Swarm Intelligence for Analyzing Opinions in Online Communities

Carolin Kaiser, Johannes Kröckel, Freimut Bodendorf
Department of Information Systems, University of Erlangen-Nuremberg, Germany
{carolin.kaiser, johannes.kroeckel, freimut.bodendorf}@wiso.uni-erlangen.de

Abstract

Web 2.0 platforms change the collaboration within online communities. A new way of organizing and opinion exchanging derives from increased social interactions and networking among community members. These members join together in self-organizing groups where opinions are forming by social swarming. Explaining and predicting the evolutionary process of opinion formation by social swarming is not only a powerful instrument for opinion research but also a great challenge. A new approach is presented which enables the recognition of opinions of swarm members and the analysis of opinion formation in the overall swarm by combining methods from text mining and swarm intelligence. The concept is illustrated by an example.

1. Introduction

Today’s Internet is becoming increasingly interactive and networked. Web 2.0 platforms such as wikis, weblogs and social networks empower the people and change their way of organizing within communities. Without any central guidance self-organizing groups arise in which members collaborate by following simple rules of communication and interaction. This new form of collaboration emerging from collective intelligent behavior changes the process of information sharing and opinion formation [36]. In contrast to the top-down approach of spreading information by classic media, opinions are exchanged and formed in the bottom-up approach of Web 2.0 by social swarming [28].

Understanding the process of opinion formation in social swarms provides great potential for opinion research. Opinion trends in swarms can be predicted or might even be manipulated. A new approach is introduced which identifies the opinions of single swarm members by using text mining techniques and analyzes opinion formation with respect to the underlying swarming behavior by applying methods associated with swarm intelligence. Swarm intelligence is a discipline of artificial intelligence and aims at developing algorithms based on the swarming behavior of social insects [9]. A new algorithm inspired by the commonly known ant colony optimization metaheuristic is presented which allows the prediction of opinion forming in human swarms. The approach is illustrated by an exemplary online community in which opinions on computer games are exchanged and discussed.

2. Related Work

Text mining enables the detection of patterns in texts [26]. The mining process transforms unstructured text into meaningful structured data. Text mining comprises four main fields of research: information retrieval, text classification, text clustering, and information extraction [48]. Traditional research in these fields concentrated on mining facts, whereas recent research focuses on mining opinions. Many papers deal with the detection of opinions on the Internet. For example, Dave et al. [14], Liu et al. [29] as well as Popescu and Etzioni [35] outline systems which analyze customer opinions in product reviews. There is also some research on monitoring opinions on the Internet. Holzmann [22] as well as Tong and Yager [45] describe systems which detect and keep track of Web opinions over a period of time. However, the development of opinion formation over time is not analyzed on the basis of the interactions between users which have a high influence on opinion evolution.

Social network analysis examines relationships among interacting users [47]. It comprises an individual analysis for studying single users’ habits and a collective analysis for observing the activity structure of the entire network [40]. There are many papers which apply social network analysis to reveal opinions in online communities. For example, Chang et al. [12] identify leading authors and provocative articles in newsgroups. Glance et al. [18] analyze clusters of persons discussing certain topics in forums. Agrawal et al. [4] divide authors of newsgroups into opposite camps depending on their opinion. Adamic et al. [3] study the correlation between the interests of forum users and their behavior in replying. Gloor et al. [19] investigate the evolution of online social networks over
time with respect to communication flow and content. Discussion boards in online communities can be considered as social networks, where community members are linked to one another by communication relationships. However, our approach deals with a different perspective on the relationships among community members. Instead of considering the network structure built by the relationships, the swarming process arising from the relationships is examined.

Agent-based models provide computational models to simulate how interactions among individuals lead to the emergence of a collective organization [20]. It is based on the findings that the behavior of a group cannot be explained by the independent behavior of individuals ([5], [38]). Interaction among the individuals results in a high-level organization which crystallizes without knowledge of the individuals. Often, they are even unable to estimate their own actions and feelings ([30], [11]). Agent-based models are applied to explain the collective behavior in the fields of organization, contagion and cooperation [20]. For example, Schelling [41] simulates how individual movements according to neighborhood similarity lead to the formation of segregated groups. Axelrod [6] shows how the individuals’ adoption of neighborhood traits brings forth global polarization. Berger and Heath [8] study how ideas are spread throughout discussion boards depending on environmental cues. Rosenkopf und Abrahamson [37] demonstrate how innovations diffuse across organizations with respect to reputational and informational influences. Sakamoto et al. [39] examine how human choices develop under varying social influences within online communities. The approach presented in this paper can be considered as an agent-based model for simulating opinion formation by social swarming. The process of opinion formation is also based on the interactions among individuals and the orientation towards neighboring discussion partners. However, our approach differs in aim and method. Our aim is to address the phenomenon of opinion formation by employing an ant algorithm from swarm intelligence.

