INSTITUT FÜR INFORMATIK
Arbeitsgruppe Machine Learning

Universitätsstr. 1 D–40225 Düsseldorf

Heinrich Heine
Universität
Düsseldorf

Transformer Powered Two Level Matrix Factorization for Recommender Systems

Ricardo Plesz

Bachelorarbeit

Date of issue: 01. Oktober 2021
Date of submission: 01. Januar 2022
Reviewers: Prof. Dr. Stefan Harmeling
Prof. Dr. Stefan Conrad

Erklärung

Hiermit versichere ich, dass ich diese Bachelorarbeit selbstständig verfasst habe. Ich habe dazu keine anderen als die angegebenen Quellen und Hilfsmittel verwendet.

Düsseldorf, den 01. Januar 2022

____

Ricardo Plesz

Abstract

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 **gomez2015netflix**.

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 **HowIArgue**. 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 **HowIArgue**. The arguments were used to build an argumentation tree to eventually obtain numerical similarity values for the argumentation behaviour of users **brenneis2020much**. Hence, there were no content-based elements involved in the predictions.
In the first step of this bachelor thesis I will re-implement a Two-level Matrix Factorization (TLMF) algorithm to improve upon the presented baseline algorithms **li2016two**. In the first level 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 second 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.
In the second step of this bachelor thesis I will extend the TLMF algorithm by calculating the semantic similarities between arguments using a pre-trained BERT-model

devlin2018bert.
In the third step of this bachelor thesis I will implement and train a BERT-model from scratch. 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 novel symmetric sentence prediction pre-training task where also Previous-Sentence-Prediction (PSP) is applied instead of only applying the usual Next-Sentence-Prediction (NSP) **xu2020symmetric**. 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.
In the last step of this bachelor thesis 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 **thakur2020augmented**. The AugSBERT-architecture applies both, a pre-trained, cross-encoder BERT-model as well as a siamese, bi-encoder Sentence-Bert (Sbert)-model **reimers2019sentence**. 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.

Contents

1 De Bello Gallico

1.1 Gallia est omnis divisa

incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. Horum omnium fortissimi sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos animos pertinent important, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proeliis cum Germanis contendunt, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una, pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. Belgae ab extremis Galliae finibus oriuntur, pertinent ad inferiorem partem fluminis Rheni, spectant in septentrionem et orientem solem. Aquitania a Garumna flumine ad Pyrenaeos montes et eam partem Oceani quae est ad Hispaniam pertinet; spectat inter occasum solis et septentriones.

1.2 Apud Helvetios longe

Apud Helvetios longe nobilissimus fuit et ditissimus Orgetorix. Is M. Messala, [et P.] M. Pisone consulibus regni cupiditate inductus coniurationem nobilitatis fecit et civitati persuasit ut de finibus suis cum omnibus copiis exirent: perfacile esse, cum virtute omnibus praestarent, totius Galliae imperio potiri. Id hoc facilius iis persuasit, quod undique loci natura Helvetii continentur: una ex parte flumine Rheno latissimo atque altissimo, qui agrum Helvetium a Germanis dividit; altera ex parte monte Iura altissimo, qui est inter Sequanos et Helvetios; tertia lacu Lemanno et flumine Rhodano, qui provinciam nostram ab Helvetiis dividit. His rebus fiebat ut et minus late vagarentur et minus facile finitimis bellum inferre possent; qua ex parte homines bellandi cupidi magno dolore adficiebantur. Pro multitudine autem hominum et pro gloria belli atque fortitudinis angustos se fines habere arbitrabantur, qui in longitudinem milia passuum CCXL, in latitudinem CLXXX patebant.

