Association for Information Systems
AIS Electronic Library (AISeL)

ECIS 2012 Proceedings

European Conference on Information Systems (ECIS)

5-2-2012

REDUCING INFORMATION OVERLOAD: DESIGN AND EVALUATION OF FILTERING & RANKING ALGORITHMS FOR SOCIAL NETWORKING SITES

Ksenia Koroleva
Humboldt-University Berlin

Antonio José Bolufé Röhler
University of Havana, Cuba

REDUCING INFORMATION OVERLOAD: DESIGN AND EVALUATION OF FILTERING & RANKING ALGORITHMS FOR SOCIAL NETWORKING SITES

Koroleva, Ksenia, Humboldt-University Berlin, Unter den Linden 6, Berlin, Germany, koroleks@wiwi.hu-berlin.de

Bolufe-Röhler, Antonio, University of Havana, Colina Universitaria, Havana, Cuba, bolufe@matcom.uh.cu

Abstract

Recognizing the detrimental impact of information overload on user participation, in this paper we design and evaluate several algorithms to filter and rank the information on Social Networking Sites (SNS). As a first step we identify the factors that impact user evaluations of information shared through these networks in a set of regression analyses. Second, we use a Neural Network algorithm to predict three dimensions of user evaluations: affective, cognitive and instrumental value of information shared. Moreover, we design algorithms that allow not only to filter out the irrelevant information, but also rank the information on SNS in order of its relevance. As a result, the filtering algorithm accurately predicts in 73% of the cases, whereas for more than 70% of the users the individual ranking accuracy lies over 70%. The designed algorithms can be implemented by SNS providers in order to present users with more relevant and better structured information.

Keywords: Information Overload, Information Filtering, Information Ranking, Neural Network Algorithm, Social Networking Sites, Facebook Newsfeed

1 INTRODUCTION

The new generation of CMC – Social Networking Sites (SNS) (Boyd and Ellison, 2008) are widely utilized for information generation and exchange. For example, on Facebook users share ca. 30 billion pieces of content every month (Facebook, 2011a). Users rely on the information exchanged through these networks for news (Glynn et al., 2012), purchase decisions and other personal issues (Lampe et al., 2012), relationship development with friends (Köbler et al. 2010) and even the benefits of social capital (Ellison, 2007). However, due to the increasing amount and variegated quality, information sharing is bounded by the problem of *information overload* (Eppler and Mengis, 2004). Usual consequences of information overload include confusion, stress, anxiety (Schick et al., 1990), as well as diminishing decision quality (Chen et al., 2009a). On SNS, users feel dissatisfied and thus may reduce their activity (Koroleva et al., 2010), which is detrimental for the longevity of SNS providers.

Recognizing this problem, on most SNS specific algorithms are responsible for determining the most relevant information for the user. Scattered insights suggest that on Facebook algorithms prefer posts which have received more feedback, as well as from those friends with whom users previously interacted (TechCrunch, 2010). However, algorithms rarely consider such factors as length of post or friend posting frequency which might cause information overload on SNS (Koroleva et al., 2010). At the same time, user feedback might not impact evaluations in the same way: affirmations are positively, whereas comments – rather negatively correlated with information value (Koroleva et al. 2011a). Moreover, preferring information from those with whom the user interacts often online is not optimal. First, users may prefer other means to communicate with their close friends (Vitak et al., 2011). Second, the value of recommender systems lies in discovering new content outside of the user's usual social circle (Chen et al., 2010). Designing an optimal information filtering is thus a difficult task and therefore an interesting venue for research.

As SNS users are largely dissatisfied with the existing information filtering algorithms (Tonkelowitz, 2011) and unable to cope with information overload themselves (Koroleva et al., 2010), more insights are required to design filtering algorithms for SNS. First of all we need to know which factors are driving information relevance on SNS, as using all the data available on the network may reduce the speed and efficiency of a filtering algorithm. Therefore, in our paper we first identify which factors significantly impact information value on SNS. We then include only those as input factors into a Neural Network algorithm in order to predict different classes of information relevance for the users. Moreover, most efforts so far have been concentrated on filtering out the irrelevant information, rather than ranking the information in order of importance. Our set-up allows us thus to additionally perform individual ranking of the information. In this way we give practical suggestions for the design of filtering and ranking algorithms that can be used by SNS and other social media providers.

2 RELATED WORK

Recommending content to users has always been an important task of the information systems, which requires both filtering and ranking of information. Several approaches to information filtering exist: i) *collaborative filtering* based on the similarity of preferences between users widely used in e-commerce to recommend products (Konstan et al., 1997); ii) *content relevance* approach matching the topic interests of the user and the content vector of presented information used to suggest news items (Pazzani et al., 1996); iii) *social voting* based on the frequency of mentioning or rating of information by other users (Hill and Terveen, 1996); and iv) *social matching* system that recommends people to each other on SNS (Chen et al., 2011).

