

基于 Seq2seq 模型的推荐应用研究

陈俊航 徐小平 杨恒泓

(广东技术师范大学电子与信息学院 广州 510000)

摘 要 日常生活的信息纷繁复杂,因此需要推荐系统来帮助人们进行信息筛选。传统的推荐系统将推荐过程看成是静态的,缺少对序列数据短期或长期的依赖关系的研究。循环神经网络由于在处理序列化数据时有突出的表现,因此可应用到具有序列特征的推荐数据中。文中采用循环神经网络的 seq2seq 模型来构造这种推荐系统,将推荐过程看作一个序列的翻译过程或答案生成的过程,利用大量用户以往的交互数据,找出其中的频繁模式,将其应用到其他用户对物品的行为预测中。实验在两个常用数据集上进行,使用 BLEU 衡量推荐结果,实验结果表明:该方法可以做出序列化的推荐。该方法只需要用户和物品的互动数据,摆脱了评分矩阵,避免了数据稀疏性的问题。

关键词 推荐系统,循环神经网络,Seq2seq 模型

中图法分类号 TP391.4 **文献标识码** A

Research on Recommendation Application Based on Seq2seq Model

CHEN Jun-hang XU Xiao-ping YANG Heng-hong

(School of Electronics and Information,Guangdong Polytechnic Normal University,Guangzhou 510000,China)

Abstract There is enormous information around us in daily basis which lead to the recommender systems to filter out the pure gold. The traditional recommender systems have been regarded as static,and lack of the research about the long or short term dependency of data. Considering the outstanding perform of recurrent neural network in tackling the sequence data,recommender system based on seq2seq model was built. The process of recommendation can be viewed as a process of sequence translation or a process of answer generation,and the model make uses of the used interactive sequence data to learn the inherent frequent patterns,then makes the prediction of other users' actions with items. Two datasets usually used for recommender system test are involved in the experiments,which measured by the BLEU. The results show that the method can make the sequence recommendation. The model only needs the interactive data between users and items,and gets rid of the rating matrix,thus avoids the sparsity problem.

Keywords Recommender system,Recurrent neural networkm,Seq2seq model

1 引言

推荐系统是基于用户兴趣并使用机器学习算法向用户推荐其喜欢的潜在相关事物的工具^[1]。文献[2]将推荐系统分为主要的 3 类:协同过滤、基于内容、混合推荐系统。其中,协同过滤算法^[3]是目前使用最为广泛的一种,许多更为复杂和准确的模型都是基于协同过滤算法的。协同过滤算法利用了用户和物品以往的交互信息,找出与目标用户 A 相似的用户 B,并将用户 B 感兴趣的物品推荐给用户 A。这是一种基于邻域的方法。还有部分推荐系统在找出相似用户之前对用户进行了聚类,显著地提高了推荐的准确率。

为了解决传统推荐系统存在的冷启动、数据稀疏性、准确率低等问题,基于机器学习的推荐模型得到了广泛关注,主要有基于贝叶斯的模型和基于矩阵分解的模型用于解决数据稀疏性和可扩展的问题,基于马尔科夫链和马尔科夫决策过程的推荐系统用于解决用户兴趣动态变化问题。深度学习模型在图像分类、语言识别和自然语言处理等方面有相当突出的

表现。而其中循环神经网络主要用于对序列数据进行建模并做出预测,这正好符合了我们想要捉住用户兴趣动态变化的需要。越来越多研究将深度学习与推荐算法结合,与传统的推荐算法相比,深度学习可以更好地挖掘出用户的需求物品的特征,以及两者之间的交互信息^[4]。协同过滤系统则利用其他用户与物品之间的交互信息做出推荐,我们提出的模型也基于这一原理。特别的是,利用循环神经网络将交互信息压缩成一个向量,再对该向量做进一步处理,可使代表用户历史信息的向量得到解释和利用。

本文首先介绍利用序列数据的推荐系统的相关工作;然后描述本文的数据集的基本信息及 seq2seq 推荐模型;接下来给出实验的过程和衡量标准,并分析实验结果;最后总结全文。

2 技术基础

为了提高推荐的效果,摆脱数据稀疏性和冷启动等问题,许多机器学习的方法被应用到了推荐系统中。例如,朴素贝

本文受广东省科技厅项目(粤科产学研字[2016]176 号,粤科规财字[2014]140 号)资助。
陈俊航(1991—),男,硕士,主要研究方向为机器学习、推荐系统,E-mail:chenjunhang2013@163.com;徐小平(1965—),教授,硕士生导师,CCF 高级会员,主要研究方向为网络信息安全、大数据与云计算,E-mail:cathy.xu@163.com;杨恒泓(1994—),女,硕士,主要研究方向为推荐系统。

