Topic Models for Semantically Annotated Document Collections

Markus Bundschus
Institute for Computer Science
Ludwig-Maximilians-Universität München
Oettingenstr. 67, 80538 München, Germany
bundschu@dbs.ifi.lmu.de

Volker Tresp
Siemens AG, Corporate Technology
Otto-Hahn-Ring 6, 81739 München, Germany
volker.tresp@siemens.com

Hans-Peter Kriegel
Institute for Computer Science
Ludwig-Maximilians-Universität München
Oettingenstr. 67, 80538 München, Germany
kriegel@dbs.ifi.lmu.de

Abstract

Increasingly, web document collections such as PubMed and DBPedia, but also social bookmarking systems, are annotated with semantic meta data. Given that the number of semantically annotated document collections is expected to increase in the near future, it is of interest to analyze if topic models might be able to play a larger role. Since most of the time, annotations are noisy and even human experts annotate inconsistently, a probabilistic view, as provided by topic models, is appropriate. Besides a number of interesting knowledge discovery tasks, representing topics by meta data has an additional advantage: if the concepts refer to real-world objects, the readability of the topics is greatly improved. In this paper, we present several suitable strategies to model this type of data and show experiments on two large semantically annotated document collections.

1 Introduction

Web document collections annotated with meta data such as concepts (potentially from an ontology), named entities, relations extracted from text, or noisy semantic meta data in form of tags, are expected to represent a major fraction of the web in the future. Meta data describe content concisely and support search and information retrieval. On one side we have high-quality annotations generated by trained professionals. An example here is *PubMed*, a huge biomedical collection of abstracts annotated with Medical Subject Headings (MeSH) terms. On the other extreme are meta data or tags generated by a social network community. Here, tags can be chosen freely, they are of lower quality and contain spelling errors and might have other problems as well. Topic models are useful for the (semi-) automated generation of annotations, they can be used to analyze the content of a document corpus, for knowledge discovery in general, for organizing a document corpus in taxonomies and for navigation and browsing in a document corpus. Examples of existing research in the area of topic modeling dealing with meta data are the Correspondence Latent Dirichlet Allocation (LDA) model [2], topic models for entities [8] and topic models for social networks [4]. The work of [7] presents a principled way to model meta data with help of Dirichlet Multinomial Regression. Another interesting approach is to use existing ontologies to improve the predictive performance of topic models for the words in a document collection [5].

In this paper we compare different topic modeling approaches with respect to their ability to model annotations. We apply several suitable existing topic models and compare them on two large docu-

Table 1: Corpora statistics for the two data sets used in this paper.

	PubMed corpus	**CiteULike corpus**
Documents	**50.000**	**18.628**
Unique Words	**22.531**	**14.489**
Total Words	**2.369.616**	**1.161.794**
Unique Meta Data	**17.716**	**3.411**
Total Meta Data	**470.101**	**125.808**
Unique Users	**—**	**1.393**
Total Users	**—**	**18.628**

ment collections: PubMed and CiteULike[1]. PubMed is the largest biomedical document collection today, consisting of about 17 million abstracts most of them annotated with MeSH terms. There are approx. 22.000 MeSH terms arranged in a taxonomy. PubMed annotations are of high quality. CiteULike, is a social bookmarking system that allows researchers to manage their scientific reference articles. Researchers upload references they are interested in and assign tags to the reference. Since users can annotate freely and are not forced to use a specific vocabulary, annotations are noisy and error prone. The rest of the paper is organized as follows: in Section 2 we briefly describe the topic models used in our analysis. Section 3 gives information about the data sets, a perplexity analysis on both corpora and a user modeling analysis on CiteULike. Section 4 provides conclusions.

2 Models

Here, we briefly describe the models used in the experiments. For all models, we use Gibbs Sampling for inference.
LDA[1]: Instead of deriving a topic model of the word tokens in the document —as it is done in classical LDA— we form a topic model of only the meta data of the document. Depending on the type of meta data, we refer to this model as *Tag LDA* or *Concept LDA*. With this model we can analyze the topic structure of the meta data and we can predict additional meta data given a set of previously assigned meta data.

Link-LDA[6]: Originally, the purpose here was to model the relationship between document content and document hyperlinks. Applied to our context, a LDA models the document-word topic structure and a second latent structure models the meta data distribution. Mutual coupling is achieved since one single multinomial distribution Θ is used to assign topics to words and to assign topics to meta data (see [6] for more information).

