Pragmatic metadata matters: How data about the usage of data effects semantic user models

Claudia Wagner[1], Markus Strohmaier[2], and Yulan He[3]

[1] JOANNEUM RESEARCH, Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
claudia.wagner@joanneum.at
[2] Graz University of Technology and Know-Center
Inffeldgasse 21a, 8010 Graz, Austria
markus.strohmaier@tugraz.at
[3] The Open University, KMi
Walton Hall, Milton Keynes MK7 6AA, UK
yhe@open.ac.uk

Abstract. Online social media such as wikis, blogs or message boards enable large groups of users to generate and socialize around content. With increasing adoption of such media, the number of users interacting with user-generated content grows and as a result also the amount of *pragmatic metadata* - i.e. data about the usage of content - grows.
The aim of this work is to compare different methods for learning topical user profiles from Social Web data and to explore if and how pragmatic metadata has an effect on the quality of semantic user models. Since accurate topical user profiles are required by many applications such as recommender systems or expert search engines, learning such models by observing content and activities around content is an appealing idea.
To the best of our knowledge, this is the first work that demonstrates an effect between pragmatic metadata on one hand, and the quality of semantic user models based on user-generated content on the other. Our results suggest that *not all types of pragmatic metadata are equally useful* for acquiring *accurate* semantic user models, and some types of pragmatic metadata can even have detrimental effects.

Keywords: Semantic Analysis, Social Web, Topic Models, User Models

1 Introduction

Online social media such as Twitter, wikis, blogs or message boards enable large groups of users to create content and socialize around content. When a large group of users interact and socialize around content, *pragmatic metadata* is produced as a side product. While *semantic metadata* is often characterized as *data about the meaning of data*, we define *pragmatic metadata* as *data about the usage of data*. Thereby, pragmatic metadata captures how data/content is used

by individuals or groups of users - such as who authored a given message, who replied to messages, who "liked" a message, etc. Although the amount of pragmatic metadata is growing, we still know little about how these metadata can be exploited for understanding the topics users engage with.

Many applications, such as recommender systems or intelligent tutoring systems, require good user models, where "'good"' means that the model accurately reflects user's interest and behavior and is able to predict future content and activities of users. In this work we explore to what extent and how pragmatic metadata may contribute to semantic models of users and their content and compare different methods for learning topical user profiles from Social Web data.

To this end, we use data from an online message board. We incorporate different types of pragmatic metadata into different topic modeling algorithms and use them to learn topics and to annotate users with topics. We evaluate the quality of different semantic user models by comparing their predictive performance on future posts of user. Our evaluation is based on the assumption that "better" user models will be able to predict future content of users more accurately and will need less time and training data.

Generative probabilistic models are a state of the art technique for unsupervised learning. In such models, observed and latent variables are represented as random variables and probability calculus is used to describe the connections that are assumed to exist between these variables. Only if the assumptions made by the model are correct, Bayesian inference can be used to answer questions about the data. Generative probabilistic models have been successfully applied to large document collections (see e.g. [1]). Since for many documents one can also observe metadata, several generative probabilistic models have been developed which allow exploiting special types of metadata (see e.g., the Author Topic model [10], the Author-Recipient Topic model [8], the Group Topic model [14] or the Citation Influence Topic model [2]). However, previous research [10] has also shown that incorporating metadata into the topic modeling process may lead to model assumptions which are too strict and might overfit the data. This means that incorporating metadata does not necessarily lead to "better" topic models, where "better" means, for example, that the model is able to predict future user-generated content more accurately and needs less trainings data to fit the model.

Our work aims to advance our understanding about the effects of pragmatics on semantics emerging from user-generated content and specifically aims to answer the following questions:

1. Does incorporating pragmatic metadata into topic modeling algorithms lead to more accurate models of users and their content and if yes, what types of pragmatic metadata are more useful?
2. Does incorporating behavioral user similarities help acquiring more accurate models of users and their content and if yes, which types of behavioral user similarity are more useful?

