Characterizing Social Relations Via NLP-based Sentiment Analysis

Georg Groh, Jan Hauffa
Fakultät für Informatik
Technische Universität München, Germany
grohg,hauffa@in.tum.de

Abstract

We investigate and evaluate methods for the characterization of social relations from textual communication context, using e-mail as an example. Social relations are intrinsically characterized by the Cartesian product of weights on various axes (we employ valuation and intensity as examples). The prediction of these characteristics is performed by application of unsupervised learning algorithms on meta-data, communication statistics, and the results of deep linguistic analysis of the message body. Classification of sentiment polarity is chosen as the means of linguistic analysis. We find that prediction accuracy can be improved by introducing limited amounts of additional information.

Introduction

Social relations become substantially manifest in interpersonal communication, which is accessible on the Social Web and thus allows for the evaluation and application of methods for characterizing the relations on the basis of communication data. Meta-data (e.g. sender, recipients, time stamps), content of individual artifacts, and the statistical properties of an aggregation may be analyzed. Instead of an ontology- or Folksonomy-based classification of relations, which own preliminary studies showed to be less adequate for socio-psychological reasons, relations are characterized as a Cartesian product of weights along characterizing axes. Characterizing relations with weights may improve e.g. the quality of inference on social networks, as demonstrated by (Barrat et al. 2004). Goal of this work is the accurate prediction of these relationship characteristics using all information provided by the communication artifacts, with special focus on linguistic analysis of the message content. We selected two axes which cover important characteristics of a relation: *Emotional intensity* is considered to be an indicator of tie strength by (Granovetter 1973). The *valence* of a relationship is a generalization of the binary emotional polarity in a signed social network. Sentiment analysis, specifically the classification of sentiment polarity, is chosen as the method of linguistic analysis, because of the connection between sentiment polarity and relationship valence.

Related Work

Extracting a social network graph from a collection of email messages is covered in several studies, which mostly focus on "outer statistics" (such as message frequency, reciprocity, thread structure etc.) and mostly neglect the deep linguistic analysis of message content: (Bird et al. 2006) use a public mailing list archive as a data source and identify common problems in obtaining a clean data set. (Matsuo, Mori, and Hamasaki 2006) present a system for finding and quantifying arbitrary relationships between persons on the Web, using specially crafted search queries to estimate tie strength. (Kunegis, Lommatzsch, and Bauckhage 2009) analyze the social network induced by threaded discussions on the website Slashdot. Their unsupervised analysis generates a signed social network from users' lists of "friends" and "enemies". (Gilbert and Karahalios 2009) develop a predictive statistical model for measuring the strength of relationships between users of online social network services. Communication statistics, similarity of user profiles, and frequency of words associated with emotion and intimacy in the exchanged messages are used as variables. The predictions are compared to the results of a user survey. (Wilson, Wiebe, and Hoffmann 2009) perform sentiment analysis with rich linguistic features on the level of individual words. A comprehensive review of sentiment analysis techniques can be found in (Prabowo and Thelwall 2009).

Data Acquisition

In order to evaluate a method for extracting weighted social networks from email, a corpus of email messages and a matching weighted social network graph is acquired by means of a survey. The concepts of emotional intensity and valence are explained to the otherwise untrained participants. Since the full message body and headers are analyzed, special care must be taken to preserve the privacy of the participants. A special purpose GUI application performs the anonymization entirely on the participant's PC. Any occurrence of a name or email address is replaced with a random, but unique identifier (UUID) generated for the person. From the meta-data of the emails, the application extracts an egocentric social network with the survey participant as the ego, and asks him or her to rate each relationship in terms of emotional intensity and valence. This induces a weighted and

directed social network. Weights are represented as Cartesian products of real numbers in the interval $[0, 1]$. Since information about the alters' assessment of the relationship is not available, all edges are directed towards the alters. Each message is associated with one or more edges.

Data Processing

Given a social network graph and the associated messages provided by the survey application, the task is to predict the weights corresponding to emotional intensity and valence. The human assigned weights are held back for evaluation.

Preprocessing

Not all parts of the message body are of equal value to linguistic analysis. When replying to an email message, it is customary to quote the specific parts of the original message the reply refers to. It is not unusual that parts of the message at the beginning of a conversation are quoted in each following message. To avoid giving an inappropriate weight to these parts of text, quotations are detected heuristically and removed. Some email systems automatically append a signature to the message body, which is discarded as well.