Topic-Concept LDA: Topic-Concept (TC) LDA provides a principled coupling between the topic distribution of a document and its meta data. First, a standard LDA step is performed, where each word is assigned to a topic. Second, a word index is uniformly sampled and the current topic assignment of that word index is used to sample a concept from a concept-specific multinomial distribution. Each concept is conditioned on the topic that generated the uniformly sampled word. This principled coupling has been sucessfully apllied to modeling images and their captions [2].

User-Topic-Tag LDA [3]: Finally we use our recently developed model for collaborative tagging systems, which can model the most important entities in social bookmarking system, i.e., the users, its resources and their corresponding tags. Users are modeled in a similar way as authors in a related publication [9]. Tags are sampled with the same generative process as in [2]. The User-Topic-Tag LDA showed encouraging results in a personalized tag recommendation task by creating a personalized view on a document by sampling the document-specific topic distribution through the user-specific topic distribution.

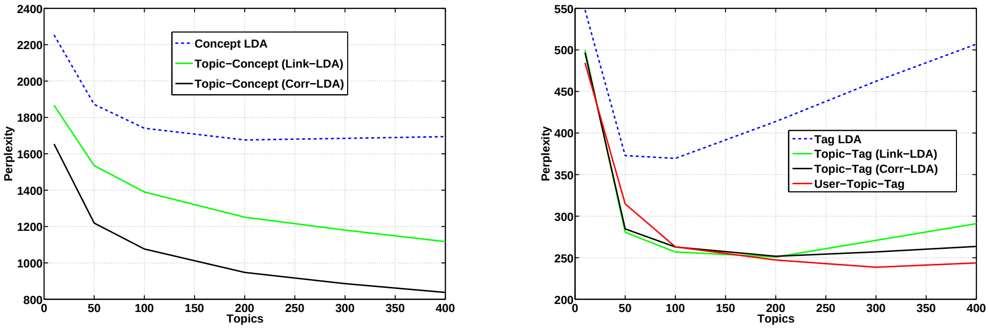
3 Experiments

The first data set consists of PubMed abstracts randomly selected from the MEDLINE 2006 baseline database provided by the NLM. The second data set is a snapshot provided by the social bookmarking system CiteULike. Word tokens from title and main text were stemmed with a Porter stemmer and stop words were removed. In both data sets, word tokens occurring less than five times were filtered out. Table 1 summarizes the corpus statistics.

[1] http://www.citeulike.org/

Table 2: Selected topics from a TC model with T=200. Top words and MeSH concepts are shown.

HIV Topic				Phosphorylation Topic			
Word	Prob.	Concept	Prob.	Word	Prob.	Concept	Prob.
viru	**0.118**	**Humans**	**0.06**	**phosphoryl**	**0.130**	**Phosphorylation**	**0.123**
viral	**0.064**	**HIV-1**	**0.06**	**kinas**	**0.118**	**Prot.-Serine-Threon. Kin.**	**0.075**
infect	**0.058**	**HIV Infections**	**0.059**	**activ**	**0.060**	**Proto-Oncogene Prot.**	**0.060**
hiv-1	**0.047**	**Virus Replication**	**0.045**	**akt**	**0.060**	**Proto-Oncogene Proteins c-akt**	**0.047**
virus	**0.035**	**RNA, Viral**	**0.042**	**tyrosin**	**0.036**	**1-Phosphatidylinositol 3-Kin.**	**0.047**



(a) *Pubmed Corpus*

(b) *CiteULike Corpus*

Figure 1: Perplexity on the test set. 50% of the meta data per document were chosen as held-out.

Parameters were estimated by averaging samples from ten randomly-seeded runs, each running over 100 iterations, with an initial burn-in phase ranging from 500 to 1.500 depending on the trained model. We found the number of burn-in iterations to be a convenient choice by observing a flattening of the log likelihood. Instead of estimating the hyperparameters α, β and γ, we fix them to 50/T, 200/W and 200/C respectively in each of the experiments (W words, C concepts).

3.1 Concept/Tag Perplexity and Topic Structure

We measure meta data quality in terms of perplexity and follow the evaluation procedure of [9]. All perplexity values were computed by averaging over ten different samples. Figure 1 plots the perplexity over the held-out meta data of each model for different values of T. We observe that the models, which include the word tokens into the computation of the likelihood clearly outperform the standard LDA model, which only analyzes the structure in the annotations. On the PubMed Corpus the Corr-LDA model performs much better than the Link-LDA model. This is still the case for the CiteULike corpus, but the difference in performance is smaller.