The remainder of the paper is organized as follows: Section 2 gives an overview of the related work, while Section 3 describes our experimental setup. In Section 4 we report our results, followed by a discussion of our findings in Section 5.

2 Related Work

From a machine learning perspective, social web applications such as Boards.ie provide a huge amount of unlabeled training data for which usually many types of metadata can be observed. Several generative probabilistic models have been developed which allow exploiting special types of metadata (such as the Author Topic model [10], the Author-Recipient Topic model [8], the Group Topic model [14] or the Citation Influence Topic model [2]). In contrast to previous work where researchers focused on creating new topic models for each type of metadata, [9] presents a new family of topic models, Dirichlet-Multinomial Regression (DMR) topic models, which allow incorporating arbitrary types of observed features . Our work builds on the DMR topic model and aims to explore the extent to which different types of pragmatic metadata contribute to learning topic models from user generated content.

In addition to research on advancing topic modeling algorithms, the usefulness of topic models has been studied in different contexts, including social media. For example, [5] explored different schemes for fitting topic models to Twitter data and compared these schemes by using the fitted topic model for two classification tasks. As we do in our work, they also point out that models trained with a "'User"' scheme (i.e., using post aggregations of users as documents) perform better than models trained with a "'Post"' scheme. However, in contrast to our work they only explore relatively simple topic models and do not take any pragmatic metadata (except authorship information) into account when learning their models.

In our own previous work, we have studied the relationship between pragmatics and semantics in the context of social tagging systems. We have found that, for example, the pragmatics of tagging (users' behavior and motivation in social tagging systems [11, 6, 4]) exert an influence on the usefulness of emergent semantic structures [7]. In social awareness streams, we have shown that different types of Twitter stream aggregations can significantly influence the result of semantic analysis of tweets [12]. In this paper, we extend this line of research by (i) applying general topic models and (ii) using a dataset that offers rich pragmatic metadata.

3 Experimental Setup

The aim of our experiments is to explore to what extent and how pragmatic metadata can be exploited when semantically analyzing user generated content.

3.1 Dataset

The dataset used for our experiments and analysis was provided by Boards.ie,[4] an Irish community message board that has been in existence since 1998. We used all messages published during the first week of February 2006 (02/01/2006 - 02/07/2006) and the last week of February 2006 (02/21/2006 - 02/28/2006). We only used messages authored by users who published more than 5 messages and replied to more than 5 messages during this week. While we performed our experiments on both datasets, the results are similar. Consequently, we focus on reporting results obtained on the first dataset which consists of 1401 users and 27525 posts which were authored by these users and got replies.

To assess the predictive performance of different topic models we estimate how well they are able to predict the content (i.e. the actual words) of future posts. We generated a test corpus of 4007 held out posts in the following way: for each of the 1401 user in our training corpus we crawled 3 future posts which were authored by them and to which at least one user of our training corpus has replied. From here on, we refer to this data has *hold-out* data.

3.2 Methodology

In this section we first introduce the topic modeling algorithms (LDA, AT-model and DMR topic model) on which our work is based and then proceed to describe the topic models which we fitted to our training data, their model assumptions and how we compared and evaluated them.

Latent Dirichlet Allocation (LDA) The idea behind LDA is to model documents as mixtures of topics and force documents to favor few topics. Therefore, each document exhibits different topic proportions and each topic is defined as a distribution over a fixed vocabulary of terms. That means the generation of a collection of documents is modeled as a three step process: First, for each document d a distribution over topics θ_d is sampled from a Dirichlet distribution α. Second, for each word w_d in the document d, a single topic z is chosen according to this distribution θ_d. Finally, each word w_d is sampled from a multinomial distribution over words ϕ_z which is specific for the sampled topic z.

The Author Topic (AT) model The Author Topic model [10] is an extension of LDA, which learns topics conditioned on the mixture of authors that composed the documents. The assumption of the AT model is that each document is generated from a topic distribution which is specific to the set of authors of the document. The observed set of variables are the words per document (similar as in LDA) and the authors per document. The latent variables which are learned by fitting the model, are the topic distribution per author (rather than the topic distribution per document as in LDA) and the word distribution per topic.