Swarm intelligence provides problem-solving algorithms which are inspired by the swarming behavior of animals and which can be transcribed to human behavior. Swarm intelligence consists of two main meta-algorithms: Ant colony optimization and particle swarm optimization [9]. Ant colony optimization is inspired by the foraging behavior of ants. It enables the incremental solving of discrete optimization problems. Solutions are found by digital ants following pheromone trails which indicate the quality of a tentative solution. Particle swarm optimization imitates the behavior of birds searching for food. It allows population-based solving of continuous optimization problems. Solutions are represented by birds (particles) which are flying around the solution space by following the best birds so far. In the past, algorithms of these two types have been developed for solving traditional optimization problems such as route planning and time scheduling. Nowadays new challenges such as data mining and Web mining are being faced. A lot of papers describe ant colony optimization algorithms and particle swarm optimization algorithms for data classification [2], data clustering [15], feature selection [23] and fuzzy-rule induction [16]. Some research deals with the development and application of such algorithms to Web mining. Abraham and Ramos [1] present a cluster algorithm for detecting Web usage patterns. Ujin and Bentley [46] propose an algorithm which guides online shoppers and visitors to entertaining Web sites by individual recommendations according to their preferences. Jensen [23] describes an algorithm for classifying Web pages based on their subject. Palotai et al. [33] introduce an algorithm which finds news on the Internet. However, so far there are no algorithms for analyzing the swarming behavior of Web users during the evolutionary process of opinion formation.

3. Approach

Colonies of social insects like ants, bees, wasps, and termites are able to achieve complex tasks such as picking up material or finding food by means of cooperation. The collective intelligent behavior emerging from relatively simple interactions between colony members is called swarm intelligence [10].

In general terms, swarm intelligence can be defined as a phenomenon which arises from the social structure of interacting agents over a period of time if the sum of the problems solved collectively is higher than the sum of the problems solved individually [44]. Two preconditions must be fulfilled in order for swarm intelligence to develop: The agents must interact with each other and must be capable of problem-solving [44]. Characteristics of emerging swarm intelligence are self-organization, robustness, and flexibility [10]. The members of the swarm interact without supervision or centralized control. The swarm is capable of achieving its task even if some members fail and is able to adapt to a changing environment.

This phenomenon of collective intelligence cannot only be observed in colonies of social insects but also in collaborative groups of humans. By exchanging experiences, correcting mistakes, and inspiring one another, collaborative groups are in a better position of solving problems than individuals [21]. Collaboration can be understood as an act of collective information
processing [42]. Discussion is one of its basic forms [44]. Due to the collective process of exchanging information and opinions during a discussion, the sum of the combined knowledge of the community becomes more valuable than the sum of the knowledge of all individual community members [28]. Web 2.0 platforms increase this effect of knowledge enhancement [24]. A wider range of people can connect more easily and more rapidly to reach a common opinion in an online discussion.

The Web also provides an advantage for opinion research. The process of opinion formation can be traced by applying mining techniques. A new approach based on text mining and swarm intelligence is presented which is capable of analyzing the evolutionary process of opinion formation by social swarming. Text mining enables the recognition of opinions of single community members. An algorithm associated with swarm intelligence, especially the commonly known ant colony metaheuristic, allows the prediction of the opinion trend during the collective intelligent process of opinion formation in online communities.

4. Opinion Mining

4.1 Method

The goal of opinion mining is to recognize the attitude of single swarm members towards an object mentioned in postings. Attitudes are classified according to their polarity as “positive”, “negative”, or “no opinion”.

Opinion mining in texts involves two steps, the extraction of features from the text and the application of a learning algorithm to identify the polarity of the text [48].

The extraction of features comprises a combined linguistic and statistical analysis. First, a posting is decomposed into single words. After removing unimportant words (e.g. “the”), the remaining words are reduced to their stem and their frequency is calculated. Those word stems which are especially distinctive for each of the classes (meaning that they appear often in one class but rarely in others) are used as the features of the postings.

Based on the extracted features, the postings are classified according to their polarity by a learning algorithm. In general, machine learning provides three categories of learning algorithms [17]. Supervised learning algorithms use input and desired output data for learning to produce the correct output data. Reinforcement learning algorithms learn to generate actions by getting rewards and punishments [25]. Unsupervised learning algorithms, in contrast, receive no feedback and find patterns within the input data which can be used to produce output data [17]. Supervised and unsupervised learning methods are often used for text classification [27]. Supervised learning requires more effort for pre-classifying texts (desired output) but enables enhanced classification results. It is, therefore, employed in this approach.