叶斯协同过滤算法^[5]使用朴素贝叶斯分类器做预测。假设在给定类别(即用户对物品的评分)的情况下,特征(其他用户对该物品的评分)之间是相互独立的,因此在知道所有特征之后,某个用户对某物品的评分大小的概率是可计算的。朴素贝叶斯协同过滤算法可以被看作基于记忆的协同过滤算法。区别于使用相似度的推荐方法,该方法并没有利用用户的动态序列数据特征。此外,推荐过程还可以被看成是矩阵的分解过程^[6]。其将用户物品的互动关系(评分矩阵)分解成两个代表隐因子的低维矩阵,再将训练后的两个矩阵做内积,得到用户对物品评分的预测值。此推荐过程是静态的,相当于只考虑了长期依赖关系。但是一般的评分矩阵都是很稀疏的,因此难免了引入数据稀疏性的问题,使预测的准确率降低。而且并没有考虑到用户的动态数据,为了弥补这一不足,出现了 time-SVD++^[7]等模型,进一步将序列数据的性质应用到了推荐过程中。马尔科夫决策过程方法将推荐过程看作序列的最优化问题^[8]。MDP 的状态是物品,MDP 的动作是某件物品的推荐,激励是该件物品是否被购买。假设一个用户购买一件物品取决于他现在的状态以及该物品是否被推荐,但不取决于其他被推荐的物品。这样就可以预测出用户在购买某一物品之后的接下来几步的最可能的购买行为。MDP 利用了数据的序列性质,显著地提高了推荐系统的性能。

基于深度学习模型的推荐越来越受到工业界和学术界的关注。He 等^[9]提出了基于多层感知机的推荐系统,Sedhain 等^[10]提出了基于自编码器的推荐系统,Gong 等^[11]提出了基于卷积神经网络的推荐系统,Fischer 等^[12]提出了基于受限玻尔兹曼机的推荐系统,Wu 等^[13]提出了基于循环神经网络的推荐系统。

而且越来越多研究将 RNN 作为推荐系统。Musto 等^[14]提出了基于循环神经网络的推荐系统,通过从物品文本信息中学习到的特征,该系统能够产生 top-N 的物品给用户。Ko 等^[15]提出了基于用户隐式反馈数据的 RNN 推荐系统,可以捉住用户的动态信息。Wu 等^[13]提出了能够预测用户未来的行为轨迹的模型。Hidasi 等^[16]提出的基于 RNN 的推荐系统利用了用户的对话数据来做推荐。循环神经网络(RNN)是一种前向网络,在序列数据方面效果显著。假设序列的输入为 $(x_1, \dots, x_T)$,RNN 通过以下公式计算出输出序列 $(y_1, \dots, y_{T'})$ 。

$$h_t = \text{sigm}(W_{hx}x_t + W_{hh}h_{t-1})$$
$$y_t = W_{yh}h_t$$

RNN 突出的作用领域是 language modeling^[17],即预测序列的下一个输出。因此可以使用该模型来搭建推荐系统,在 language modeling 模型中将每个物品看作一个句子中的单词,使用 word embedding^[18]将整个物品集合中的各个物品之间联系起来。将该物品序列输入到 RNN 中,让 RNN 自动地提取序列中的特征,记录用户的购买习惯。然后预测用户接下来的购买行为^[19-20]。使用 language modeling 模型来进行推荐时一次只能预测一件物品,而不能产生多件物品,不能很好地体现用户接下来的购买序列的可能性。

3 基于 Seq2seq 的推荐模型

3.1 Seq2seq 模型

本文使用 RNN 的 Seq2seq 模型^[21],该模型在机器翻译

领域有卓越的效果,使用的是流行的编码器、解码器模式^[22],输入序列为 $(x_1, \dots, x_T)$,输出序列为 $(y_1, \dots, y_{T'})$ 。编码器可以很好地捉住短句子的特征,该特征被压缩到了 RNN 最后时刻的隐藏层输出 v ,解码器根据这个输出,模仿这两种语言的翻译特点,将一种语言翻译成另一种语言。为了提高 Seq2seq 的效果,现目前多数研究使用了 Attention 机制^[23-24],以提高结果的稳定性,让模型知道该在哪个句子的哪部分提取特征。也有研究者将 dropout 用在 RNN 上以避免网络训练过程中出现的过拟合现象^[25]。最后的预测是使用 softmax 层中最大的概率做分类^[26]。Seq2seq 模型在生成对话领域也有显著的表现^[27],这种模型的特性刚好符合我们推荐系统思想的需要:模型学习得出用户以往的购买特征,我们将该特征用于接下来的预测。模型最终的目标就如以下公式所示,在已知输入 x 的条件下,求出条件概率 $p(y_1, \dots, y_{T'} | x_1, \dots, x_T)$ 。