Useful insights for the design of filtering algorithms are provided in the studies on microblogging applications such as Twitter, where the problem of information overload is even more acute than on Facebook as the average frequency of posting is even higher (Chen et al., 2010). The best performing algorithm which selects posts from the outer circle of followees (as opposed to direct followees) and

ranks them by both content relevance and social voting achieves an accuracy[1] of ca. 72% (Chen et al., 2009b). In the follow-up study, authors extend the input factors to account for thread length and tie strength which further improve the accuracy of the algorithm (Chen et al., 2010).

Similar efforts have been taken to recommend information on SNS, such as Facebook. A social matching system developed for SNS to recommend friends to each other uses the available social network information and matching of user-generated content (Chen et al., 2011). The easiest heuristic is to recommend friends of friends, however more complicated systems are usually employed to decide which friends to recommend first. To predict the importance of Newsfeed posts, Paek et al. (2010) use Support Vector Machine (SVM) algorithm and use all possible factors as input (in total over 50), including communication and post characteristics, message text and corpus features, as well as shared background information (Paek et al., 2010). In a binary classification the highest prediction accuracy of the algorithm lies at 69.7%. The most important factors are tie strength with a friend (implied using a myriad of indicators) and content relevance (matched using topic vectors): if any of these factors is not considered, classification accuracy drops to 63%.

In contrast to presented studies, we assume that already with fewer factors than those used by Paek et al. (2010) a satisfactory level of prediction accuracy can be achieved. Therefore, we explore what are the necessary input factors to filter the posts on SNS. To our knowledge, no study so far has conceptualized the inclusion of factors into a filtering algorithm, but simply used all data. Although social network data is usually available on the network, this results in large amounts of data to process for the algorithm which may reduce its efficiency, especially if an on-line ranking system is implemented for SNS. Our first research question is: *What are the necessary input factors to achieve a satisfactory level of filtering accuracy of the information on SNS?*

Moreover, previous authors have implemented binary classifications as a type of a spam filter, but the recent developments show that filtering out the irrelevant posts does not solve the problem of information overload. Therefore our goal is not only to filter the posts, but also to rank the posts for the user in order of their relevance. Ranking of posts allows users to determine which amount of information they want to process themselves, and thus reduces the feelings of information overload. Ranking requires a finer-grained classification of posts into classes of relevance, which is only possible if a moderate amount of data is used (Paek et al., 2010). Our second research question is: *Which ranking accuracy of the information on SNS can be achieved?*

Studies confirm the importance of tie strength in determining the relevance of the information on SNS (Morris et al., 2010). However, the information about the underlying tie strength is not available on the network and can only be inferred. Gilbert and Karahalios (2009) gather an extensive amount of social network data, including such complex characteristics as inbox thread depth or the frequency of positive and negative words exchanged between users, in order to distinguish between strong and weak ties on Facebook, achieving an accuracy of over 85%. Aiming to reduce the amount of data used, a secondary question of our paper is: *What information can best predict tie strength on SNS?*

3 CONCEPTUAL DESIGN

3.1 Identifying the Input Factors

In our study users evaluate the information on SNS along several dimensions, which reflect the affective and cognitive components of attitude widely used by authors in psychology (Voss et al., 2003) as well as in IS (e.g. Yang and Yoo, 2004). Cognitive attitude refers to evaluations of the attitude object and the qualities it possesses, whereas affective focuses on how much the person likes the object and is emotionally attached to it (Ajzen, 2005). In our study *affective value* of information

[1] - Classification accuracy measures the frequency with which the algorithm correctly classifies an item.

shows how much the user likes the information shared on SNS, *cognitive value* reflects the level of interest of this person in the information shared, whereas *instrumental value* exemplifies the utilitarian dimension of how useful the information is to the user. We want to explore which factors pertaining to information on SNS impact these dimensions of information value. The factors can either characterize the *information* exchanged, such as the number of affirmations and comments it received, or characteristics of the *relationship* with the user who shared the information (Koroleva et al. 2010). Moreover, these factors can be *objective*, the same for all the "receivers" of information, (e.g. post type or word count), as well as *subjective*, unique to a particular relationship between the "sender" and the "receiver" of information (e.g. understandability or similarity of interests).

Objective *information characteristics* include word count, comments and affirmations, and type of post, whereas understandability can only be measured subjectively. First, we hypothesize a u-shape relationship between the *word count* and user evaluations, as each additional word possesses marginally lower value while incurring the same processing costs (Schroder et al., 1967), resulting in information overload (Schneider, 1987), corroborated in numerous studies (Chen et al. 2009). Second, the number *affirmations* by summarizing the positively directed evaluations from others increases information value (Koroleva et al. 2011a). As *comments* are not necessarily positively directed and may cause information overload, they trigger rather negative evaluations (Koroleva et al. 2011a). In fact, thread length reduces the accuracy of filtering when other features, such as topic interest, are absent (Chen et al., 2009a). Third, we control for the post type, as *photos* or *links* might be differently perceived by users. Finally, *understandability* should correlate positively with information value (Koroleva et al., 2011b).

