measures the connection of a network and is an indicator for communication within the network [14]. The higher the calculated value of density the more information can be exchanged within the network. Randic connectivity shows the level of branching within the network [12]. A low value indicates a high level of branching which means that many users have direct relationships to a high number of users. Closeness centralization indicates the variation of the closeness centrality values of the network's users [16]. A network is strongly centralized if there are only a few central and many peripheral users.

Density and Randic connectivity indicators show if there is a trend of the overall forum opinion. The higher the density and the lower the Randic connectivity, the more forum users are connected with each other and communicate directly. In this way, opinions can be exchanged increasingly. Thus, there is a high probability of a trend of the opinion of all users regarding a specific product. The trend becomes stronger the more one opinion class outweighs the others. It is calculated as the product of the deviations of the largest opinion class from the other two.

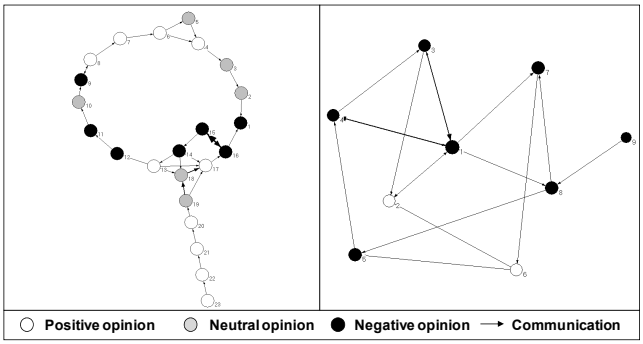

Figure 2: Threads from the forums slashfm.de and pcfreunde.de

Figure 2 shows two threads regarding the iPhone. The thread from the forum slashfm.de (Figure 2, left side) has a relatively low density of 0.08 and a relatively high Randic connectivity of 11.33. It also shows a low opinion trend of 0.03. 44% of the users have a positive, 30% have a negative, and 26% a neutral opinion on the iPhone. The thread from pcfreunde.de (Figure 2, right side), in contrast, has a relatively high densitiy of 0.22 and a low Randic connectivity of 4.22. The opinion trend is relatively high with a value of 0.43. 78% of the users think negatively about the iPhone and 22% think positively.

Centralization indicates whether a trend arises from an opinion leader. If there is a high value of closeness centralization, there is one or a few opinion leaders in the center of the network and many users on the periphery. The opinion of the leader can thus be spread easily. If the network is, in addition, strongly linked the leader's opinion can be spread among many other forum users which could start a trend following the leader's opinion. The pcfreunde.de thread in Figure 2 illustrates this. The closeness

centralization is relatively high with a value of 0.4. 78% of users share the negative opinion of the opinion leader (user 1) who has the highest closeness centrality of 0.7.

The first validation was based on 17 threads about the iPhone. The key figures density, Randic connectivity, closeness centralization, opinion of the opinion leader, and tendency of the overall opinion were calculated. The coherence was then measured by using Pearson's correlation coefficient and Cramer's contingency coefficient. [1]. The correlation coefficient shows an important positive correlation between density and tendency with a value of 0.54. There is also a significant, positive correlation between Randic connectivity and tendency with a value of 0.50. Cramer's contingency coefficient of 0.55 shows a significant, positive correlation between high density, low Randic connectivity, and high centralization as well as a trend evoked by the opinion leader.

8. CONCLUSIONS

The Internet is changing more and more from a mere source of information to an interactive platform. Customers form social networks where they discuss products and influence each other's opinions. For companies, it is of great importance to learn about opinions on their products in order to assess chances and risks. A manual analysis is only possible on a very limited scale. An automated computer supported analysis is necessary given the large number of virtual communities with huge amounts of postings. A new approach for analyzing opinions in forums enables an automated extraction of opinions and relationships from forums and identification of opinion leaders and opinion trends. The approach is realized by combining text mining and social network analysis. Thus existing positive opinions can be reinforced via marketing measures. Negative opinions can be counteracted in time e.g. by product improvements or appropriate marketing measures.

The approach can be considered as a basic concept for analyzing opinions in social networks. A future task is the expansion of the validation database which should also include forums about different product categories. Moreover, a dynamical network analysis can be added to the static social network analysis. Snapshots of social networks at different, consecutive points in time can be taken and the changes can be analyzed. Thus, opinion changers can be identified on user level and changes in opinion trend can be detected on network level.

9. REFERENCES

[1] Bamberg, G., Baur, F., and Rapp, M. 2007 Statistik, Oldenburg Verlag, München.

[2] Chang, C. L., Chen, D. Y., and Chuang, T. R., 2002 Browsing Newsgroups with a Social Network Analyzer, in: Proceedings of the Sixth International Conference on Information Visualization, London.

[3] Cortes, C.; Vapnik, V. N. 1995 Support Vector Networks, In Machine Learning, Vol. 20 (1995), 273-297.

[4] Dave, K., Lawrence, S., and Pennock, D. M. 2003 Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, In Proceedings of the Twelfth International Conference on World Wide Web, ACM Press, Budapest, 519–528.

[5] Gomez, V., Kaltenbrunner, A., and Lopez, V. 2008 Statistical Analysis of the Social Network and Discussion Threads in Slashdot, In Proceedings of the International World Wide Web Conference, ACM Press, Beijing.

[6] Goyal A, Bonchi F., and Lakshmanan L.V. 2008 Discovering leaders from community actions. In Proceedings of the International Conference on Information and Knowledge Management, Napa Valley, ACM:

[7] Java A., Kolari P., Finin T., Oates T. 2006 Modeling the spread of influence on the blogosphere, In Proceedings of the 15th International Conference on WWW, Edinburgh.

[8] Kao, A.; Poteet, S. 2007 Overview, In Kao, A.; Poteet, S. R. (eds.), Natural Language Processing and Text Mining, Springer Verlag, London, 1-7.

[9] Liu, B. Hu, M., Cheng, J. 2005 Opinion Observer: Analyzing and Comparing Opinions on the Web. In Proceedings of the 14th international conference on WWW, ACM Press, New York, 342-351.

[10] Popescu, A. M., Etzioni, O. 2007 Extracting Product Features and Opinions from Reviews. In: A. Kao und S. R. Poteet (Eds.): Natural Language Processing and Text Mining, Springer Verlag, London, 9-28.

[11] Pang, P, Lee, L, and Vaithyanathan, S. 2002 Thumps up? Sentiment Classification using Machine Learning Techniques, Proceedings of the Conference on Empirical Methods in Natural Language Processing, ACM, 79-86.

[12] Randic, M., On Characterization of Molecular Branching, In Journal of the American Chemical Society, vol. 97, 1975, 6609-6615.

[13] Rogers, E: M. 2003 Diffusion of Innovations. Free Press, New York.

[14] Scott, J. 2007 Social Network Analysis, a Handbook, Sage Publications, London.

[15] Song X., Chi Y., Hino K., Tseng B.L 2007 Identifying opinion leaders in the blogosphere. In Proceedings of 16th ACM Conference on Information and Knowledge Management, Lisboa, 971–974.

[16] Wassermann, S.; Faust, K. 1999Social Network Analysis – Methods and Applications. Cambridge University Press, Cambridge,.

[17] Weiss, S. M.; Indurkhya, N.; Zhang, T.; Damerau, F. J.: Text Mining – Predictive Methods for Analyzing Unstructured Information, Springer Verlag, New York 2005.

[18] Welser, H. T.; Gleave, E.; Fisher, D.; Smith, M., Visualizing the Signatures of Social Roles in Online Discussion Groups, in: Journal of Social Structure, Vol. 8, 2007.