CTI Project: Deep Networks as a Semantic Platform for Modeling User Behavior Data Milestone 2

1plusX AG and ETH Zürich

August 16, 2017

1 Introduction

In this CTI project we intend to model user behavior on the web. In the first milestone, we took a closer look at embeddings of data in multiple languages, embeddings of documents (a collection of words), and embeddings that represent a user based on his/her website visits.

In the second milestone documented in this report, we extend our work on user embeddings by analyzing different features based on a website's content. We then evaluate the trained embeddings on a demographics prediction task.

2 User Embeddings

We propose to represent a user by the sites he or she visited. Similar to a document, which is a collection of words that describe its topics, a user can be seen as a collection of website visits that describe the user's interests. We therefore employ the Neural Variational User Model (NVUM) based on [1] introduced in milestone 1, and further investigate different input features in the following.

2.1 Recap: Embeddings from Site Visits

In milestone 1, we use a user's site visits as input features to the NVUM. This model assumes that the identity of the website alone provides enough information to get an accurate picture of a user's interests. The data for the website visits is binary for the moment (has visited the site or not), and does not take into account the length of the visit, what parts of the website have been looked at or any interactions with the website (clicking links or filling in forms).

2.2 Embeddings from Site Content

A different model puts the emphasis on the textual content of a website. In this model, a user is seen as the collection of all words on the visited websites, similar

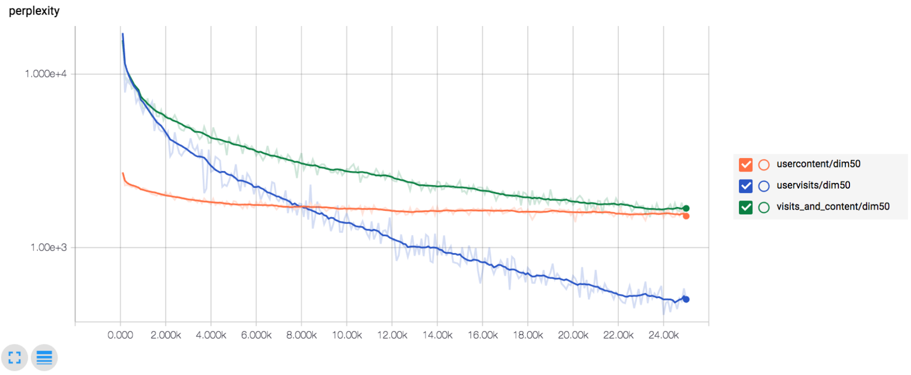

Figure 1: Perplexity during training for different user features.

to a giant document. This assumption makes the similarity to document models even clearer, although the resulting *user documents* are potentially much larger (and more diverse) than regular documents, such as a single website.

2.3 Embeddings from Visits and Content

Finally, we also investigate a model that combines the site information identity with its content by simple concatenation of the features. Since the feature dimensions have no explicit meaning in the used document models anyway, this can be done very naturally.

2.4 Experiments

We performed several experiments to evaluate the proposed models. We found that all models proved to be very robust to selection of hyperparameters, and did not show much difficulty in the training process. We summarize the findings of the most interesting experiments below.

Data. The data contains the website visits of 60M users in the month of June to various publisher's websites. The number of visits varies greatly between users, and so does the number of distinct sites these visits have been to. In preprocessing, we only keep the sites with at least 100 visits from at least 10 distinct users. Vice versa, we keep users with at least 10 visits to at least 5 different sites. This leaves us with 207k sites and 9M users. Of these 9M users, we randomly sample 100k to reduce training times and allow for more experimentation.

Different input features. First off, we compare the perplexity (a measure for how well the model has captured the true data distribution; lower is better) of the different input features. The results are shown (in log-scale) in Figure 1.

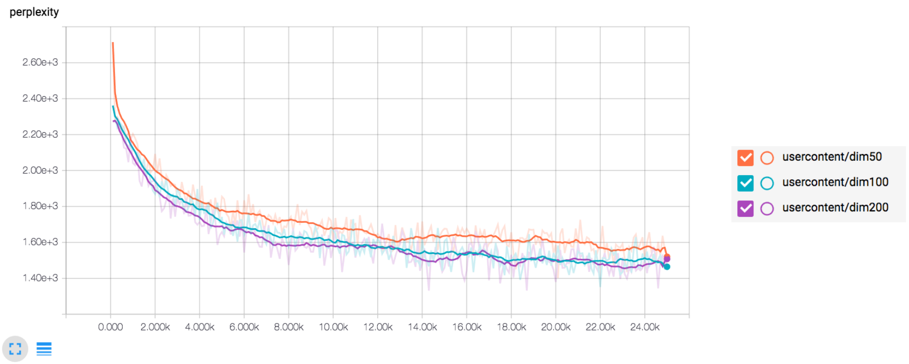

Figure 2: Perplexity of content features during training for different latent dimensionality.

While the content features start out at a lower perplexity, the model only slowly improves its estimate of the data distribution. In comparison, the visit-based model shows stronger learning. The combination of input features translates straightforwardly into the perplexity evaluation as approximately the addition of the two separate models (keep in mind the log-scale of the y-axis).