[4] `http://www.boards.ie/`

We implemented the AT-model based on Dirichlet-multinomial Regression (DMR) Models (explained in the next section). While the original AT-model uses multinomial distribution (which are all drawn from the same Dirichlet) to represent an author-specific topic distributions, the DMR-model based implementation uses a "fresh" Dirichlet prior for each author from which then the topic distribution is drawn.

Dirichlet-multinomial Regression (DMR) Models Dirichlet-multinomial regression (DMR) topic models [9] assume not only that documents are generated by a latent mixture of topics but also that mixtures of topics are influenced by an additional factor which is specific to each document. This factor is materialized via observed features (in our case pragmatic metadata such as authorship or reply user information) and induce some correlation across individual documents in the same group. This means that e.g. documents which have been authored by the same user (i.e., they belong to one group) are more likely to chose the same topics. Formally, the prior distribution over topics α is a function of observed document features, and is therefore specific to each distinct combination of feature values. In addition to the observed features we add a default feature to each document, to account for the mean value of each topic.

Fitting Topic Models In this section we describe the different topic models which we fitted to our training datasets (see table 1 and 2). Each topic model makes different assumptions on what a document is (see column 3), takes different types of pragmatic metadata into account (see column 4) and makes different assumptions on the document-specific topic distributions θ which generates each documents (see column 5).

For all models, we chose the standard hyperparameters which are optimized during the fitting process: $\alpha = 50/T$ (prior of the topic distributions), $\beta = 0.01$ (prior of the word distributions) and $\sigma^2 = 0.5$ (variance of the prior on the parameter values of the Dirichlet distribution α). For the default features $\sigma^2 = 10$. Based on the empirical findings of [13], we decided to place an asymmetric Dirichlet prior over the topic distributions and a symmetric prior over the distribution of words. All models share the assumption that the total number of topics used to describe all documents of our collection is limited and fixed (via hyperparameter T) and that each topic must favor few words (as denoted by hyperparameter β which defines the Dirichlet distribution from which the word distributions are drawn - the higher β the less distinct the drawn word distributions).

Following the model selection approach described in [3], we selected the optimal number of topics for our training corpus by evaluating the probability of held out data for various values of T (keeping $\beta = 0.01$ fixed). For both datasets (each represents one week boards.ie data), a model trained on the "'Post"' scheme (i.e., using each post as a document) gives on average (over 10 runs) the highest probability to held out documents if $T = 240$ and model trained on the "'User"' scheme (i.e., using all posts authored by one user as a document) gives on av-

erage (over 10 runs) the highest probability to held out documents if $T = 120$. We kept T fixed for all our experiments.

Evaluation of Topic Models To compare different topic models we use perplexity which is a standard measure for estimating the performance of a probabilistic model. Perplexity measures the ability of a model to predict words on held out documents. In our case a low perplexity score may indicate that a model is able to accurately predict the content of future posts authored by a user. The perplexity measure is defined as followed:

$$perplexity(d) = exp[-\frac{\sum_{i=0}^{N_d} lnP(w_i|\hat{\phi}, \alpha)}{N_d}] \quad (1)$$

In words, the perplexity of a held out post d is defined as the exponential of the negative normalized predictive likelihood of the words w_i of the held out post d (where N_d is the total number of words in d) conditioned on the fitted model.

ID	Alg	Doc	Metadata	Model Assumption
M1	LDA	Post	-	A post is generated by a mixture of topics and has to favor few topics.
M2	LDA	User	-	All posts of one user are generated by a mixture of topics and have to favor few topics.
M3	DMR	Post	author	A post is generated by a user‘s authoring-specific mixture of topics and a user has to favor few topics he usually writes about.
M4	DMR	User	author	All posts of one user are generated by a user‘s authoring-specific mixture of topics and a user has to favor few topics he usually writes about.
M5	DMR	Post	user who replied	A post is generated by a user‘s replying-specific mixture of topics and a user has to favor few topics he usually replies to.
M6	DMR	User	user who replied	All posts of one user are generated by a user‘s replying-specific mixture of topics and a user has to favor few topics he usually replies to.
M7	DMR	Post	related user	A post is generated by a user‘s authoring- or replying-specific mixture of topics and a user has to favor few topics he usually replies to and he usually writes about.

