Edda Leopold, Jörg Kindermann, Gerhard Paaß

Analysis of E-Discussions Using Classifier Induced Semantic Spaces

We categorise contributions to an e-discussion platform using Classifier Induced Semantic Spaces and Self-Organising Maps. Analysing the contributions delivers insight into the nature of the communication process, makes it more comprehensible and renders the resulting decisions more transparent. Additionally, it can serve as a basis to monitor how the structure of the communication evolves over time. We evaluate our approach on a public e-discussion about an urban planning project, the Berlin Alexanderplatz, Germany. The proposed technique does not only produce high-level-features relevant to structure and monitor computer mediated communication, but also provides insight into how typical a particular document is for a specific category.

1 Introduction

E-discussion platforms facilitate the moderated collaboration between persons distributed in time and space. In order to support a focussed and goal-oriented discussion it is desirable to provide condensed information about the ongoing discussion process in order to monitor and to influence the discourse.

We propose to analyse e-discussion contributions on the basis of Classifier Induced Semantic Spaces (CISSs) (Leopold et al., 2004; Leopold, 2005), and visualise the resulting semantic spaces with Self-Organising Maps (SOMs)(Kohonen, 1995). CISSs can be constructed from any supervised classifier, that determines the membership of a given entity to a pre-defined category on the basis of some numerical threshold. Here, we use a Support Vector Machine (SVM) (Vapnik, 1998) as classifier.

A semantic space is a metric space whose elements are representations of signs of a semiotic system. The metric of the space quantifies some semantic dissimilarity of the signs. If the semantic space is a vector space then its dimensions are associated with some kind of meaning.

Self-Organising Maps are a technique to map elements of a vector-space with three or more dimensions into a two-dimensional "map" by preserving the original distance relationships as far as possible. They therefore allow to represent the structure of a semantic space in the two dimensions of a computer screen.

The remainder of the paper is organised as follows. In Section 2, we discuss the integral parts of the proposed method of structuring the components of a communication network built from e-discussion contributions. In Section 4, we describe the experimental evaluation of our approach, and in Section 5, we conclude.

2 Discourse Grammar and E-Discourse

According to Turoff et al. (1999) a discourse is a deliberative, reasoned communication, which is focused and intended to culminate in decision making. Turoff et al. (1999) argued that building a discourse grammar, which allows individuals to classify their contributions according to their pragmatic function within the discourse is a collaborative effort and is an integral part of the discussion process.

Inspired by the Bülerian Organon-Model (Bühler, 1934), we decided to consider a discourse grammar, that consists the following pragmatic functions: 'giving information', 'making an objection', 'asking a question', and 'giving a reply' . So the contributions of the e-discourse can be assigned to four different classes $z_k, k = 1 \ldots 4$, corresponding to the linguistic functions they fulfil. Multiple class assignment is supported.

The assignment to the different classes is performed inductively based on the judgement of human experts. So in contrast to a rule-based approach we avoid to explicitly construct rules that define the pragmatic functions. We think that this makes our approach more flexible as the language-system changes. It may be, however, interesting to combine both inductive learning and deductive construction of rules.

2.1 Classifier Induced Semantic Spaces

A *classifier induced semantic space* (CISS) is generated in two steps: In the training phase classification rules $\vec{x}_j \rightarrow z_k$ are inferred from the training data. In the classification phase these decision rules are applied to possibly not-annotated documents.

Any supervised classifier, that internally calculates a quantity and bases its classification on whether this quantity exceeds a given threshold or not, can be employed to construct a CISS. So Linear Classifier, Naive Bayes classifiers as well as Support Vector Machines are applicable for the construction of as classifier induced semantic space. We decided to use Support Vector Machines (SVM)s because of its efficiency and its ability to handle high-dimensional input spaces.

An SVM is a supervised binary classifier, that takes as input a set $E = \{(\vec{x}_1, y_1), \cdots, (\vec{x}_n, y_n)\}$ of positive and negative training examples, where each entity x_i belongs to an instance space X, and each y_i belongs to a set $Y = \{-1, +1\}$ of binary class labels. In the classification phase an SVM generates a numerical value $v(\vec{x})$ for each instance $\vec{x} \in X$. The instance $\vec{x}$ is considered to belong to the positive class $y = +1$ if $v(\vec{x})$ is above a certain threshold, and to belong to the negative class $y = -1$ otherwise.

In order to construct a CISS, the SVM-classifier is used as follows: *In the training phase* for each contribution of the e-discussion a word-frequency vector $\vec{x}$ is computed. Since the SVM is a binary classifier, one separate SVM is trained for each of the K class-labels. To this end we consider all contributions belonging to the category in question as positive examples ($y = +1$), whereas all others are considered as negative examples ($y = -1$). This approach is usually referred to as a 1 vs. $K - 1$ setting.