Mainly subjective *relationship characteristics* concern the underlying tie strength between the "poster" and the "receiver" of information. In fact, users prefer information from their strong rather than weak ties on SNS (Koroleva et al., 2011a), although weak ties are known to possess more potential in these environments (boyd and Ellisson, 2008). Tie strength, however, can only be implied using the information available on the platform (Gilbert and Karahalios, 2009). First, *similarity* can serve as a good indicator of tie strength, as users tend to have similar interests with those they are close with. Shared interests or shared background information are often used to filter information on social applications (Paek et al., 2010). Second, *communication intensity*, reflected in public and private communication as well as passive following, positively impacts information value on SNS (Koroleva et al., 2011b) and has been used to imply tie strength (Paek et al., 2010). Third, *posting frequency* is often used by filtering systems to estimate the weight of information (Chen et al., 2010). However, high posting frequency, by overloading other users with information, may lead to negative evaluations (Jones et al. 2008; Koroleva et al., 2011b). Finally, the *location* of the user has been identified as an important determinant of information relevance (Koroleva et al., 2010).

3.2 Survey Design and Scale Development

The data for this study was collected from Facebook users with the help of a specially programmed application. In order to take part, users had to log-in to their Facebook accounts and install the application, after which they were asked for permission to access 6 posts on their Newsfeed in real time. The posts were retrieved from the Facebook database using Facebook query language (structure similar to SQL), which is an API (application programming interface) provided by Facebook (Facebook, 2011a). Out of all available posts on the user's Newsfeed over the last 72 hours, up to 6 posts of different types were randomly selected and presented for evaluation together with an integrated survey tool. In total, 158 people completed the survey. As each user evaluated up to 6 posts, 948 observations were obtained. After deleting the outliers (those with more than 50 comments or affirmations), a total of 925 observations from 154 users were obtained.

Most of the items used in the integrated survey had to be adapted to the Facebook context. First, respondents were asked to evaluate the presented information, which was measured along three different dimensions (all on a 6pt ordinal scale): affective value (*dislike very much – like very much*),

cognitive value (*very boring – very interesting*), and instrumental value (*very useless – very useful*). Moreover, the objective data was collected by the application automatically at the time the information was accessed: post type (status update, photo or link), word count, comments and affirmations. All other subjective variables were additionally elicited from users, including communication intensity (private, public and following) and posting frequency (all on 5pt, *almost never – almost always*), similarity (5pt, *nothing in common – very much in common*), understandability (3 pt scale, *not at all – very well*), as well as tie strength (5 pt, *don't know at all – very well*). Due to limitations of space, full survey is obtainable from the authors upon request.

4 DETERMINING THE INPUT FACTORS

4.1 Methodology

We use regression analyses in order to explore which factors are significant predictors of the different dimensions of information value on SNS. As users evaluated information on an ordinal scale, we estimate an Ordered Probit regression (Greene, 2000) tailored to use with the dependent variables of this type. Moreover, as each respondent evaluated six different posts, we apply a panel-data specification via the inclusion of user-specific random effects (Buttler and Moffitt, 1982). The dependent variables are the different dimensions of user evaluations – affective, instrumental and cognitive, whereas the independent – the information and relationship characteristics presented in the section "conceptual design". In order to standardize the independent variables (as they were measured on ordinal scales), we create dummy variables, which are equal to one for the high levels of these variables, and zero in all other cases. We determine the split of the variables into high vs. low levels based on the median of the sample. For post type, we explore the impact of links and pictures with respect to status updates. We add word count squared as we hypothesize that the number of words will have an inverted u-shape relationship with user evaluations. The main purpose of our analysis is not to prove certain relationships, but to identify the siginificant factors to include into the algorithm.

Variable	**Affective**		**Cognitive**		**Instrumental**	
Information Characteristics						
Word count	**0.007 (0.00)**	**	0.004 (0.003)		**0.010 (0.003)**	***
Word count squared	-0.000 (0.00)		-0.000 (0.00)		**-0.000 (0.00)**	**
Affirmations	**0.042 (0.01)**	***	**0.032 (0.01)**	***	**0.038 (0.01)**	***
Comments	**-0.016 (0.01)**	**	-0.011 (0.01)		**-0.021 (0.01)**	**
Photos (w.r.t status)	**0.329 (0.11)**	***	**0.274 (0.11)**	***	**0.389 (0.11)**	***
Links (w.r.t status)	0.045 (0.08)		**0.217 (0.08)**	***	**0.483 (0.09)**	***
Understandability	**0.730 (0.08)**	***	**0.694 (0.08)**	***	**0.633 (0.085)**	***
Relationship Characteristics						
Similarity	**0.552 (0.11)**	***	**0.59 (0.11)**	***	**0.463 (0.11)**	***
Public Communication	0.161 (0.13)		**0.267 (0.128)**	**	**0.326 (0.13)**	**
Private Communication	**0.299 (0.13)**	**	0.199 (0.131)		0.087 (0.13)	
Passive Following	**0.479 (0.11)**	***	**0.483 (0.11)**	***	**0.274 (0.11)**	**
Posting Frequency	**-0.171 (0.08)**	**	**-0.159 (0.08)**	**	**-0.164 (0.08)**	**
Location	0.106 (0.09)		0.088 (0.09)		-0.001 (0.092)	
Pseudo-R^2	**10%**		**8.9%**		**7.6%**	