His rebus adducti et auctoritate Orgetorigis permoti constituerunt ea quae ad proficiscendum pertinerent comparare, iumentorum et carrorum quam maximum numerum coemere, sementes quam maximas facere, ut in itinere copia frumenti suppeteret, cum proximis civitatibus pacem et amicitiam confirmare. Ad eas res conficiendas biennium sibi satis esse duxerunt; in tertium annum profectionem lege confirmant. Ad eas res conficiendas Orgetorix deligitur. Is sibi legationem ad civitates suscipit. In eo itinere persuadet Castico, Catamantaloedis filio, Sequano, cuius pater regnum in Sequanis multos annos obtinuerat et a senatu populi Romani amicus appellatus erat, ut regnum in civitate sua occuparet, quod pater ante habuerit; itemque Dumnorigi Haeduo, fratri Diviciaci, qui eo tempore principatum in civitate obtinebat ac maxime plebi acceptus erat, ut idem conaretur persuadet eique filiam suam in matrimonium dat. Perfacile factu esse illis probat conata perficere, propterea quod ipse suae civitatis imperium obtenturus esset: non esse dubium quin totius Galliae plurimum Helvetii possent; se suis copiis suoque exercitu illis regna conciliaturum confirmat. Hac oratione adducti inter se fidem et ius iurandum dant et regno occupato per tres potentissimos ac firmissimos populos totius Galliae sese potiri posse sperant.

Conrad (1997) hat ein Buch geschrieben. Es gibt auch andere Arbeiten (Patterson and Hennessy, 2004) die referenziert sind. In Abbildung 1 ist ein Sachverhalt dargestellt.

1 Autor: Conrad (1997) (Conrad, 1997)
2 Autoren: Liebeck and Conrad (2015) (Liebeck and Conrad, 2015)
3 Autoren: Liebeck et al. (2016) (Liebeck et al., 2016)

Online resource: ILSVRC2016 (2017)

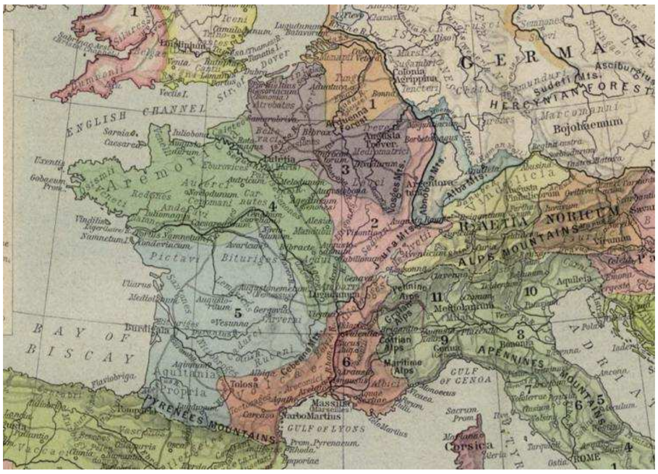

Figure 1: Gallien zur Zeit Caesars

Jahr	Erster Consul	Zweiter Consul
1	C. Caesar	L. Aemilius Paullus
2	P. Vinicius	P. Alfenus Varus
3	L. Aelius Lamia	M. Servilius
4	Sex. Aelius Catus	C. Sentius Saturninus
5	L. Valerius Messalla	Cn. Cornelius Cinna
suff.	C. Vibius Postumus	C. Ateius Capito
6	M. Aemilius Lepidus	L. Arruntius

Table 1: Römische Konsulen

2 De Bello Hispaniensi

Pharnace superato, Africa recepta, qui ex his proeliis cum adulescente Cn. Pompeio profugissent, cum . . . et ulterioris Hispaniae potitus esset, dum Caesar muneribus dandis in Italia detinetur, . . . quo facilius praesidia contra compararet, Pompeius in fidem uniuscuiusque civitatis confugere coepit. Ita partim precibus partim vi bene magna comparata manu provinciam vastare. Quibus in rebus non nullae civitates sua sponte auxilia mittebant, item non nullae portas contra cludebant. Ex quibus si qua oppida vi ceperat, cum aliquis ex ea civitate optime de Cn. Pompeio meritus civis esset, propter pecuniae magnitudinem alia qua ei inferebatur causa, ut eo de medio sublato ex eius pecunia latronum largitio fieret. Ita pacis commoda hoste +hortato+ maiores augebantur copiae. +Hoc crebris nuntiis in Italiam missis civitates contrariae Pompeio+ auxilia sibi depostulabant.