Different supervised learning methods such as Naïve Bayes or Maximum Entropy can be used for text classification [32]. Support Vector Machines [13] are applied because of their ability to process a high number of features and their success in related projects [32]. Their input are sample data records which consist of various forum postings with their features and classes. By analyzing the sample data Support Vector Machines learn the parameters of a rule which classifies the postings best. The rule allows a binary classification. If there are three classes, three rules must be learned: “positive” versus “not positive”, “negative” versus “not negative”, and “opinion” versus “no opinion”. A posting will be assigned to the class which has the highest probability. In the simple two-dimensional case the rule can be depicted as a straight line (linear rule). Postings lying on one side of the line belong to the first class and those lying on the other side belong to the second class. Figure 1 shows how a straight line divides the postings into the classes ‘positive’ and ‘not positive’ depending on the mentioning of the word stems ‘love’ and ‘best’.

4.2 Application

For validation, opinions posted to the German online gaming community of Gamestar.de were classified. Gamestar.de is the online platform of Europe’s most popular magazine for PC and video games. 4010 postings were extracted and assigned to the three classes “positive”, “negative”, and “no opinion” by a human annotator.

In order to examine the classification results a stratified ten-fold cross validation is applied. This means that all postings are divided into ten equally
sized parts containing the same proportions of class labels. There are ten validation loops. In each loop nine parts are used for learning the classification rules and the remaining part for testing the classification rules learned. After ten runs the average precision and recall are calculated. While precision describes how many of the recognized opinions are correct, recall shows how many of the opinions are really recognized. The results of validation are shown in Table 1. They indicate that learning was more successful for negative opinions and no opinions than for positive opinions. Lessons learned from misclassification show that quite often postings are not recognized as positive if they contain several negative arguments or neutral information but a positive introduction or a positive conclusion. This problem is planned to be solved by attaching more weight to the words at the end and the beginning of the postings.

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>62.96%</td>
<td>62.52%</td>
</tr>
<tr>
<td>Negative</td>
<td>86.35%</td>
<td>86.05%</td>
</tr>
<tr>
<td>No Opinion</td>
<td>81.65%</td>
<td>81.07%</td>
</tr>
</tbody>
</table>

5. Opinion Analysis

5.1 Method

The aim of opinion analysis is to gain a better understanding of opinion formation in social swarms. With this knowledge opinion trends can be predicted and the process of opinion formation might even be manipulated.

Opinion analysis is inspired by the collective intelligent behavior of living ants finding the shortest path between their nest and their source of food. This intelligent behavior emerges from the ants’ indirect manner of communicating by leaving and following pheromone trails in their environment – a phenomenon called stigmery [31]. While ants are moving about they drop chemical substances called pheromones on their paths. The more the same path is frequented, the more the pheromone intensity increases and the more likely this path will be followed by other ants. If the same path is only followed by a few ants, the pheromone intensity of the path decreases due to evaporation. As a result of this feedback loop, the probability that the path will be followed by an ant depends on the number of ants having taken this path before.

The ants’ behavior seems to resemble in some ways the behavior of human swarming within online communities and can be used as a simplified model to simulate the process of opinion formation. Members of online communities communicate indirectly with each other by posting messages to a discussion thread. In their postings they can express positive or negative opinions. The more messages of the same opinion are posted, the more other people are attracted by this opinion and the more likely they are to follow this opinion.

The collective intelligent behavior of ant colonies is also the basic idea of the ant colony optimization metaheuristic, from which an algorithm for simulating the process of opinion formation can be derived. In ant colony optimization algorithms, possible solutions for a given problem are represented by paths [34]. If a path is followed by an ant, a certain amount of pheromones is deposited on it depending on the quality of the solution. Evaporation gradually decreases the pheromone amounts on those paths which are not traversed frequently. This means that the corresponding solution is not particularly appreciated.

When simulating the process of opinion formation the problem is to predict the polarity of the next posted opinion in a discussion thread. Possible solutions are represented by two different paths: one for positive and one for negative opinions. An ant predicts the next posted opinion by following the corresponding path and drops a certain amount of pheromones on this path depending on the correctness of the prediction. Evaporation depends on the sequence of postings in the thread and leads to a reduction of the pheromone amount on the path of the less frequently mentioned opinion. The ant is more likely to predict the opinion class whose corresponding path has a higher amount of pheromones. Figure 2 illustrates this prediction model derived from the ant colony optimization metaheuristic.

