(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109241424 B
(45)授权公告日 2019.08.27

(21)申请号 201810999367.7

(22)申请日 2018.08.29

(65)同一申请的已公布的文献号
申请公布号 CN 109241424 A

(43)申请公布日 2019.01.18

(73)专利权人 陕西师范大学
地址 710000 陕西省西安市雁塔区长延堡办长安南路东侧

(72)发明人 王小明 庞光垚 郝飞 谢杰航 王新燕 林亚光 秦雪洋

(74)专利代理机构 北京中济纬天专利代理有限公司 11429
代理人 覃婧婵

(51)Int.Cl.
G06F 16/9535(2019.01)
G06N 3/04(2006.01)

(56)对比文件
CN 105868317 A,2016.08.17,
CN 107330115 A,2017.11.07,
CN 105577197 A,2016.05.11,
CN 107665254 A,2018.02.06,

审查员 张乾桢

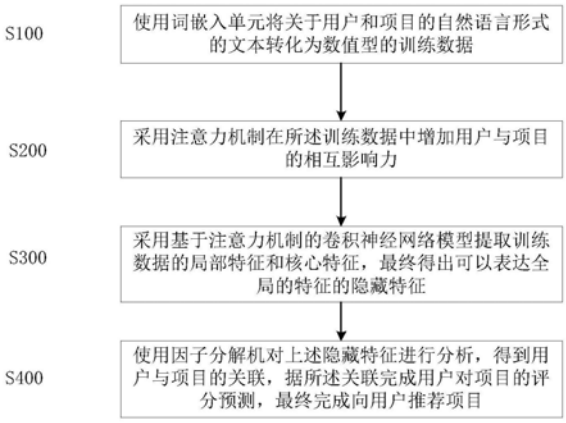
权利要求书3页 说明书10页 附图13页

(54)发明名称

一种推荐方法

(57)摘要

一种推荐方法，包括如下步骤：S100：使用词嵌入单元将关于用户和项目的自然语言形式的文本转化为数值型的训练数据；S200：采用注意力机制在所述训练数据中增加用户与项目的相互影响力；S300：采用基于注意力机制的卷积神经网络模型提取训练数据的局部特征和核心特征，最终得出可以表达全局的特征的隐藏特征；S400：使用因子分解机对上述隐藏特征进行分析，得到用户与项目的关联，据所述关联完成用户对项目的评分预测，最终完成向用户推荐项目。该方法相较于现有方法，提高了推荐的精度和准确度，提高了数据利用率。



CN 109241424 B

1.一种推荐方法,包括如下步骤:

S100:使用词嵌入单元将关于用户和项目的自然语言形式的文本转化为数值型的训练数据;

S200:采用注意力机制在所述训练数据中增加用户与项目的相互影响力;

S300:采用基于注意力机制的卷积神经网络模型提取训练数据的局部特征和核心特征,最终得出可以表达全局的特征的隐藏特征;

S400:使用因子分解机对上述隐藏特征进行分析,得到用户与项目的关联,据所述关联完成用户对项目的评分预测,最终完成向用户推荐项目;

步骤S100进一步包括,

所述词嵌入单元包括规格化和数值化两个步骤,其中,规格化是指对自然语言形式的文本进行分词、去除停用词和无用词,数值化是指使用多维分布向量对文本进行数值化操作;

采用评论信息作为训练数据,则该方法具体为:

S101:使用词嵌入单元将自然语言形式的用户评论信息和项目评论信息数值化为用户的表达特征向量和项目的表达特征向量;

S201:通过计算用户评论信息与项目评论信息之间的相似度,得到表达用户评论与项目评论之间影响度的注意力矩阵;将注意力矩阵与用户的权值矩阵进行运算,得到用户的注意力特征向量,将注意力矩阵与项目的权值矩阵进行运算,得到项目的注意力特征向量;把用户的表达特征向量与用户的注意力特征向量进行拼接形成带记忆能力的新的用户的表达特征向量;把项目的表达特征向量与项目的注意力特征向量进行拼接形成带记忆能力的新的项目的表达特征向量;