Feature Generation

Table 1 lists the features for prediction of the different relationship attributes. There are two kinds of features: Message features are generated by analyzing each message associated with the relationship and computing the arithmetic mean of all per-message feature values. Relationship features represent statistic properties of the whole set of messages and are computed once per relationship. The choice of features depends on the attribute of the relationship that is to be predicted. An entry of "inv." indicates that the feature value is inverted: $v' = 1 - v$. There are two relationship features: The *relative message frequency* is the number of messages sent and received relative to the overall number of messages. This feature is an approximation of relationship activity, but does not take into account the temporal spacing of the messages. The *message frequency balance* feature is derived from the ratio of incoming and outgoing messages. Its value is maximal in a perfectly balanced relationship, where the actors send and receive the same amount of messages. The *message length* feature is the length of a message relative to the longest message. For the *use of first name* feature, the different forms of addressing a person by name are identified in the message body: Any sequence of words starting with a title or name component (UUID) followed by one or more name components is classified as a full name. Any name component that does not match the pattern is a first name. The feature value is the frequency of first names relative to the overall number of name sequences. The *elongated words* feature originates from an analysis of MySpace comments by (Thelwall 2009), where emotion carrying words were often found to be emphasized by repetition of one or more letters, e.g. "really" becoming "reeeallly". The feature value is the relative frequency of elongated words. A similar stylistic device is *word obfuscation*, where one or more letters are replaced with punctuation characters. The feature *words with pos. / neg. polarity* is the relative frequency of sentiment carrying words. Sentiment directed towards the recipient has to be distinguished from sentiment directed towards the subject of the message when computing the feature value: Under the assumption that each message has a central subject matter, sentiment towards the recipient is concentrated in the beginning and end. The influence of a word's polarity on the feature value is weighted by a gaussian function of the position of the word within the message body. A different approach is to look for "cue words", which identify a sentence as a statement about the relationship. The present implementation uses a small list of personal and possessive pronouns. The feature *colloquial expressions with pos. / neg. polarity* counts single or multi word expressions that occur on a list. The highest ranking definitions for 3640 popular expressions from UrbanDictionary.com were processed with a sentiment polarity classifier. For each expression, the ratio of positive words to polar (positive and negative) words was computed. The feature value is the average polarity ratio of all words in the message body. In addition, a list of emoticons was compiled from Internet resources and manually classified by polarity. Another word list feature is the relative frequency of *text message abbreviations*. It is intended as an indicator of the amount of colloquial language.

Feature	Type	Intensity	Polarity
relative message frequency	rel.	yes	
message frequency balance	rel.	yes	
message length	msg.	yes	
use of first name	msg.	yes	
elongated words	msg.	yes	
obfuscated words	msg.	inv.	
text message abbreviations	msg.	yes	
words with pos. polarity	msg.	yes	yes
words with neg. polarity	msg.	yes	inv.
colloquial expr., pos. polarity	msg.	yes	yes
colloquial expr., neg. polarity	msg.	yes	inv.

Table 1: Composition of feature vectors for the prediction of relationship attributes

Dimensionality Reduction

The problem of obtaining a rating for a relationship from a feature vector can be formulated as reducing its dimensionality from n to 1, assuming that each component of a feature vector is positively correlated with the rating. A simple method is to compute the arithmetic mean of the components, so that each feature is given the same weight, and each feature vector is treated individually. For more sophisticated methods, multiple feature vectors are concatenated to form a matrix. Principal Component Analysis (PCA) is a method for transforming a data set into a lower dimensional space while minimizing the loss of variance: The data is transformed linearly, so that the direction of highest variance (principal component) coincides with the first axis, etc. The stronger the linear correlation between the features, the lower is the variance of the higher dimensions after transformation. All dimensions but the first are discarded.

A third method of dimensionality reduction uses Self-Organizing Maps (SOM) developed by (Kohonen 1982). A SOM consists of a fixed number of vectors in the input data space ("neurons"), which are arranged as a "map" in a lower dimensional space. Neurons are initialized randomly and move towards areas with a high density of data points in an iterative process. Whenever a neuron moves, its neighbors on the map move as well. If the map is one-dimensional, the neurons represent the principal curve, a non-linear generalization of the principal component. For each feature vector, the closest neuron in feature space (BMU) is determined, and a scalar representation is derived from the neuron's location on the map. A constraint is added to the SOM fitting process, to ensure that the component average of the neurons rises in correlation with their horizontal position on the map. Given a data point x, a neuron v located at position (X, Y) on the map is only moved if it is located to the left of the BMU ($X < X_{BMU}$) and the movement would raise its component average ($\sum_{i=0}^{n} x_i - v_i > 0$), or conversely $X \geq X_{BMU}$ and $\sum_{i=0}^{n} x_i - v_i < 0$.

Transformation

Dimensionality reduction generates a scalar value for each relationship, which is assumed to be highly correlated with the unobserved true rating. To make true and predicted ratings comparable, one has to make assumptions about their respective distributions. Then the predictions can be transformed so that their distribution approximates or matches the parameters of the hypothetical distribution. We propose four methods of transformation, each based on a linear function $f_{a,b}(x) = a \cdot (x + b)$.