$$p(y_1, \dots, y_{T'} | x_1, \dots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \dots, y_{t-1})$$

3.2 基于 Seq2seq 模型的推荐算法

考虑将每个物品当作一个单词,所有的物品组成一个词汇表,每个用户的购买历史当作一个序列样本。随机将每个序列样本分成两部分,其中第一部分作为 encoder 的输入,第二部分作为 decoder 的输入和目标。RNN 的 encoder 对一个用户的购买物品序列进行编码,并将物品序列的第一部分输入到 RNN 的 decoder,每一步的输入是一个 embedding 层,用来对输出的物品进行 embedding。然后将 encoder 最后的隐藏层输出作为 decoder 的输入。同时将序列样本的第二部分逐个输入到 decoder 中,decoder 的输出是一个 softmax 层,用来对输出的结果进行分类,概率最大的分量为预测的单词(物品)。损失函数使用 negative log cross-entropy,训练过程就是最小化损失函数,提高了推荐结果的准确率。RNN 目前有 GRU 和 LSTM 两种主要的变体,由文献^[28]可知,GRU 在计算量上比 LSTM 要少,但性能差异不大,因此选择 GRU 作为我们模型的单元。图 1 表示了模型的训练过程。

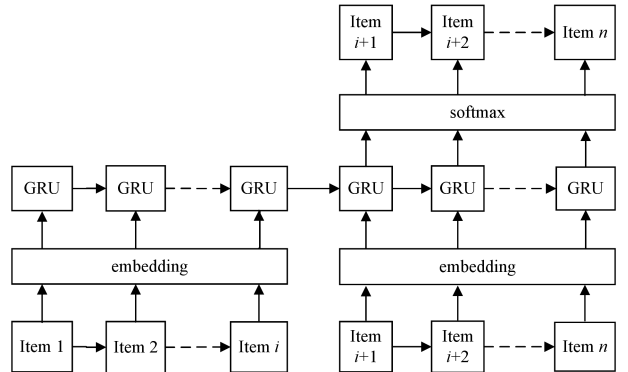


图 1 Seq2seq 推荐系统模型的训练过程

4 实验及结果分析

4.1 推荐数据集

MovieLens 10M 是一个专门用来测试推荐系统的数据集,这个数据集还包括了用户(性别、年龄、职业等)和电影(日期、题材等)等附属信息。Netflix 数据集是 Netflix 公司为了其举办的推荐系统大赛而发布的一个较大的数据集。这两个数据集包含了用户对电影评分的时间,体现了数据的序列性,适

合应用到我们的模型中。表 1 列出两个数据集的基本信息。

表 1 数据集信息			
数据集	用户数量/万个	电影数量/万部	评分数量/万次
Movielens 10M	7.15	1.06	100
Netflix	48	1.77	100

4.2 实验

4.2.1 训练过程

使用用户的观看记录来表示每个用户,将每个用户观看电影的序列随机地分成相应长度的序列对,每个用户在训练过程中生成大约 20~40 个序列对,随机选择序列可以增加实验数据的复杂度,提高模型的泛化能力。学习率采用 0.01,使用神经网络常用的训练方法(dropout,shuffle 等)来防止训练过程出现过拟合,使用随机梯度下降法来优化损失函数。

4.2.2 衡量标准

- (1)准确率:与序列作对比,预测准确的物品数除以总的预测数,代表此模型基本的性能。
- (2)BLEU:在机器翻译中,常使用 BLEU^[29] 作为翻译结果的评测指标,可以表明一个句子的词汇、短语顺序的准确性。由于这里使用的模型主要用于机器翻译,因此使用 BLEU 来衡量推荐的效果,即衡量推荐出来的物品顺序的准确程度。