*Table 1. Estimation Results of Ordered Probit Regression (***-1%, **5%, standard error in brackets)*

4.2 Estimation Results

The results presented in the table 1 reveal underlying dynamics between the different dimensions of information value on SNS. For example, the number of affirmations, understandability of information, similarity of interest between users as well as passive following of a user are significantly positively related with all dimensions of information value. Moreover, photos are evaluated significantly better

than status updates, no matter which dimension is considered. On the negative side, posting frequency of other users reduces information value overall. Contrary to expectations, location does not have any significant impact on any dimension of information value. Additionally, affective value increases as the word count and private communication with the user increase. At the same time, users prefer links to status updates as well as posts from those with whom they communicate frequently in public when estimating instrumental and cognitive value of information. What concerns instrumental value, additionally the inverted u-shape relationship with the word count, as well as the negative impact of the number of comments on information value, are proven. To predict different dimensions of information value using the algorithms presented in the following sections, we use solely the factors that have proven significant. According to the pseudo-R^2 measure of fit (MacFadden, 1974) our regression model is better at explaining the affective (pseudo-R^2: 0.10) and the cognitive (pseudo-R^2: 0.089) than the instrumental value of information (pseudo-R^2: 0.076).

5 FILTERING ALGORITHM

5.1 Methodology and Design

In order to filter the posts for the user, we propose to use Neural Networks (NN). NN is a well known method that has been successfully applied to real-world classification problems, similar to information classification on SNS. NN are flexible and robust, which is important when classifying noisy data we are dealing with: either the subjective information provided in surveys or objectively obtained from Facebook. NN allow us not only to do an effective binary classification, but also to classify posts into 3 and 6 classes. As the number of classes increases, it is possible to obtain a more "fine grained" ranking, although the accuracy of the classification is expected to decrease. At the same time, the classification mistakes tend to be less severe (for example, ranking the post as "very useful", when it is, in fact, "slightly useful"), compared to the failure of an accurate binary yes-no classification. Moreover, a fine-grained classification is a first step in the direction of post ranking – which is a more challenging, yet the final goal of classification systems.

For the design of the filtering algorithm, three design dimensions are taken into account (see table 2): the combinations of input factors (2), the different target variables (3) and the different number of classes (3), resulting in total 18 algorithms. Specifically, the experiments were run with two combinations of input variables: one considering only the objective factors, and the other with all the factors, including subjective. Only the significant input factors (results presented in table 1) were included into the algorithm for the corresponding dimension of information value. The algorithm predicts different number of classes, where two classes were obtained by merging the corresponding positive and negative categories, three classes – by merging the following categories: very (-) and quite (-), slightly (-) and slightly (+), quite (+) and very (+).

Design Dimension	**Possible Design Choices**		
Input Factors	objective	all: objective + subjective	
Target Variable	affective	cognitive	instrumental
Number of classes	2 classes	3 classes	6 classes

Table 2. Overview of possible algorithm designs

For the classification, the FeedForward network provided by the Neural Network Toolbox in Matlab was used (Demuth and Beale, 1997). After several tests, the best results were obtained with the following configuration of the network: one hidden layer with ten neurons; a log-sigmoid transfer function for the hidden layer; a linear transfer function for the output layer; a gradient descent with momentum weight and bias learning function; the Levenberg-Marquardt back propagation function for training the network. For classification, the 925 posts were divided into three sets: training (557 posts), validation (184 posts) and testing (184 posts). The selection of the posts for each set, as well as the initialization of the networks weights is random. Thus, the performance of the network depends on the

random seed and varies from one execution to another. Therefore, the reported accuracy of the network classification corresponds to the average value of 100 independent runs.

5.2 Computational Results

We use the classification accuracy metric as well as the mean standard error to compare each of the 18 implemented algorithms. The algorithms are compared based on: the different degrees (classes) of importance (2, 3 or 6 classes), for each of the target variables (affective, cognitive and instrumental information value), as well as when using only the objective vs. all input factors. Classification accuracy measures the frequency with which a recommender system correctly classifies an item (Herlocker et al., 2004) and thus can be considered a reliable measure for comparison.

Figure 1. Classification accuracy of posts in two classes (binary classification)

When classifying the posts into two classes (Figure 1), the average relative prediction improvement is 21% compared to a baseline of 50% (completely random classification). The maximum achieved accuracy of the algorithm is 73.1% when predicting affective value using all data, whereas both instrumental and cognitive value are a bit harder to predict. When using only the objective data, the algorithm achieves an average accuracy of 62%. By including subjective data into the algorithm we can achieve an average increase in accuracy of 9.2%. However, this increase varies depending on which dimension is predicted: objective data allows a good classification when targeting affective and instrumental value, whereas for cognitive value the objective data alone is not enough.