The first method naively assumes that the true ratings come from a uniform distribution over the interval $[0, 1]$. If a mailbox contains messages from a sufficient number of relationships, there will be some with true ratings on the extremal points of the scale. If the predictor is accurate, the predicted ratings for these messages will also be extreme relative to the other predictions. Given the minimum and maximum predicted ratings p_{min} and p_{max}, we choose offset $b = -p_{min}$ and scale factor $a = 1/(p_{max} - p_{min})$.

The second method assumes that the true ratings come from a normal distribution with a mean of 0.5, which corresponds to moderate emotional intensity / neutral polarity, and a variance of $1/36$, so that $\overline{x} + 3 \cdot SD = 1$ (SD is the standard deviation). The predicted values are assumed to come from a normal distribution with yet unknown parameters. From the sample estimate of variance, the standard deviation SD' is computed. We choose $b = 3 \cdot SD' - \overline{x}$ and $a = 1/6 \cdot SD'$.

For the third scaling method, the assumption of normality of the true ratings is relaxed by not requiring specific parameters. The true ratings are transformed in the same way as the predicted ratings. This means discarding information about the original distribution of the true ratings, and thus overestimating the performance of the predictor. It is an indication of how well a predictor would perform if more information about the distribution of true ratings was provided.

The fourth method simulates knowledge about the correspondence between the true and the predicted rating. For the n highest and lowest rated relationships, $\overline{r}_t$ and $\overline{r}_p$ are the arithmetic mean of the true and predicted ratings. The predictions for these $2n$ relationships are discarded to avoid biasing the subsequent evaluations. First, ratings are transformed to $[0, 1]$ by choosing $b = -\overline{r}_{p,min}$ and $a = 1/(\overline{r}_{p,max} - \overline{r}_{p,min})$. A second transformation with $b = \overline{r}_{t,min}/(\overline{r}_{t,max} - \overline{r}_{t,min})$ and $a = \overline{r}_{t,max} - \overline{r}_{t,min}$ moves them into $[\overline{r}_{t,min}, \overline{r}_{t,max}]$.

Sentiment Analysis

Sentiment polarity classification on the level of individual words can be considered a sequence labeling problem, where each word in a sentence is to be assigned a label from a set of polarity classes. Conditional Random Fields (CRF) as devised by (Lafferty, McCallum, and Pereira 2001) were chosen as the learning method because of their efficiency in handling large feature vectors according to the maximum entropy principle. The CRF implementation of the MALLET toolkit 2.0-RC4 (McCallum 2002) is used.

A linear-chain CRF is trained on a labeled dataset. Training of a sentiment polarity classifier on the expression level requires a corpus where individual words and phrases have been manually labeled with their sentiment polarity. Version 2.0 of the MPQA opinion corpus created by (Wiebe, Wilson, and Cardie 2005) provides these and other high-level semantic annotations with a focus on subjective language. We convert the corpus to a simpler representation, retaining only the polarity information. Sentences consisting of only neutral words are discarded, but the resulting distribution of labels is still highly imbalanced, with 91.6% of words being classified as neutral.

General linguistic features with rising levels of abstraction (statistics, morphology, syntax, semantics) were evaluated. A combination of statistical features (n-grams, word / sentence length, word position and number of occurrences in sentence) and morphological features (part-of-speech tags, pre- and suffixes, word stems, capitalization) was found to perform best, surpassing even models with more complex linguistic features.

Experimental Evaluation

Email messages were collected from five persons. Two additional datasets from previous work ((Richter and Groh 2007)) are only annotated for emotional intensity. In total, 399 messages exchanged between 122 actors were collected, an average number of 3.5 messages per relationship. Figure 1 illustrates the distribution of ratings. Each circle corresponds to one or more relationships with a specific combination of intensity and valence ratings, indicated by the location of the center. The radius is proportional to the number of relationships. Shading indicates missing valence ratings. There are two clusters, the first approximately located at an intensity and valence of (0.1, 0.6), the second at (0.8, 0.9). This implies a bimodal distribution of intensity and valence. The first cluster corresponds to relationships with little emotional intensity and slightly positive valence, e.g. brief ac-