C. Caesar dictator tertio, designatus dictator quarto multis +iterante diebus coniectis+ cum celeri festinatione ad bellum conficiendum in Hispaniam cum venisset, legatique Cordubenses, qui a Cn. Pompelo discessissent, Caesari obviam venissent, a quibus nuntiabatur nocturno tempore oppidum Cordubam capi posse, quod nec opinantibus adversariis eius provinciae potitus esset, simulque quod tabellariis, qui a Cn. Pompeio dispositi omnibus locis essent, qui certiorem Cn. Pompeium de Caesaris adventu facerent . . . multa praeterea veri similia proponebant. Quibus rebus adductus quos legatos ante exercitui praefecerat Q. Pedium et Q. Fabium Maximum de suo adventu facit certiores, utque sibi equitatus qui ex provincia fuisset praesidio esset. Ad quos celerius quam ipsi opinati sunt appropinquavit neque, ut ipse voluit, equitatum sibi praesidio habuit.

Erat idem temporis Sex. Pompeius frater qui cum praesidio Cordubam tenebat, quod eius provinciae caput esse existimabatur; ipse autem Cn. Pompeius adulescens Uliam oppidum oppugnabat et fere iam aliquot mensibus ibi detinebatur. Quo ex oppido cognito Caesaris adventu legati clam praesidia Cn. Pompei Caesarem cum adissent, petere coeperunt uti sibi primo quoque tempore subsidium mitteret. Caesar - eam civitatem omni tempore optime de populo Romano meritam esse - celeriter sex cohortis secunda vigilia iubet proficisci, pari equites numero. Quibus praefecit hominem eius provinciae notum et non parum scientem, L. Vibiurn Paciaecum. Qui cum ad Cn. P praesidia venisset, incidit idem temporis ut tempestate adversa vehementique vento adflictaretur; aditusque vis tempestatis ita obscurabat ut vix proximum agnoscere possent. Cuius incommodum summam utilitatem ipsis praebebat. Ita cum ad eum locum venerunt, iubet binos equites conscendere, et recta per adversariorum praesidia ad oppidum contendunt. Mediisque eorum praesidiis cum essent, cum quaereretur qui essent unus ex nostris respondit, ut sileat verbum facere: nam id temporis conari ad murum accedere, ut oppidum capiant; et partim tempestate impediti vigiles non poterant diligentiam praestare, partim illo responso deterrebantur. Cum ad portam appropinquassent, signo dato ab oppidanis sunt reccepti, et pedites dispositi partim ibi remanserunt, equites clamore facto eruptionem in adversariorum castra fecerunt.

3 Weiteres Kapitel

3.1 Unterkapitel

3.2 Unterkapitel

A Anhang

Zusatzteil 1

Dies ist ein Anhang.

References

Stefan Conrad (1997). *Föderierte Datenbanksysteme: Konzepte der Datenintegration*. Springer Verlag.

ILSVRC2016 (2017). *ILSVRC 2016*. URL: http://image-net.org (visited on 07/17/2017).

Matthias Liebeck and Stefan Conrad (July 2015). "IWNLP: Inverse Wiktionary for Natural Language Processing". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 414–418.

Matthias Liebeck, Katharina Esau, and Stefan Conrad (Aug. 2016). "What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld". In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. Association for Computational Linguistics, pp. 144–153.

David A. Patterson and John L. Hennessy (2004). *Computer Organization and Design*. Morgan Kaufmann Publishers.

List of Figures

List of Tables