M8	DMR	User	related user	All posts of one user are generated by a user‘s authoring- or replying-specific mixture of topics and a user has to favor few topics he usually replies to and he usually writes about.

Table 1: Overview about different topic models which incorporate different types of pragmatic metadata.

ID	Alg	Doc	Metadata	Model Assumption
M9	DMR	Post	top 10 forums of author	A post is generated by a mixture of topics which is specific to users who show a similar forum usage behavior as the author of the post.
M10	DMR	User	top 10 forums of author	All posts are generated by a mixture of topics which is specific to users who show a similar forum usage behavior as the author of the post-aggregation.
M11	DMR	Post	top 10 communication partner of author	A post is generated by a mixture of topics which is specific to users who show a similar communication behavior as the author of the post.
M12	DMR	User	top 10 communication partner of author	All posts are generated by a mixture of topics which is specific to users who show a similar communication behavior as the author of the post-aggregation.

Table 2: Overview about different topic models which incorporate different types of smooth pragmatic metadata based on behavioral user similarities.

4 Experimental Results

Our experiments were set up to answer the following questions:

1. Does incorporating pragmatic metadata into topic modeling algorithms lead to more accurate models of users and their content and if yes, what types of pragmatic metadata are more useful?

To answer this question, we fit different models to our training corpus and tested their predictive performance on future posts authored by our trainings users.

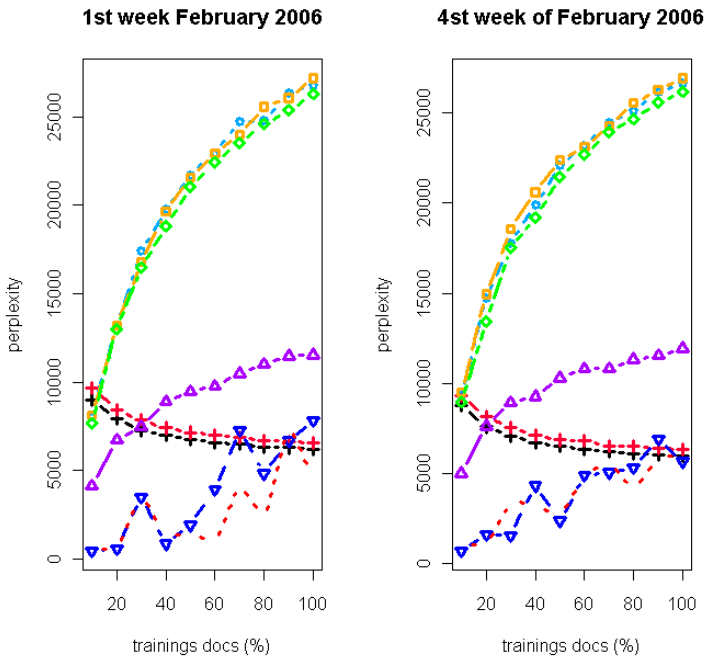

Fig. 1. Comparison of the predictive performance of different topic models on held out posts. The y-axis shows the average perplexity (over 10 runs) and the x-axis indicates the percentage of whole dataset used as training data. As baseline we use 2 versions of LDA (M1and M2).

Figure 1 shows that the predictive performance of semantic models of users which are either solely based on the users (i.e., aggregations of users‘ posts) to whom these users replied (M6) or which take in addition also the content authored by these users (M8) into account, is best. Therefore, our results suggest that it is beneficial to take user‘s reply behavior into account when learning topical user profiles from user generated content.

We also noted that all models which use the “User” training scheme (M4, M6 and M8) perform better than the models which use the “Post” training scheme (M3, M5 and M7). One possible explanation for this is the sparsity of posts which consist of only 66 tokens on average.