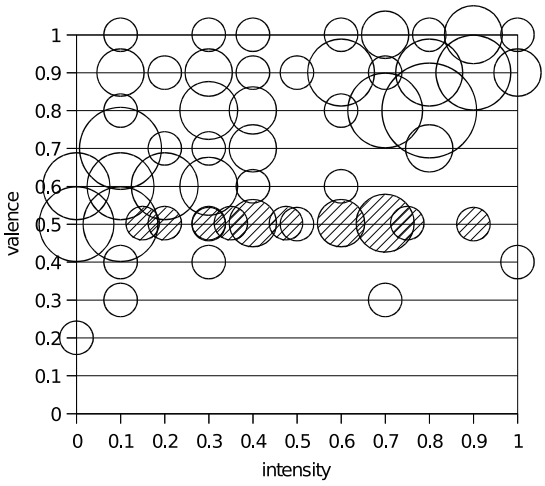

Figure 1: Distribution of intensity and valence ratings

quaintances. The second cluster corresponds to emotionally intense relationships of positive valence, like close friendships. Relationships with negative valence are underrepresented.

The root mean square error (for two vectors p and q of n ratings: $RMSE(p,q) = (\frac{1}{n}\sum_{i=1}^{n}(p_i - q_i)^2)^{\frac{1}{2}})$ is chosen as a measure of accuracy. Tests are performed for all three methods of dimensionality reduction. Three SOM configurations are considered: 11×1, 11×11, and 40×3 neurons. Including variations of other parameters, 38 configurations are evaluated. The predictors are compared to two baseline methods: The first is to choose a fixed value of 0.5 as the prediction. The second is to submit five random values drawn from a uniform distribution over $[0, 1]$ to the transformation. Table 2 shows the improvement over baseline of the best performing predictor configuration for each scaling method (average RMSE over all datasets and attributes).

method	avg. RMSE	Δ constant	Δ random
1	0.403	-38.2%	-23.2%
2	0.311	-6.5%	5.1%
3	0.196	32.9%	40.2%
4	0.250	14.4%	23.7%

Table 2: Improvement over baseline of the best predictor for each scaling method

With the first scaling method, performance is below both baselines. The assumption of a uniform distribution of true ratings does not hold, especially in the case of valence, where no examples for the lowest ratings are present. The second scaling method assumes a normal distribution and shows bad performance for similar reasons. The next two scaling methods were designed to gauge the effect of introducing small amounts of knowledge about the distribution of the true ratings: The third method achieves the best results. It is equivalent to knowledge about the distribution parameters mean and variance of the ratings. The fourth method simulates knowledge of the correspondence between the range of true ratings and the range of predicted ratings. It yields a smaller improvement over both baselines.

Conclusion

An improvement of prediction accuracy over the baseline could only be achieved by providing additional information about the distribution of the true ratings, which then acts as a frame of reference for the prediction. One way to integrate such information is supervised training of the predictor, either directly by annotating some relationships, or indirectly via feedback about the prediction accuracy. This is often undesirable, especially in the case of large scale data mining. The bimodal distribution of valence and intensity shown in figure 1 suggests that a simple binary classification of a relationship as professional / private, or acquaintance / friend could provide enough information to make an accurate prediction. Yet, the observed distribution might be an artifact of the small number of datasets evaluated. A larger study of online communication is necessary to gain further insight.

References

Barrat, A.; Barthélemy, M.; Pastor-Satorras, R.; and Vespignani, A. 2004. The architecture of complex weighted networks. *PNAS* 101(11):3747–3752.

Bird, C.; Gourley, A.; Devanbu, P.; Gertz, M.; and Swaminathan, A. 2006. Mining email social networks. In *Proc. Int'l Workshop on Mining Software Repositories*, 137–143.

Gilbert, E., and Karahalios, K. 2009. Predicting tie strength with social media. In *Proc. 27th Int'l Conf. on Human Factors in Computing Systems*.

Granovetter, M. 1973. The strength of weak ties. *American Journal of Sociology* 78(6):1360–1380.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43:59–69.

Kunegis, J.; Lommatzsch, A.; and Bauckhage, C. 2009. The Slashdot zoo: Mining a social network with negative edges. In *Proc. 18th Int'l Conf. on World Wide Web*, 741–750.

Lafferty, J.; McCallum, A.; and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th Int'l Conf. on Machine Learning*, 282–289.

Matsuo, Y.; Mori, J.; and Hamasaki, M. 2006. POLYPHONET: An advanced social network extraction system from the web. In *Proc. 15th Int'l Conf. on World Wide Web*, 397–406.

McCallum, A. 2002. MALLET: A machine learning for language toolkit. Accessed on July 16, 2009.

Prabowo, R., and Thelwall, M. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics* 3(2):143–157.

Richter, N., and Groh, G. 2007. Analysis of social relationships via email. Software development project, TUM.

Thelwall, M. 2009. MySpace comments. *Online Information Review* 33(1):58–76.

Wiebe, J.; Wilson, T.; and Cardie, C. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39(2–3):164–210.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics* 35(3):347–354.