![Figure 2: Prediction model](image-url)
A heuristic function which evaluates the quality of the solution found by an ant
A rule for pheromone updating which describes how to reinforce pheromones on paths
A rule for pheromone evaporation which specifies how pheromones on paths diminish over time
A decision function which finds solutions by considering the value of the heuristic function and the amount of pheromones on paths

In order to develop an algorithm based on ant colony optimization for predicting opinions in online swarms, these components must be specified and integrated into a procedure.

Figure 3 shows the flowchart of the developed procedure. First, all variables are initialized. The pheromone values of both paths representing the positive and negative opinions are given equal amounts of pheromones. While the discussion is going on, the opinion trend in terms of the next posted opinion class is predicted by an ant. The ant predicts the opinion class by choosing a path according to the decision function. As soon as the next message is posted to the thread its content is checked. If no opinion is expressed in the posting evaporation takes place. However, if the posting contains an opinion the correctness of the ant’s prediction is evaluated. In case the predicted opinion differs from the posted opinion, the average error ratio is increased. Otherwise the average error ratio is decreased. Thereafter, the heuristic values of the decision function are adjusted depending on the dynamics of the opinion discussion. In addition, the pheromone value of the path representing the predicted opinion is updated. Finally, evaporation takes place which decreases the pheromone values of all paths.

According to the procedure, the functions for decision making, pheromone updating and evaporation have to be defined.

The decision function determines the predicted opinion at time $i$ by comparing the weighted sum of pheromones on the positive ($\tau^p_i$) and negative path ($\tau^n_i$). It is implemented as a signum function based on the difference of the weighted pheromone sums of both paths (positive and negative opinions). By this means the opinion of the path with the highest weighted sum of pheromones is predicted.

$$d_i = sgn \left( (\eta^n_i \tau^n_i) - (\eta^p_i \tau^p_i) \right)$$

If the decision function yields the value 1, a positive opinion is predicted by the ant. In the case of value -1 a negative opinion is forecasted. The pheromone sums of both paths are weighted by the iteratively computed heuristic values $\eta^n_i$ and $\eta^p_i$. The heuristic values are incremented by a certain factor $x$ if the actual opinion ($o_i$) equals the previous one ($o_{i-1}$) and decremented by $x$ if both opinions are different. By doing so, the dynamics of opinion changes during opinion prediction is taken into account.

$$\eta_i = \begin{cases} 
\eta_{i-1} + x & \text{if } o_i = o_{i-1}, \\
\eta_{i-1} - x & \text{if } o_i \neq o_{i-1}, \eta_i \in ]0; 1]\n\end{cases}$$

If a lot of consecutive messages with no opinion are posted to the thread, the influence of the last posted opinions on the future opinion trend becomes insignificant. In such a case, the prediction should be based on all opinions posted to the thread. In the ant algorithm evaporation leads to a rapid diminution in pheromone values of the negative and positive path if there is a sequence of postings without opinions. A minimum-rule derived from the Max-Min Ant System of Stützle and Hoos [43] is implemented to change the prediction basis in this case. According to this rule, the decision function predicts the future opinion trend based on the opinion class most frequently mentioned in the thread, if the pheromone values of both paths fall below a minimum value. The rule is also inspired by the biological archetype of behavior of real ants. If the amount of pheromones on a path cannot be smelled any more, the ants rely on their instinct when choosing a path.

In order to reinforce predicted opinions the pheromone values of the corresponding paths are updated. Pheromone updating is realized by adding a predefined amount of pheromones $\rho$ to the current pheromone value of the selected path.

---

**Figure 3: Flowchart**
\[ \tau_i = \tau_i + \rho \]

Since the opinion trend in a forum thread can change from time to time more weight should be added to recent opinions than to past opinions. Pheromone evaporation enables this weighting. It is realized by multiplying the pheromone value \( \tau_i \) with a certain factor \( e < 1 \).

\[ \tau_{i+1} = \tau_i e \]

5.2 Application

In order to validate the ant algorithm for opinion prediction the German online gaming community Gamestar.de was analyzed. Three threads in which community members discussed their opinions on games were extracted. All opinions mentioned in the postings were classified as ‘positive’, ‘negative’ or ‘no opinion’.

Before applying the ant algorithm the parameters of the functions involved must be determined. Tests with different combinations of parameters revealed the following best set:

- Pheromone update \( \rho \): 0.5
- Heuristic factor \( x \): 0.3
- Minimum threshold: 0.0001
- Evaporation rate \( e \): 0.8

Based on this set of parameters, the average error ratio is measured over the entire period of time. It is calculated as the fraction of all incorrectly predicted opinions to all opinions. Table 2 depicts average ratios for the three discussion threads. The low error rate of 17.2% for the thread discussing the game “Dead Space” indicates high prediction accuracy. The error rate for the game “Far Cry 2” is higher indicating that prediction was less successful.