4.3 实验结果分析

4.3.1 序列对大小

从图 2 中可以看出,当输入的序列长度比输出的序列长度短时,结果的 BLEU 和准确率都不高,只是因为当输入序列长度较短时,模型并不能从输入中提取到足够的信息,使得推荐结果不准确;相反地,当输入序列过长时,模型会因为提取能力不够而丢失信息。因此通过以上实验对比,为了计算方便,本文决定采用 10-10 的序列对来进行实验。

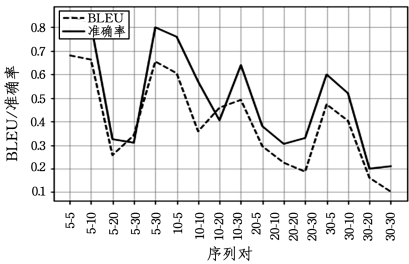


图 2 序列对大小

4.3.2 神经元数量

从图 3 中两个数据集的走势可以看出:随着神经元数量的增加,BLEU 递增。RNN 隐藏层神经元的数量的多少决定了整个模型的复杂度。从图中可以推出:在没有过拟合的情况下,更复杂的模型有更强的提取数据特征的能力,从而可以做出更好的预测。

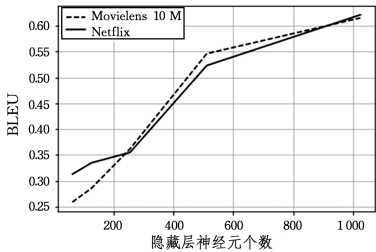


图 3 神经元数量对模型的影响

4.3.3 数据量

从图 4 中可以看出:随着数据量的增加,BLEU 下降,网络的复杂度不够支撑数据量的增加,因此目前只适用于小数据。

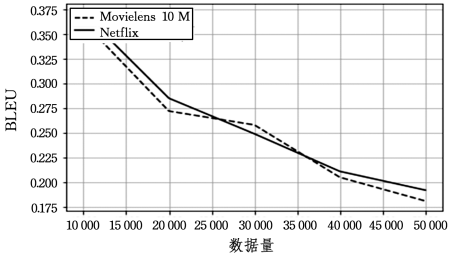


图 4 数据量与模型的影响

有时会出现有趣的推荐结果,系统会向你推荐一个系列的电影。例如:Police Academy 3:Back in Training (1986)<->Police Academy 4:Citizens on Patrol (1987)<->Police Academy 6:City Under Siege (1989)。这增加了推荐结果的多样性,而且也符合人们看电影会看一个系列的倾向。

4.3.4 准确率的比较

根据上述实验结果选取神经元个数 1000,序列对 10-10,以及数据量 15000 来定义 Seq2seq 推荐模型,并将实验结果与 MF,BPR,MDP,RNN 等模型的推荐准确率进行对比。从图 5 中可以看出:采用 Seq2seq 模型的推荐系统的准确率比其他几种算法要高。因为模型使用了用户的历史信息,相比于 MF 等静态方法,其准确率得到了提高;同时,同样利用用户历史信息的 MDP 和 RNN 等方法也取得了与 Seq2seq 相似的准确率。

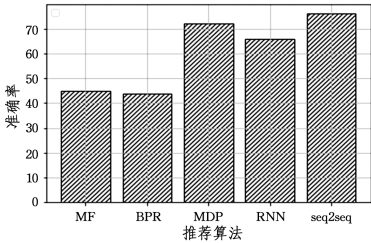


图 5 准确率对比

4.3.5 BLEU 的比较

Seq2seq 模型采用了上述实验参数,与 MDP,RNN 等序列化模型的 BLEU 结果进行对比,结果如图 6 所示。从图 6 可以看出,采用 Seq2seq 模型的推荐系统在两个数据集上的 BLEU 值更高,这说明 Seq2seq 能够更好地利用用户的历史信息,给出可能性更高的推荐序列。

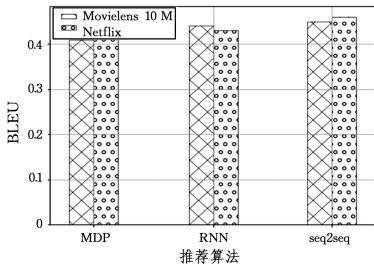


图 6 BLEU 对比

结束语 本文模型通过使用神经网络的自动提取特征功能,将数据中的序列化特征提取出来,通过对序列数据中潜在的用户物品关系的学习,将高维稀疏的特征空间映射到相对

稠密的低维特征空间,并做出一系列的推荐,序列数据的长期依赖关系导致了推荐结果的多样性。同时,由于只使用到数据的序列信息,摆脱了评分矩阵,因此避免了数据稀疏性的问题。但是就目前该模型还不适用于太大的数据量。下一步将增加模型的复杂度以增强模型处理大规模序列化数据的能力。

参 考 文 献

[1] PESKA L, VOJTAS P. Using implicit preference relations to improve recommender systems[J]. Journal on Data Semantics, 2017,6(1):15-30.