Since we were interested in how the predictive performance of different models change depending on the amount of data and time used for training, we split our training dataset randomly into smaller buckets and fitted the model on different proportions of the whole training corpus. One would expect that as the percentage of training data increases the predictive power of each model would improve as it adapts to the dataset. Figure 1 however shows that this is only true for our baseline models M1 and M2 which ignore all metadata of posts. The model M3 which corresponds to the Author Topic model exhibits a behavior that is similar to the behavior reported in [10]: When observing only few training data, M3 makes more accurate predictions on held-out posts than our baseline models. But the predictive performance of the model is limited by the strong assumptions that future posts of one author are about the same topics as past posts of the same author. Like M3, also M5 (and M7) seem to over-fit the data by making the assumptions that future posts of a user will be about the same topics as posts he replied to in the past (and posts he authored in the past).

To address these over-fitting problems we decided to incorporate smoother pragmatic metadata into the modeling process which we get by exploiting behavioral user similarities. The pragmatic metadata we used so far capture information about the usage behavior of individuals (e.g., who authored a document), while our smoother variants of pragmatic metadata capture information about the usage behavior of groups of users which share some common characteristics (e.g., what are the forums in which the author of this document is most active). Our intuition behind incorporating these smoother pragmatic metadata which are based on user similarities is that users which behave similar tend to talk about similar topics.

2. Does incorporating behavioral user similarities help acquiring more accurate models of users and their content and if yes, which types of behavioral user similarity are more useful?

From Figure 2 one can see that indeed all models which incorporate behavioral user similarity exhibit lower perplexity than our baseline models, especially if only few training samples are available. The model M12, which is based on the assumption that users who talk to the same users talk about the same topics, exhibits the lowest perplexity and outperforms our baseline models in terms of their predictive performance on held out posts.

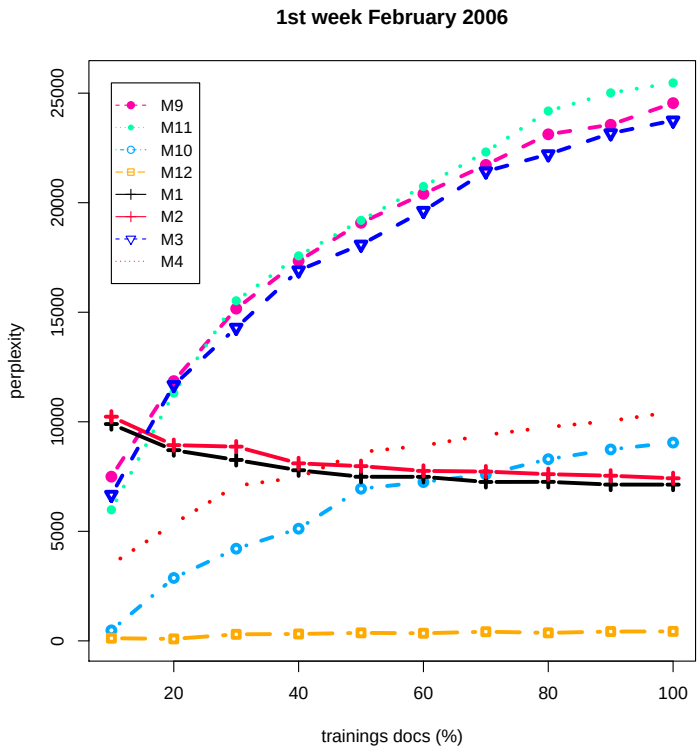

Fig. 2. Comparison of the predictive performance of topic models which take smooth pragmatic metadata into account by exploiting user similarities. The y-axis shows the average perplexity (over 10 runs) and the x-axis indicates the percentage of whole dataset used as training data. As baseline we use 2 versions of LDA (M1and M2).

For the model M10 which assumes that users who tend to post to the same forums talk about the same topics, we can only observe a lower perplexity than our baseline models when only few trainings data are available, but it still outperforms other state of the art topic models such as the Author topic model.