Latent dimensions for content features. Since the NVUM operates as an autoencoder, it compresses the input signal and tries to reconstruct it from the compressed latent representation. For our next experiment, we hypothesize that due to the larger input signal (one site visit corresponds to many content words on that site) for the content features, the compression to 50 dimensions is too restrictive. In Figure 2, we investigate increasing the latent dimensionality to 100 and 200. We indeed observe lower perplexity than with 50 dimensions, but diminishing returns set in when moving to 200 dimensions. More importantly, learning does not seem to be enabled any more than with 50 dimensions, so we stick to the original setting for the remainder of this report.

Decreasing vocabulary size for sites. The vocabulary size for content words is restricted to the 50k most common words, whereas the equivalent vocabulary for sites is 207k. Although the user-visit matrix is much sparser than the user-word matrix, we want to investigate the effect of reducing the site vocabulary to 50k as well. The content-based method is only affected by this vocabulary reduction insofar that only visits to the 50k most common sites are included in creating a user's read content. In Table 1 we list the perplexity on a held-out test set of users for the three models and the different vocabulary sizes. As expected, the visit-based model finds it much easier to predict a user's visit distribution. The content-based model is unaffected, showing that the uncommon sites do not pose the biggest problems to model learning, presumably because they are easier or similarly difficult to model than the more common

Features	Perplexity
Site visits (50k sites)	519.143
Site visits (full; 207k sites)	908.130
Site content (50k sites)	1569.014
Site content (full)	1569.786
Visits and content (full)	2279.643

Table 1: Perplexity on hold-out test set.

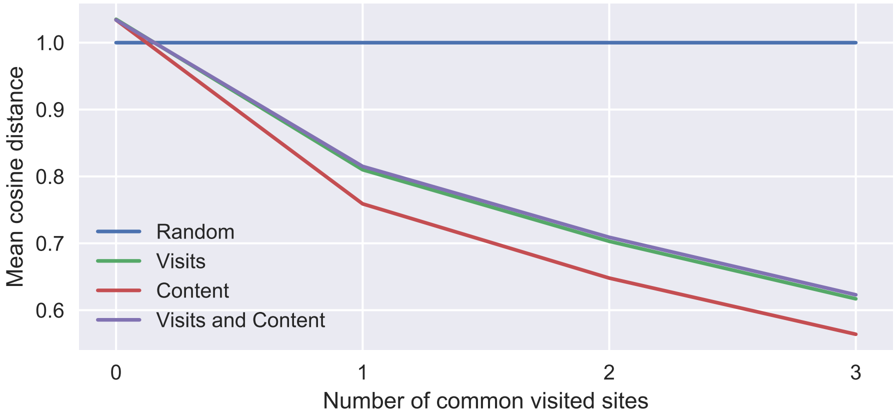

Figure 3: Proximity of test users for different user features.

sites. Again, the concatenation of features also adds to the complexity of learning the distribution. We decide to keep the entire site vocabulary to not miss out on sites that could be less popular but highly discriminative for users.

Proximity. Finally, we employ the clustering metric from milestone 1, called proximity, that compares the mean cosine distances between users with 0 to 3+ common site visits. From Figure 3, we see that all input features cluster users with more common visits more closely together. Content features do an especially good job on this task. Conversely, they do not improve the performance of the joint feature model. It seems that the discriminative characteristics of the content model for this task are not incorporated into the latent representations of the combined input features.

3 User Demographics Prediction

With the findings from Section 2, we set out to evaluate the user embeddings, that were learned in an unsupervised manner, in a supervised classification task. We employ the task of predicting the demographics of a user.

Features	Gender	Age
Random	50.00%	14.29%
Site visits	74.34%	33.59%
Site content	76.66%	32.01%
Visits and content	78.76%	34.90%

Table 2: Accuracy in predicting the demographics of users.

3.1 Experiments

We use anonymized ground truth data from the same time period of our training data for the user embeddings. In total, around 19k users appear in our ground truth and have visited a publisher's website in the month of June. We look at two demographic properties, namely age and gender. For gender, we perform a binary classification with the classes male and female. For gender, we bucket the users into 7 buckets with age ranges: 0-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65+. We train a logistic regression directly on the user embeddings and use 5-fold cross-validation on 80% of the data, to then evaluate the performance of our classifier on the remaining 20% of test data. The results of the three different input features are shown in Table 2, alongside a random baseline. In both cases, the concatenated input features perform best, and content features seem to provide more information for gender, whereas visit features help more with age prediction.

4 Conclusion

We have extended our work on user embeddings from milestone 1 to take into account a visited website's content. Experiments have shown that the combination of visit and content features provides better results, both for intrinsic evaluation of the unsupervisedly learned embeddings as well as for downstream prediction tasks. All models have shown to perform robustly under different settings and tasks, which makes them more attractive to employ in a production setting. Since the methods applied in this investigation are general, further extension to other features and models is straightforward.

References

[1] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *Proc. ICML*, 2016.