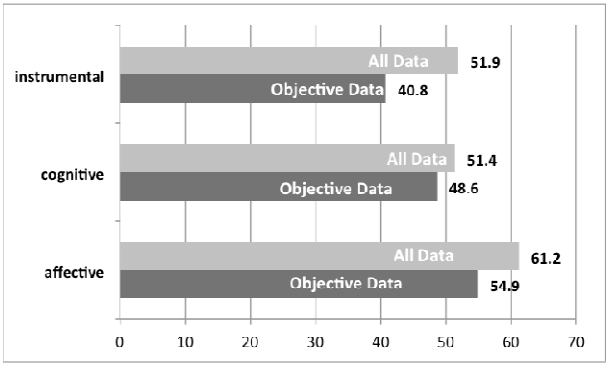

Figure 2. Classification into three classes

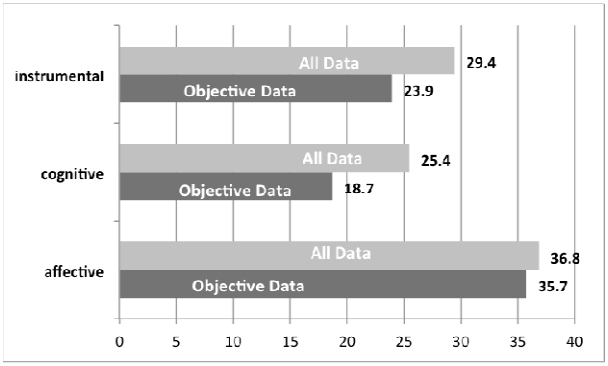

Figure 3. Classification into six classes

When classifying into three classes (figure 2), the average relative improvement compared to a random baseline (33.3%), is 21.5%, which is slightly higher than with the two classes classification. The maximum accuracy achieved is 61.2% with affective value of information. The other two dimensions perform much worse in this classification. We note an increase in accuracy, on average by 6.2%, when all data is taken into account as opposed to using only objective data. Interestingly, the relative increase in accuracy varies: objective data is now enough to predict the cognitive value, whereas it is not sufficient for the instrumental, and performs moderately with the affective dimension.

When classifying into six classes (Figure 3) the average improvement is at 14.4% compared against a random baseline (16%), which is lowest compared to other classifications. The highest accuracy of 37% is again achieved with affective value when all data is taken into account, whereas the cognitive dimension is hardest to predict. When subjective data is considered, an average improvement of only 4.4% can be achieved, as opposed to using objective data only. The most interesting result is with affective value: the difference between using objective or subjective data is almost nonexistent, whereas instrumental and cognitive dimensions show a stepwise increase in accuracy levels.

$$\overline{|E|} = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N} \quad (1)$$

To further validate our results, we use the Mean Absolute Error metric (Herlocker et al. 2004). MAE measures the average absolute deviation between the predicted classification and the user's classification using equation (1), where p_i is the predicted class and r_i is the users classification, N is the total amount of post predicted by the system. The results presented in table 3 show that the smallest difference between using objective data and all data is achieved for the affective dimension, while the cognitive dimension is the one that benefits the most from using all data. Therefore we can conclude that using objective vs. all data depends on the type of evaluation: to predict affective value of information, using all data is not significantly better than using just the objective data, whereas for instrumental and cognitive evaluations, all data is needed for more accurate predictions.

Number of Classes	**Affective**		**Cognitive**		**Instrumental**	
	Objective	**All Data**	**Objective**	**All Data**	**Objective**	**All Data**
2 Classes	0.31	0.27	0.46	0.29	0.36	0.29
3 Classes	0.55	0.44	0.77	0.59	0.81	0.68
6 Classes	0.89	0.79	1.18	0.98	1.23	1.08

Table 3. Mean Absolute Error classification accuracy

To determine the difference in processing cost when using all data vs. objective data, the time required for the NN to do the classification was measured. The experiment was done with an Intel Core2 Duo CPU at 2.4Ghz and 2Gb RAM. As a result, we observe a clear decrease in processing cost: when using only objective data execution time is on average 316 seconds, whereas when using all data ca. 354 seconds are required. The attained decrease in the necessary processing time for objective data vs. all data for the different dimensions of information value is: 11.1% for affective, 11.3% for instrumental and 10.1% for affective.

5.3 Predicting the Tie Strength

Tie strength is a direct measure of the relationship between the "poster" and the "receiver" of information on SNS. However, in order to quantify the tie strength using the information available on the network, the algorithm would need to process large amounts of data, similar to the model of Gilbert and Karahalios (2009). Therefore we explore whether communication intensity (TechCrunch, 2010) or the similarity between users (Gilbert and Karahalios, 2009) is a better predictor of tie strength on SNS. By determining which factor better predicts tie strength, we can reduce the necessary processing effort for the algorithm. Already by the distributions depicted in figures 4 and 5 we can see that similarity follows a more normal distribution similar to the one of tie strength: most users have something in common as well as almost as much have very much and nothing in common at all. The skewed distribution of the frequency of communication, however, hints that on average users do not communicate with a majority of their friends, putting the underlying tie strength in question.