5 Discussion of Results and Conclusion

While it is intuitive to assume that incorporating metadata about the pragmatic nature of content leads to better learning algorithms, our results show that not all types of pragmatic metadata contribute in the same way. Our results confirm previous research which showed that topic models which incorporate pragmatic metadata such as the author topic model tend to over-fit data. That means incorporating metadata into a topic model can lead to model assumptions which are too strict and which yield the model to perform worse.

Summarizing, our results suggest that:

- **Pragmatics of content influence its semantics:** Integrating pragmatic metadata information into semantic user models influences the quality of resulting models.
- **Communication behavior matters:** Taking user's reply behavior into account when learning topical user profiles is beneficial. Content of users to which a user replied seems to be even more relevant for learning topical user profiles than content authored by a user.
- **Behavioral user similarities improve user models:** Smoother versions of metadata based topic models which take user similarity into account always seem to improve the models.
- **Communication behavior based similarities matter:** Different types of proxies for behavioral user similarity (e.g., number of forums they both posted to, number of shared communication partners) lead to different results. User who have a similar communication behavior seem to be more likely to talk about the same topics, than users who post to similar forums.

Acknowledgments. The authors want to thank Boards.ie for providing the dataset used in our experiments and Matthew Rowe for pre-processing the data. Furthermore we want to thank David Mimno for answering questions about the DMR topic model and Sofia Angelouta for fruitful discussions. Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
2. Dietz, L., Bickel, S., Scheffer, T.: Unsupervised prediction of citation influences. In: International Conference on Machine Learning. pp. 233–240 (2007)
3. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101(Suppl. 1), 5228–5235 (April 2004)

4. Helic, D., Trattner, C., Strohmaier, M., Andrews, K.: On the navigability of social tagging systems. In: The 2nd IEEE International Conference on Social Computing (SocialCom 2010), Minneapolis, Minnesota, USA. pp. 161–168 (2010)
5. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics. pp. 80–88. SOMA '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1964858.1964870
6. Koerner, C., Kern, R., Grahsl, H.P., Strohmaier, M.: Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In: 21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT 2010), Toronto, Canada, ACM. ACM, New York, NY, USA (June 2010)
7. Koerner, C., Benz, D., Strohmaier, M., Hotho, A., Stumme, G.: Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In: Proc. of the 19th International World Wide Web Conference (WWW 2010). ACM, Raleigh, NC, USA (Apr 2010), http://www.kde.cs.uni-kassel.de/benz/papers/2010/koerner2010thinking.pdf
8. Mccallum, A., Corrada-Emmanuel, A., Wang, X.: The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. Tech. rep., UMass CS (December 2004)
9. Mimno, D., McCallum, A.: Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI '08) (2008), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.140.6925
10. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. pp. 487–494. UAI '04, AUAI Press, Arlington, Virginia, United States (2004), http://portal.acm.org/citation.cfm?id=1036843.1036902
11. Strohmaier, M., Koerner, C., Kern, R.: Why do users tag? Detecting users' motivation for tagging in social tagging systems. In: International AAAI Conference on Weblogs and Social Media (ICWSM2010), Washington, DC, USA, May 23-26. AAAI, Menlo Park, CA, USA (2010)
12. Wagner, C., Strohmaier, M.: Exploring the wisdom of the tweets: Knowledge acquisition from social awareness streams. In: Proceedings of the Semantic Search 2010 Workshop (SemSearch2010), in conjunction with the 19th International World Wide Web Conference (WWW2010), Raleigh, NC, USA, April 26-30, ACM (2010)
13. Wallach, H.M., Mimno, D., McCallum, A.: Rethinking LDA: Why priors matter. In: Proceedings of NIPS (2009), http://books.nips.cc/papers/files/nips22/NIPS2009_0929.pdf
14. Wang, X., Mohanty, N., Mccallum, A.: Group and topic discovery from relations and text. In: In Proc. 3rd international workshop on Link discovery. pp. 28–35. ACM (2005)