<table>
<thead>
<tr>
<th>Discussion thread on game</th>
<th>Amount of postings</th>
<th>Average error ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dead Space</td>
<td>1242</td>
<td>17.2 %</td>
</tr>
<tr>
<td>Fallout 3</td>
<td>3639</td>
<td>26.7 %</td>
</tr>
<tr>
<td>Far Cry 2</td>
<td>460</td>
<td>41.3 %</td>
</tr>
</tbody>
</table>

Besides the average error ratio of the entire period, the development of the error ratio over time is important as well. It reveals whether the decision function enables incremental learning.

Figure 4 depicts the error curve associated with the discussion on the game “Fallout 3”. The descending error curve indicates a successful learning process. This effect results from the heuristic values which adapt incrementally to the dynamics of the discussion.

![Figure 4: Error curve of game “Fallout 3”](image)

The necessity of the minimum-rule becomes apparent when looking at the development of the pheromone amounts on the positive and negative paths over time. For example, Figure 5 shows the pheromone curve associated with the discussion on the game “Fallout 3”. 12.5% of the pheromone values fell below the minimum threshold so that prediction was based on the most frequent opinion class.

In addition, the pheromone development shows the underlying swarming behavior of opinion formation. The amounts of pheromones decrease over time which indicates a leveling in discussion. At the same time opinions are getting more homogeneous. During discussion a clear opinion trend emerges from the initially differing opinions of the swarm members. At the end of the discussion there are only a few opinions differing from the overall opinion trend which have less influence. This can be interpreted as a sign of robustness of the swarming behavior.

Application results of the ant algorithm show that prediction is more successful for discussion threads in which the length of the sequences of equal opinions vary to a high degree (see Table 3).

<table>
<thead>
<tr>
<th>Game</th>
<th>Error ratio</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dead Space</td>
<td>17.2%</td>
<td>19.5</td>
</tr>
<tr>
<td>Fallout 3</td>
<td>26.7%</td>
<td>17.2</td>
</tr>
<tr>
<td>Far Cry 2</td>
<td>41.3%</td>
<td>3.3</td>
</tr>
</tbody>
</table>

The discussion threads of the games “Dead Space” and “Fallout 3”, for which prediction was successful, contain many sequences with equal opinions of different length (see Figure 6). In contrast, the
discussion thread of the game “Far Cry 2” only consists of short sequences of equal opinions which results in a less successful prediction. More threads are being prepared for testing the validation of this result.

6. Conclusion

A new form of collaboration arises within Web 2.0 communities. Increased social interactions and networking among community members change their way of self-organizing and exchanging opinions. Internet users join online communities where opinions are forming by social swarming. A new approach is presented which allows the identification of opinions of swarm members and the analysis of opinion formation within the swarm. By applying text mining techniques opinions of single swarm members are classified according to their polarity. With the aid of an adjusted ant algorithm, inspired by the commonly known ant colony optimization metaheuristic, the opinion trend of the swarm is predicted.

This approach represents a basic concept for analyzing the evolutionary process of opinion formation according to the underlying swarming behavior of online community members. The first validation yielded encouraging results. However, a larger database is needed to enable a more sound validation. Therefore, the dataset is being enlarged by additional discussion threads dealing with opinions on other topics. Moreover, validation should be improved by comparing our prediction model with other models. Starting point will be the simple model of preferential attachment introduced by Barabasi and Albert [7]. According to this concept, the predicted opinion only depends on the frequency of all mentioned opinions in the thread. This means that our minimum-rule would be used as prediction function and the effects of
pheromone accumulation and evaporation would be ignored.

Future work will also concentrate on expanding the functions of the developed ant algorithm. For example, the heuristic function should not only consider the polarity of opinions but also the intensity of opinions. More importance will be attached to strong opinions than to weak opinions. In addition, the reputation of the authors who posted the opinions should also be incorporated in the heuristic function. Inspired by the findings of Rosenkopf and Abrahamson [37] as well as Skamoto et al. [39], more weight should be given to opinions expressed by authors having a better reputation. Besides, evaporation will not only depend on the sequence of the postings but also on the time-lags of the postings. Long time-lags between two postings should lead to a higher rate of evaporation than short time-lags. These enhancements are expected to augment the findings based on swarm intelligence which has already been proven to be significant in online communities.

7. References