We estimate two algorithms use a neural network (NN) classifier on the whole sample of data: one with three distinct forms of communication intensity and the other with similarity as input factors and

tie strength as the target variable in both cases. If we use the intensity of communication and similarity of interest, the prediction accuracy of the tie strength is 73% and 76% in the binary classification (weak or strong tie) and 44% and 50% in the multiple class one (corresponding to the 5pt ordinal scale), respectively. We conclude that similarity between users is a slightly better predictor of tie strength, than intensity of communication. This can be explained by the fact that users presumably use other means to communicate with their close friends (Vitak et al., 2011).

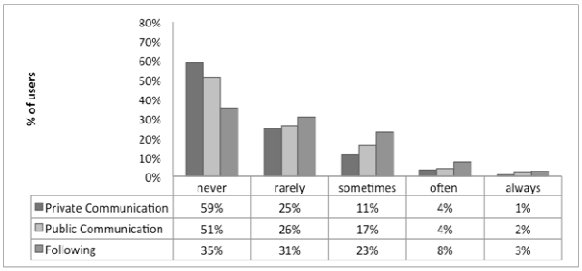

Figure 4. Communication Intensity

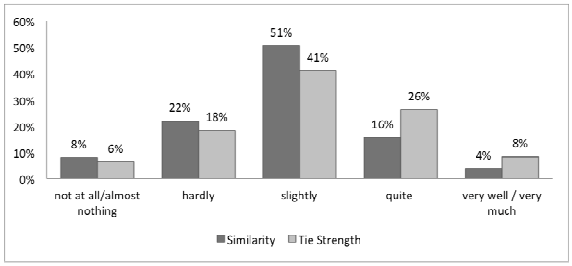

Figure 5. Similarity

6 RANKING ALGORITHM

6.1 Methodology and Design

The neural network algorithm can also be used to rank the posts according to their estimated importance and compare these values with the real ranks provided by users. For the algorithm, we use the average of all dimensions of information value (affective, cognitive and instrumental) as the target variable and all factors that have a significant impact on any dimension of value as input (all factors in table 1, except for location). The classification returns a real number which provides a measure of the post's importance which we use to assign the system's ranks to the posts for each user. The network was trained using 650 randomly selected posts, whereas the other 275 were used for validation. Once the network was trained, the posts for each user were ranked.

On the other hand, in the survey each user evaluated up to six posts which were randomly selected from the Newsfeed, as described in section 3. By summing up the evaluations along all the dimensions of information value, we were able to assign a user rank to each piece of shared information and compare these results with the ranks provided by the system. Note that we only take users who evaluated 6 posts, which results in a total of 141 users whose posts could be ranked. The user ranking and the one obtained from the system are then compared by applying the "precision of preferences" method (Carterette et al., 2008). The accuracy measure provided by this method is the proportion of the pairs correctly ranked by the algorithm. More formally, over all pairs of posts i, j such that i is preferred to j by the user, the "precision of preferences" method returns the proportion for which the system ranked i above j. If two posts have the same user rank the pair is ignored, thus the method is not affected by the presence of posts equally rated by the user.

6.2 Computational Results

The results of comparing the "real" rankings provided by users with the ones obtained through the algorithm show the ranking accuracy for the target category that comprises all dimensions of information value. The average accuracy of the implemented ranking algorithm over the posts of 141 users comprises 74.5%, where the minimal accuracy achieved is 28.6% and the maximum – 100%. The median lies at 78%. The distribution of the ranking accuracy over the posts of the users can be traced in figure 6. We notice a skewed distribution which already signals good performance of the ranking algorithm: for over 70% of the users accuracies of over 70% and higher can be achieved. For almost 30% of the users the prediction errors amount to 20% or less. We consider this a promising result for such a difficult task as ranking of information on SNS.

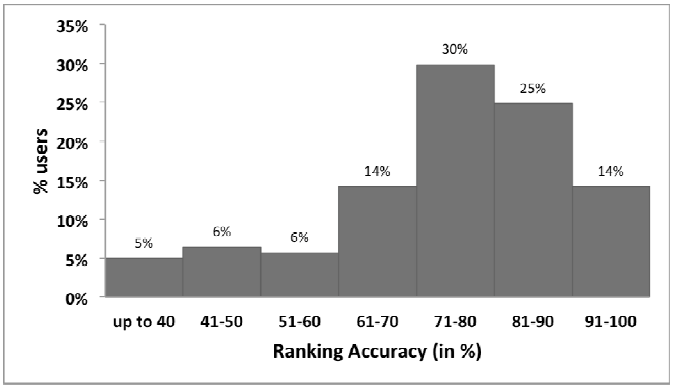

Figure 6. Distribution of the ranking accuracy

7 DISCUSSION

First and foremost, our results confirm that Neural Networks are an effective technique for classifying posts on SNS, which allow to achieve prediction accuracies of up to 73%. One of the main contributions of our work, however, is that it is possible to achieve such prediction accuracy by using much less data than has been done previously. For example, to predict affective value, only the objective characteristics are sufficient, such as number of comments and affirmations, post type and word count. This result underscores the necessity to carefully choose which factors to include into the filtering algorithms. Moreover, we show that using less data increases the efficiency and decreases processing costs of the algorithm: by using only objective data the algorithm performs on average 11% faster, which is especially significant if the classification process has to be done in real time on SNS.