In the *classification* phase, each of the K SVMs assigns a value $v_k(\vec{x}), k = 1 \ldots K$ to the contribution $\vec{x}$. A document with word-frequency vector $\vec{x}$ is represented by a

vector $\vec{v}(\vec{x}) = (v_1(\vec{x}), \ldots, v_K(\vec{x}))^T$ and the k-th component $v_k(\vec{x})$ can be interpreted as to which degree the instance $\vec{x}$ belongs to class z_k, which in our context means how much contribution $\vec{x}$ fullfils the linguistic function z_k.

This construction of a semantic space is especially useful for practical applications because (1) the space is low-dimensional (up to dozens of dimensions) and thus can easily be visualised, (2) the space's dimension possesses a well defined semantic interpretation, and (3) the space can be tailored to the special requirements of a specific application. (4) Concurrent techniques like latent semantic analysis (LSA, Landauer and Dumais, 1997), probabilistic latent semantic analysis (PLSA, Hofmann, 2001) and hierarchical latent semantic analysis (Paaß et al., 2004) base their notion of semantic nearness on features of the texts i.e. co-occurrences of words, whereas semantic nearness in a CISS is based on the judgement of a human classifier. (5) It is in principle possible to represent units of different semiotic systems (e.g. different languages or even texts and pictures) in one and the same CISS, given that a sufficiently large training set is available.

2.2 Visualisation of Semantic Spaces with Self-Organizing Maps

Self-Organising Maps (SOM) were invented in the early 80s by Kohonen (1980). They use a specific neural network architecture to perform a recursive regression leading to a reduction of the dimension of the data. For practical applications SOMs can be considered as a distance preserving mapping from a more than three-dimensional space to two-dimensions. A description of the SOM algorithm and a thorough discussion of the topic is given in (Kohonen, 1995). After having run the SVM-classification, both labelled and unlabelled contributions were used to build the Self-Organising Map from the the semantic space.

3 The Data

We evaluate our approach on a public e-discussion about an urban planning project, the Berlin Alexanderplatz, Germany. The Berlin Senate office commissioned an Internet-based civic participation in the course of planning the restructuring of one of the great city squares, the Alexanderplatz. An Internet-based discussion bulletin board was established, where citizens could express and discuss their suggestions and preferences with regard to the future shape of the square. The results of the e-discussion have in the meantime been taken into consideration by the city planners. The e-discussion was supervised by several project collaborators acting as moderators. The participants could post messages referring to a list of topics as well as reply to other participant's messages. The moderators used the same means of communication. (Roeder et al., 2005)

All contributions of the participants and moderators were recorded, yielding 1021 messages in total. 216 contributions (21%) have been annotated according to their

pragmatic functions described in section 2. The remaining contributions were left unlabelled.

Table 1: Accuracy of the trained classifiers.

function	precision	recall	F-score
information	72.1	67.4	69.7
objection	56.0	76.1	64.6
question	76.9	61.2	68.2
reply	53.5	84.7	65.6

4 Experimental Results

Training on the annotated data results in four SVMs, which are each trained to separate contributions belonging to one linguistic function (positive class) from all other contributions, which together constitute the negative class. The performance of the classifiers was tested prior to the construction of the semantic space. The results are displayed in table 1. By F-score we refer to the harmonic mean of precision and recall, i.e. $F = 2\left(\frac{1}{prec} + \frac{1}{rec}\right)^{-1}$.

The classification performance is significantly above chance ($F \approx 25\%$). Therefore the classifiers are reliable enough for the construction of a semantic space. The joint classification by the four SVMs produces a 4-dimensional classifier induced semantic space. The four dimensions of this space can be associated with the four pragmatic functions described above.

The 1021 contributions to the e-discussion (both labelled and unlabelled data) were represented in the CISS. These data were used to build the Self-Organising Map as a two-dimensional representation of the semantic space admitting a minimum distortion of the original four-dimensional distances between the data points.

Figure 4 shows an example of a SOM visualising the relations of the contributions in terms of their linguistic functions. SVMs for the four linguistic functions 'question', 'reply', 'information', and 'objections' were trained on 216 contributions (21% of the total contributions) that have been annotated according to their linguistic function. Classification and generation of the SOM was performed for the entire discourse of 1021 contributions.

The contributions of one participant of the e-discussion are displayed by white crosses. The categories are indicated by different grey tones. The SOM algorithm is applied (with 70×70 nodes using Euclidean metric) in order to map the four-dimensional document representations to two dimensions admitting a minimum distortion of the distances. The grey tone indicates the topic category. Shadings within the categories indicate the confidence of the estimated class membership.