S301:使用基于注意力机制的卷积神经网络对上述新的用户表达特征向量和新的项目表达特征向量进行卷积、池化和全连接操作,从中提取用户隐藏特征以及项目隐藏特征;

S401:使用因子分解机从用户隐藏特征及项目隐藏特征中构建出用户及项目之间的关联,并且根据所述关联完成用户对项目的评分预测,最终完成向用户推荐项目;

所述步骤S201进一步包括:

S2001:采用欧几里得距离公式来计算注意力矩阵A,$A\in R^{d\times d}$,其中,$R^{d\times d}$为维度为d×d的实数,d为每一个训练批次的数据维度;

S2002:假设用户的注意力特征向量为$F_{0,a}\in R^{d\times n}$,项目的注意力特征向量为$F_{1,a}\in R^{d\times n}$,用户的注意力权重为$W_0\in R^{d\times n}$以及项目的注意力权重为$W_1\in R^{d\times n}$,在训练初期随机初始化W_0以及W_1,再在后续的迭代训练中,根据预测结果与训练数据结果的误差情况,反向微调W_0以及W_1,在多次迭代之后得到W_0以及W_1的最优值;最优值与注意力矩阵A可以快速构建出最佳的$F_{0,a}$和最佳的$F_{1,a}$,详细的计算过程如下所示:$(F_{i,a})_{d\times n}=(W_i)_{d\times n}\cdot(A)_{d\times d}$,$i\in\{0,1\}$,其中$R^{d\times n}$为维度为d×n的实数,d为每一个训练批次的数据维度,n为句子长度;

S2003:已知用户的表达特征向量$F_{0,r}$和项目的表达特征向量$F_{1,r}$,用户的注意力特征向量$F_{0,a}$和项目的注意力特征向量$F_{1,a}$,维度均为d×n,通过公式$F_{i,nr}=\mathrm{concat}(F_{i,r},F_{i,a})$,$i\in\{0,1\}$,直接拼接构建出融合用户注意力的新的表达特征向量$F_{0,nr}$以及融合项目注意力的新的表达特征向量$F_{1,nr}$。

2.根据权利要求1的方法,步骤S300进一步包括:

所述卷积神经网络包括卷积、池化以及全连接操作。

3.根据权利要求1的方法，其中，

用户的权值矩阵以及项目的权值矩阵初期是直接随机初始化，并利用深度学习的后向传播方法进行更新。

4.根据权利要求1的方法，步骤S101进一步包括：

S1001，已知c_m^u为用户u对项目m的评论，假设有$c_m^u = \{s_{m,1}^u, s_{m,2}^u, \ldots, s_{m,i}^u, \ldots, s_{m,n}^u\}$，其中$s_{m,i}^u$表示用户u对项目m评论的第i个句子；再假设句子$s_{m,i}^u = \{w_{i1}, w_{i2}, \ldots, w_{ij}, \ldots, w_{in}\}$，其中，$w_{ij}$表示第i个句子的第j个单词，n为每个句子的单词数；为了建立单词与数值的对应关系，建立映射函数Φ(w_{ij})：w_{ij}→Z，Z∈N^*，该函数表示从单词w_{ij}到数值Z的映射关系，其中N^*为正整数集合；在此基础上，构建出以数值表达的用户u评论的多维分布向量V^u：

$$V^u = \begin{bmatrix} \phi(w_{11}) & \phi(w_{12}) & \ldots & \phi(w_{1j}) & \ldots & \phi(w_{1n}) \\ \phi(w_{21}) & \phi(w_{22}) & \ldots & \phi(w_{2j}) & \ldots & \phi(w_{2n}) \\ \ldots & & & & & \\ \phi(w_{i1}) & \phi(w_{i2}) & \ldots & \phi(w_{ij}) & \ldots & \phi(w_{in}) \\ \ldots & & & & & \\ \phi(w_{q1}) & \phi(w_{q2}) & \ldots & \phi(w_{qj}) & \ldots & \phi(w_{qn}) \end{bmatrix},$$