Table 2 shows word probabilities and concept probabilities for two typical topics found in the PubMed data set. Clearly, the concept representation is much more representative for the HIV topic, resp. Phosporylation topic if compared to the assigned words. While the HIV topic might be quite intuitive for the reader, the Phosphorylation topic is sensible, as well: the concept representation gives much more detailed information.

Since the tags in the social bookmarking system CiteULike were chosen freely and by non-professionals, the tags which are ranked highly for a topic were still quite expressive but were somewhat more noisy. An interesting observation can be made: top scored topic tags often contain identical terms with different spellings or terms and their abbreviations (consider e. g. IR vs. information retrieval). All learned topics are available online[2].

3.2 Modeling Users in Social Bookmarking Systems

Here, we validate if the modeling assumption made for the users holds in the User-Topic-Tag LDA model. We identify all users in our data set which are members of groups in CiteULike. CiteULike

[2] `http://www.dbs.ifi.lmu.de/~bundschu/NIPS_WS_supplementary/info.html`

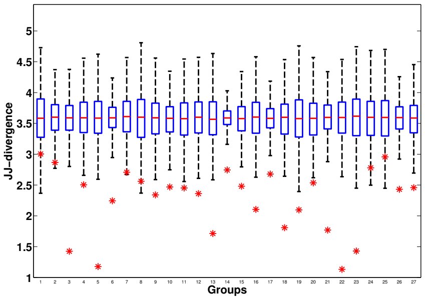


Figure 2: Boxplot over 1000 random samplings. The stars indicate the true group divergence.

groups typically share similar research interests. There are 488 users which belong to a total of 524 groups (as of November 18, 2008). We excluded all groups with less than five members. This resulted in a total of 27 groups with 160 users. 31 user belong to more than one group and the maximum number of groups for one user is five. We derive the similarity between users based on the learned user-topic distributions Θ_u. Jeffreys‘ J-divergence, a symmetric version of the Kullback-Leibler (KL) divergence, is used. Users that share the same group membership should be significantly more similar to each other than users that are randomly chosen and considered as an artificial group. Therefore, we repeat the following procedure for each group: we randomly sample n users (with n, the size of a group) and compute the mean divergence of this artificial group. This step is repeated 1000 times. These results are compared to the true group divergence. Figure 2 shows an example boxplot for each group ($T = 200$). On each box, the central red line is the median, the blue edges are the 25th and 75th percentiles. The whiskers were chosen such that all data points within $\pm 2.7\sigma$ are considered not as outliers. The stars in the plot indicate the true divergence for each group. All true group divergences fall clearly below the just mentioned percentiles. Note that this result holds for various number of topics (Results not shown for the sake of brevity).

4 Conclusion

We analyzed various topic models in the context of semantically annotated documents. In terms of perplexity, Corr-LDA shows best performance in modeling the high-quality annotations of the PubMed data. The assignment of concepts (MeSH terms) to topics is sensible and provides a much better description of the content of a topic than keywords derived from the abstracts. For the CiteULike data, the Corr-LDA models and the User-Topic-Tag LDA showed all good performance. The later exhibited slightly better performance indicating that personalized models in a hierarchical Bayesian context can be beneficial. We could also show that the user similarity derived from User-Topic-Tag LDA fits well with the structure of user groups in CiteULike.

References

[1] D. M. Blei et al. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[2] D. M. Blei et al. Modeling annotated data. *SIGIR Forum*, (SPEC. ISS.):127–134, 2003.

[3] M. Bundschus et al. Hierarchical bayesian models for collaborative tagging systems. In *ICDM 2009*.

[4] J. Chang et al. Connections between the lines: Augmenting social networks with text. In *KDD 2009*.

[5] C. Chemudugunta et al. Modeling documents by combining semantic concepts with unsupervised statistical learning. In *ISWC 2008*.

[6] E. Erosheva et al. Mixed-membership models of scientific publications, 2004.

[7] D. Mimno et al. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI 2008*.

[8] D. Newman et al. Statistical entity-topic models. In *KDD 2006*, New York, NY, USA, 2006.

[9] M. Steyvers et al. Probabilistic author-topic models for information discovery. In *KDD 2004*.