[2] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE transactions on knowledge and data engineering, 2005,17(6):734-749.

[3] SU X, KHOSHGOFTAAR T M. A survey of collaborative filtering techniques[J]. Advances in artificial intelligence, 2009, 2009:4.

[4] ZHANG S, YAO L, SUN A. Deep learning based recommender system: A survey and new perspectives[J]. arXiv preprint arXiv:1707.07435, 2017.

[5] MIYAHARA K, PAZZANI M J. Improvement of collaborative filtering with the simple Bayesian classifier[J]. Information Processing Society of Japan, 2002,43(11):679-689.

[6] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. IEEE Computer, 2009, 42(8):30-37.

[7] KOREN Y. Collaborative filtering with temporal dynamics[J]. Communications of the ACM, 2010,53(4):89-97.

[8] SHANI G, HECKERMAN D, BRAFMAN R I. An MDP-based recommender system[J]. Journal of Machine Learning Research, 2005,6(1):1265-1295.

[9] HE X, LIAO L, ZHANG H, et al. Neural collaborative filtering[C] // Proceedings of the 26th International Conference on World Wide Web. 2017:173-182.

[10] SEDHAIN S, MENON A K, SANNER S, et al. Autorec: Autoencoders meet collaborative filtering[C] // Proceedings of the 24th International Conference on World Wide Web. ACM, 2015: 111-112.

[11] GONG Y, ZHANG Q. Hashtag Recommendation Using Attention-Based Convolutional Neural Network[C] // International Joint Conference on Artificial Intelligence. 2016:2782-2788.

[12] FISCHER A, IGEL C. An introduction to restricted Boltzmann machines[C] // Iberoamerican Congress on Pattern Recognition. 2012:14-36.

[13] WU C Y, AHMED A, BEUTEL A, et al. Recurrent recommender networks[C] // Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017:495-503.

[14] MUSTO C, GRECO C, SUGLIA A, et al. Ask Me Any Rating: A Content-based Recommender System based on Recurrent Neural Networks[OL]. http://ceur-ws.org/vol-1653/paper_11.pdf.

[15] KO Y J, MAYSTRE L, GROSSGLAUSER M. Collaborative re-

current neural networks for dynamic recommender systems[C] // Asian Conference on Machine Learning. 2016:366-381.

[16] HIDASI B, KARATZOGLOU A, BALTRUNAS L, et al. Session-based recommendations with recurrent neural networks[J]. arXiv preprint arXiv:1511.06939, 2015.

[17] SUTSKEVER I, MARTENS J, HINTON G E. Generating text with recurrent neural networks[C] // Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011:1017-1024.

[18] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C] // Advances in neural information processing systems. 2013:3111-3119.

[19] DEVOOGHT R, BERSINI H. Collaborative filtering with recurrent neural networks[J]. arXiv preprint arXiv:1608.07400, 2016.

[20] DEVOOGHT R, BERSINI H. Long and Short-Term Recommendations with Recurrent Neural Networks[C] // Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization. ACM, 2017:13-21.

[21] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C] // Advances in neural information processing systems. 2014:3104-3112.

[22] CHO K, VAN B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[23] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[J]. arXiv preprint arXiv:1508.04025, 2015.

[24] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[25] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[J]. arXiv preprint arXiv:1409.2329, 2014.

[26] JEAN S, CHO K, MEMISEVIC R, et al. On using very large target vocabulary for neural machine translation[J]. arXiv preprint arXiv:1412.2007, 2014.

[27] Vinyals O, Le Q. A neural conversational model[J]. arXiv preprint arXiv:1506.05869, 2015.

[28] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C] // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002:311-318.

[29] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.

[30] LIU S J, YANG N, LI M, et al. A recursive recurrent neural network for statistical machine translation[C] // Meeting of the Association for Computational Linguistics. 2014:1491-1500.

[31] CHO K, BAHDANAU D, et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. arXiv preprint arXiv:1409.1259, 2014.