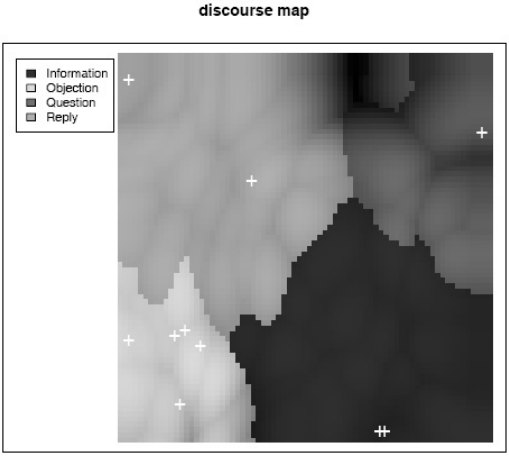

Figure 1: A discourse map generated from a CISS.

We observed that the distribution of authors in the e-discourse does not differ from what is usually observed in the bibliometrical science, namely that the number of authors who publish a given number of papers obeys a power-law. This fact is well known as Lotka's Law (Lotka, 1926). Table 2 shows the frequency distribution that we observed in the Alexanderplatz discussion.

Table 2: Frequency distribution of authors' contributions in the Alexanderplatz e-discussion

♯ contributions	♯ authors	♯contributions	♯ authors
1	73	13	1
2	22	15	1
3	18	24	3
4	4	25	1
5	5	26	1
6	3	29	1
7	3	35	1
9	2	56	1
10	1	140	1
11	2	273	1
12	3		

5 Conclusion

LSA, PLSA, and CISS map documents to the semantic space in a different manner. In the case of LSA the representation of the document in the semantic space is achieved by matrix multiplication. The dimensions of the semantic space correspond to the K largest eigen-values of the covariance matrix. PLSA maps a document to the vector of the conditional probabilities, which indicate how probable aspect z_k is, when a given document is selected. The probabilities are derived from the aspect model using the maximum likelihood principle and the assumption of multinomially distributed word frequency distributions.

The advantage of the presented technique:

1) The use of a supervised classifier makes it possible to produce high-level-features that are relevant to the problem in question (in this case to monitor the discussion process).

2) Note that classifier induced semantic spaces go beyond a mere extrapolation of the annotations found in the training corpus. It gives an insight into how typical a certain document is for each of the classes. Furthermore CISS allow to reveal unseen previously relationships between classes.

3) Concurrent techniques like latent semantic analysis (LSA) (Landauer and Dumais, 1997), probabilistic latent semantic analysis (PLSA) (Hofmann, 2001), and its hierarchical extension (Paaß et al., 2004) base their notion of semantic nearness on features of the texts i.e. co-occurrences of words, whereas semantic nearness in a CISS is based on the judgement of a human classifier.

References

Bühler, K. (1934). *Sprachtheorie*. G. Fischer, Jena.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(6):177–196.

Kohonen, T. (1980). *Content-adressable Memories*. Springer.

Kohonen, T. (1995). *Self-organising Maps*. Springer.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory. *Psychological Review*, 104(2):211–240.

Leopold, E. (2005). On semantic spaces. *LDV-Forum*, 18(3):63–86.

Leopold, E., May, M., and Paaß, G. (2004). Data mining and text mining for science & technology research. In Moed, H. F., Glänzel, W., and Schmoch, U., editors, *Handbook of Quantitative Science and Technology Research*, pages 187–214. Kluwer.

Lotka, A. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, 16(12):317–323.

Paaß, G., Kindermann, J., and Leopold, E. (2004). Learning prototype ontologies by hierarchical latent semantic analysis. In *Workshop on Knowledge Discovery and Ontologies at the joint European Conferences on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2004)*, September 20–24, Pisa (Italy).

Roeder, S., Poppenborg, A., Michaelis, S., Märker, O., and Salz, S. (2005). Public budget dialogue - an innovative approach to e-participation. In Böhlen, M., Gamper, J., Polasek, W., and Wimmer, M., editors, *Proceedings of the International Conference TCGOV 2005, Bolzano (Italy) March 2–4, 2005*, pages 48–56. Springer Lecture Notes in Computer Science, Number 3416.

Turoff, M., Hiltz, S. R., Bieber, M., Fjemestadt, M., and Ajaz, R. (1999). Collaborative discourse structures in computer-mediated group communications. *Journal of Computer-Mediated Communication*, 4:104–125.

Vapnik, V. N. (1998). *An Introduction to Computational Learning Theory*. Wiley & Sons.