Our study identifies factors that positively impact information value on SNS and can thus aid in designing algorithms. The number of affirmations, the similarity between users and the understandability of the post increase the value of information. These factors can be quantified using the available network data. For example, understandability of the post can be implied by the match in languages between the users reported on the profile or quantified using Neuro-linguistic programming. At the same time, similarity can be determined using the developed content matching techniques (Pazzani et al., 1996; Gilbert and Karahalios, 2009). Moreover, our study shows that tie strength can better be predicted by the similarity of interests between users rather than communication intensity, which is mainly utilized by Facebook. Therefore algorithm designers should be careful when using intensity of communication on SNS to recommend content to users. At the same time, we identify factors that have a detrimental impact on information value. For example, long posts with a lot of comments coming from people who post very often will most likely be negatively perceived by users. Thus, information filtering mechanisms should filter out the "spammers" on the network and treat the posts with a high number of words and comments carefully.

Moreover, our study explores different dimensions of information value on SNS. The best accuracy is clearly obtained for the affective value of information, whereas the cognitive and instrumental dimensions present a more or less similar level of difficulty to the classifier. This is in line with the results of the regression presented in section 4: the independent variables are best at explaining the affective value. This can be best explained by the fact that SNS is mainly used for entertainment and socializing (Krasnova et al., 2010), and therefore affective value of information is probably most salient on these networks. Moreover, affective value less depends on the content of the post, whereas cognitive and instrumental dimensions require extensive analysis of post content.

Furthermore, in our paper we not only filter the irrelevant posts, but also classify posts into multiple classes. This finer-grained classification allows a more precise filtering of the information. Although classifying into two classes allows us to reach higher accuracy, it does not allow us to determine the order of information presentation. If, for example, in case of the two class algorithm the results provide more candidate posts than can be presented to the user at one time, the algorithm is unable to

select which ones to present, resulting in an additional loss in the overall accuracy. On the contrary, the multiple classes algorithm would select only the highest ranked posts for presentation.

Finally, we implement an individual ranking algorithm for the users and achieve even higher accuracies than with the filtering algorithm. The ranking is effective, as the amount of posts presented to the user can be adjusted according to different parameters, such as the log-in frequency or be defined by the users themselves, who desire to regain control over information presentation (Tonkelowitz, 2011). As no other study until now has attempted a ranking of information on SNS, our results could provide a valuable benchmark for future research.

8 CONCLUSION

In our paper we design and implement several algorithms that allow to reduce information overload on SNS. First of all, we show that satisfactory levels of prediction accuracy can be achieved with a subset of available network data by using a NN algorithm. However, improving information filtering on SNS is conditional on the careful inclusion of significant factors. Second, our paper is a first attempt to rank the posts in order of their relevance and thus achieve the highest goal of effective information classification.

The main limitation of the study is that some of the variables operationalized subjectively could be measured objectively using the available network data. For example, to infer intensity of private communication, such factors as the number of personal messages sent or the number of chats initiated could be retrieved from Facebook. Collecting data from SNS objectively, however, is bounded by two problems. First, Facebook allows to collect this data only for the period of 30 days, which is not enough to generate the necessary data variability. Second, this data is raw and quite complicated interpretation techniques are necessary to obtain the variables of interest.

9 REFERENCES

Ajzen, I. (2005). Attitudes, personality, and behavior, 2nd Edition, McGraw-Hill, New York.

Boyd, D.M. and Ellisson, N.B. (2008). Social Network Sites: Definition, History and Scholarship. Journal of Computer-Mediated Communication, 13, 210-230.

Butler, J.S. and Moffitt, R. (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. Econometrica, 50, 761-764.

Carterette, B., Bennett, P. N., Chickering, D. M. and Dumais, S. T. (2008). Here or There Preference Judgments for Relevance. In Proceedings of ECIR. Lecture Notes in Computer Science, 4956.

Chen, Y., Shang., R-A. and Kao, C-Y. (2009a). The effects of Information Overload on consumers' subjective state towards buying decision in the Internet shopping environment. Electronic Commerce Research and Applications, 8, 48-58.

Chen, J., Geyer, W., Dugan, C., Muller, M., and Guy, I. (2009b). Make new friends, but keep the old: recommending people on social networking sites. In Proceedings of CHI, ACM Press.

Chen, J., Nairn, R., Nelson, L., Bernstein, M. and Chi, E.H. (2010). Short and Tweet: Experiments on Recommending Content from Information Streams. In Proceedings of CHI 2010, ACM Press.

Chen, J., Nairn, R. and Chi, E.H. (2011). Speak Little and Well: Recommending Conversations in Online Social Streams. In Proceedings of CHI 2011, ACM Press.

Demuth, H., and Beale, M. (1997). Matlab Neural Network Tool Box User's guide, The Mathworks.

Ellison, N.B., Steinfield C. and Lampe C. (2007). The benefits of Facebook "Friends": Social Capital and College Students' Use of Online Social Network Sites. Journal of Computer-Mediated Communication, 12 (4).

