Heinrich, Gregor
A generic approach to topic models and its application to virtual communities
(Ein generischer Ansatz für Topic-Modelle und seine Anwendung auf virtuelle Gemeinschaften)
Universität Leipzig, Dissertation
270 + xviii S., 90 Abb., 318 Lit., 5 Anh.

Abstract. This thesis investigates a generic model of topic models in order to facilitate their design and implementation. Topic models are probabilistic representations of grouped discrete data. Applied to text, the basic topic model represents documents as mixtures of topics – probability distributions over the vocabulary. In many cases, there exists a semantic relationship between terms that have high probability within the same topic. This phenomenon, which is rooted in the word co-occurrence patterns in the text, can be used for information retrieval and knowledge discovery in databases, and a large body of work extends the basic topic model, mostly modelling structures in the data beyond term co-occurrence or analysing different data modalities jointly to discover their inter-relations. While these approaches have been very successful individually, an analysis of topic models as a generic model class does not yet exist.

Such an analysis is undertaken in this thesis, based on the conjecture that important properties may be generic across models and that this, in turn, may lead to practical simplifications in the derivation of model properties, inference algorithms and finally design methods. As an exemplary application domain, virtual communities are considered, like those arising in large organisations, the scientific community and the "Social Web".

Work pursues a three-step strategy: In the initial Modelling part, theories of (1) virtual communities and (2) topic models are developed. For virtual communities, this thesis posits that a large part of the their available knowledge can be expressed by three types of entity and their inter-relations: actors (i.e., people or agents), media (i.e., documents and other information sources) and qualities (units of knowledge representation). For topic models, this thesis builds a generic representation that consists of networks of discrete mixtures, "networks of mixed membership" (NoMMs), and shows that this covers a large set of real-world models.

In the subsequent Inference part, generic algorithms for Gibbs sampling and variational inference are developed, revealing a general structure of model properties analogous to the structure of NoMMs. A central result of this work is a Gibbs meta-sampler that allows implementation of inference algorithms from NoMM structures directly. To improve scalability, variations of the generic sampling algorithm are studied that are based on parallelisations and accelerations of the serial algorithm, leading to significant speed-up of model implementations.

The final Application part combines the previous results to a design method that allows composition of topic models from modular NoMM structures, aligning them with structures in the data and monitoring model properties at each construction step. A case study applies this method to a scientific virtual community, and novel models for expert finding are developed that use semantic tags in addition to document text and authorship information to improve retrieval results and topic coherence.