Transformer Powered Two Level Matrix Factorization for Recommender Systems

Ricardo Plesz

August 18, 2021

In the era of Big Data it becomes increasingly difficult to find information that is relevant to oneself. Recommender systems are used to cope with this flood of information: they do this by providing personalized subsets of information to the user. Today, they are used in a wide variety of domains with great success. For instance, 80% of the streaming time on Netflix is influenced by the recommender system Netflix operates [1].

Online-argumentation is a domain where the user's opinion is influenced strongly by the specific arguments he faces. Problems such as filter bubbles or arguments that are not relevant to a specific user can be mitigated by using a recommender system which provides suitable arguments to the user, depending on the objective of the recommender system.
In the original paper, data from over 600 individuals on 900 arguments at different points in time was collected [2]. The goal was to provide a dataset that can be used to evaluate algorithms that predict how persuasive an argument is to a specific user.
Three tasks that use the known user-argument interaction data were presented to predict ratings of arguments by users that were collected at later points in time:

- Predicting a user's conviction by an argument (binary classification)

- Predicting the strength of the conviction for an argument (multiclass classification in the range $[0, 6]$)

- Predicting three convincing statements for a specific user

In this thesis I will focus on the first two tasks. In order to obtain reference performances for such an algorithm, two baseline algorithms were presented: a simple majority voter and a more sophisticated nearest-neighbor-algorithm. The goal of this thesis is to implement an algorithm that exceeds the performance of these two baseline algorithms on the provided dataset.

In the original paper the linguistic properties of the arguments were not exploited to make predictions [2]. The arguments were used to build an argumentation tree to eventually obtain numerical similarity values for the argumentation behaviour of users [3]. Hence, there were no content-based elements involved in the predictions.
In this bachelor thesis I will implement a Two-level Matrix Factorization (TLMF) algorithm to improve upon the presented baseline algorithms.[4] In the first step of the TLMF, semantic similarities of arguments are computed by applying a weighted textual matrix factorization (WTMF) that utilizes a term-document matrix. In the following step this information is integrated into another matrix factorization to calculate the user and item latent factors from which the original user-item matrix can be approximated. To extend the TLMF algorithm I will make use of the Augmented Sentence-BERT (AugSBERT) transformer architecture to incorporate in-domain knowledge of the arguments' topic into the semantic similarity prediction of argument pairs [5]. The AugSBERT-architecture applies both, a pre-trained, cross-encoder BERT model as well as a siamese, bi-encoder Sentence-Bert (Sbert)-model [6]. In a first step the BERT model will be pre-trained on different datasets and tasks to build a language model capable of applying word-level-knowledge as well as sentence-level-knowledge. To further strengthen the latter, I will apply a symmetric sentence prediction pre-training task where also Previous-Sentence-Prediction (PSP) is applied instead of only applying the usual Next-Sentence-Prediction (NSP) [7]. This pre-training task will tackle the problem of order-sensitivity of the BERT-architecture as well as forcing the model to learn deep contextual, semantic relations instead of relying on high-level semantic concepts such as topics to distinguish between different documents. Afterwards the pre-trained Bert model will soft-label the argument data, such that the siamese Sbert-architecture can be fine-tuned on this synthetic data to boost performance on predicting the similarity of arguments of the known domain. In this context,

new sampling algorithms will be evaluated to create a balanced, augmented and soft-labeled dataset to fine-tune the bi-encoder Sbert-model on.

References

[1] Carlos A Gomez-Uribe and Neil Hunt. "The netflix recommender system: Algorithms, business value, and innovation". In: *ACM Transactions on Management Information Systems (TMIS)* 6.4 (2015), pp. 1–19.

[2] Markus Brenneis, Maike Behrendt, and Stefan Harmeling. "How Will I Argue? A Dataset for Evaluating Recommender Systems for Argumentations". In: *22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue.* 2021.

[3] Markus Brenneis et al. "How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs." In: *SAFA@ COMMA.* 2020, pp. 2–13.

[4] Fangfang Li, Guandong Xu, and Longbing Cao. "Two-level matrix factorization for recommender systems". In: *Neural Computing and Applications* 27.8 (2016), pp. 2267–2278.

[5] Nandan Thakur et al. "Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks". In: *arXiv preprint arXiv:2010.08240* (2020).

[6] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019).

[7] Weidi Xu et al. "Symmetric Regularization based BERT for Pair-wise Semantic Reasoning". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2020, pp. 1901–1904.