Eppler, M.J. and Mengis, J. (2004). The Concept of Information Overload: A Review of Literature from Organization Science, Marketing, Accounting, MIS and related Disciplines. The Information Society: An International Journal, 20 (5), 1-20.

Facebook Query Language (2011a). http://developers.facebook.com/docs/reference/fql/

Facebook Statistics. (2011b). http://www.facebook.com/ press/info.php?statistics

Gilbert, E. and Karahalios, K. (2009). Predicting Tie Strength with Social Media. In Proceedings of CHI 2009, ACM Press.
Glynn, C.J., Huge, M.E. and Hoffman, L.H. (2012). All the news that's fit to post: A profile of news use on Social Networking Sites. Computers in Human Behavior, 28, 113-119.
Greene, W. H. (2000). Econometric Analysis, Prentice Hall, New Jersey.
Herlocker, J.L., Konstan, J.A., Terveen, L.G. and Riedl, J.T. (2004). Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems, 22 (1), 5-53.
Hill, W. and Terveen, L. (1996). Using frequency-of-mention in public conversations for social filtering. In Proceedings of CSCW.
Jones, Q., Moldovan, M., Raban, D., and Butler, B. (2008). Empirical Evidence of Information Overload Constraining Chat Channel Community Interactions. Proceedings of the CSCW.
Köbler, F., Riedl, C., Vetter, C., Leimeister, J.M. and Krcmar, H. (2010). Social connectedness on Facebook - an explorative study of status message usage. In Proceedings of AMCIS, paper 247.
Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R. and Riedl, J. (1997). GroupLens: Applying collaborative filtering to Usenet news. Communications of the ACM, 40 (3), 77-87.
Koroleva, K., Krasnova, H. and Günther, O. (2010). 'STOP SPAMMING ME!' – Exploring Information Overload on Facebook. In Proceedings of the AMCIS, Lima, Peru, paper 447.
Koroleva, K., Stimac, V., Krasnova, H. and Kunze, D. (2011a). I like it because I('m) like you – Measuring User Attitudes Towards Information on Facebook. In Proceedings of the International Conference on Information Systems, Shanghai, China, paper 26.
Koroleva, K., Krasnova, H. and Günther, O. (2011b). Cognition or Affect? – Exploring Information Processing on Facebook. In Proceedings of the Third International Conference on Social Informatics (SocInfo'11), Springer-Verlag Berlin, Heidelberg.
Krasnova, H., Spiekermann, S., Koroleva, K. and Hildebrand, T. (2010). Online Social Networks: Why we disclose? Journal of Information Technology, 25, 109-125.
Lampe, C., Vitak, J., Gray, R. and Ellison, N.B. (2012). Perceptions of Facebook's Value as an Information Source. To appear in Proceedings of CHI 2012, Austin, Texas, USA.
MacFadden, D.L. Conditional Logit Analysis of Qualitative Choice Analysis Frontiers in Econometrics, New York: Academic Press, 1974, 105-142.
Morris, M.R., Teevan, J., and Panovich, K. (2010). What do people ask their social network and why? A survey study of status message Q&A behavior. In Proceedings of the CHI 2010, Atlanta, USA.
Paek, T., Gamon, M., Counts, S., Chickering, D.M. and Dhesi, A. (2010). Predicting the Importance of Newsfeed Posts and Social Network Friends. In Proc. of the 24th AAAI Conference on Artificial Intelligence, Association for the Advanced Artificial Intelligence.
Pazzani, M. J., Muramatsu, J., and Billsus, D. (1996). Syskill webert: Identifying interesting web sites. In Proceedings of the 13th National Conference on Artificial Intelligence, 1, 54-61.
Schneider, S.C. (1987). Information Overload: causes and consequences. Human Systems Management, 7, 143-153.
Schick, A.G., Gorden L.A. and Haka, S. (1990). Information Overload: A temporal approach. Accounting Organizations and Society, 15, 199 – 220.
Schroder, H.M., Driver, M.J. and Steufert, S. 1967. Human information processing - Individuals and Groups functioning in complex social situations. New York: Holt, Rinehart & Winston.
TechCrunch. (2010). EdgeRank: The Secret Sauce that makes Facebook's Newsfeed Tick. http://techcrunch.com/2010/04/22/facebook-edgerank/
Tonkelowitz, M. (2011). Interesting News, Any Time You Visit. The Facebook blog, http://blog.facebook.com/blog.php?post=10150286921207131
Vitak, J., Ellison, N.B. and Steinfield, C. (2011). The Ties that bond: re-examining the relationship between Facebook Use and Bonding Social Capital. In Proceedings of the 44th Annual Hawaii International Conference on System Sciences, Computer Society Press.
Voss, K.E., Spangerberg, E., and Grohmann, B. (2003). Measuring the hedonic and utilitarian dimensions of consumer attitude. Journal of Marketing Research, 40 (3).
Yang, H., and Yoo, Y. (2004). It's all about attitude: revisiting the technology acceptance model. Journal of Decision Support Systems 38 (1).