其中，V^u中的任一个元素V_{ij}^u表示词嵌入单元对用户u评论中单词w_{ij}进行处理之后的数值化结果，q表示用户u所有评论的句子数量；同理构建出以数值表达的项目m评论的多维分布向量V_m，且$V_m = V^u$，以及同理得到V_m中任一个元素$V_{m,kq}$为项目m评论第k个句子第q个单词所对应的数值化值；

S1002，假设词嵌入单元表示某个训练批次数据的表达特征向量为$F_{i,r} \in R^{d \times n}$，i∈{0，1}，其中i等于0为用户的表达特征向量，i等于1为项目的表达特征向量，n为句子长度，d为每一个训练批次的数据维度，$R^{d \times n}$为维度为d×n的实数；如果用户评论的句子数量q大于批次的数据维度d，则需要划分多个批次，如果q小于批次的数据维度d，则需要增加(d-q)行值为零的数据；据此可得用户的表达特征向量和项目的表达特征向量

$$F_{0,r} = \begin{cases} \left(V_{1:}^u \quad V_{2:}^u \quad \ldots \quad V_{k:}^u \quad \ldots \quad V_q^u\right)^T, q \geq d, \\ \left(V_{1:}^u \quad V_{2:}^u \quad \ldots \quad V_{k:}^u \quad \ldots \quad V_q^u \quad \ldots \quad 0\right)^T, q < d, \end{cases}$$

$$F_{1,r}=\begin{cases}\left(V_{m,1:}\ V_{m,2:}\ \cdots\ V_{m,k:}\ \cdots\ V_{m,q:}\right)^{T}, q\geq d,\\ \left(V_{m,1:}\ V_{m,2:}\ \cdots\ V_{m,k:}\ \cdots\ V_{m,q:}\ \cdots\ 0\right)^{T}, q<d\end{cases}。$$

5.根据权利要求1的方法，步骤S301进一步包括：

S3001，以用户的新的表达特征向量与项目的新的表达特征向量$F_{0,nr}$和$F_{1,nr}$作为输入进行卷积操作，包括特征提取以及特征选择；假设卷积核为$x_j\in R^{d\times n}$，低纬度特征值为$C=[c_1,c_2,\ldots,c_j,\ldots,c_n]$，$C\in R^n$，其中，$c_j=f(F_{0,nr}*x_j+b_j)$，运算符*表示卷积操作，$b_j$为偏置变量，$b_j\in R$，f采用ReLU激活函数，其中$R^{d\times n}$为维度为d×n的实数，其中$R^n$为维度为n的实数，R为实数；接着从通过卷积核$x_j$扫描到的数据中聚合出核心特征：$f(c_j)=\max\{0,c_j\}$；

S3002，使用ReLU激活函数，在各个局部数据当中选取最大值，从一个卷积核中提取核心特征$o_j=\max\{c_1,c_2,\ldots,c_i,\ldots,c_{n-t+1}\}$，然后以并行的方式使用多个卷积核对数据进行特征提取，得到总特征$O=\{o_1,o_2,\ldots,o_j,\ldots,o_{n1}\}$，其中$n_1$为卷积核的数量，t为卷积核的步长；

S3003，在全连接操作中对总特征集合O重新组合，采用公式$H_u=f(W\cdot O+g)$计算得到可以表达全局的特征的用户隐藏特征Hu，其中g为卷积神经网路中全连接层的偏置变量，w为全连接层的权重矩阵；同理，从项目的新的表达特征向量$F_{1,nr}$中提取到项目的隐藏特征H_m。

6.根据权利要求1的方法，其中，

使用两个并列的基于注意力机制的卷积神经网络，同步从评论信息中提取用户隐藏特征以及项目隐藏特征。

一种推荐方法

技术领域

[0001] 本公开属于人工神经网络及个性化推荐技术领域，特别涉及一种推荐方法。

背景技术

[0002] 随着云计算、大数据、物联网等技术的快速发展，互联网和信息行业涌现了大量的诸如购物、教育和娱乐等应用平台，使得多源异构数据的规模也急速增长，预计到2020年全球数据总量将达到35.2ZB。这些大数据蕴含着丰富的价值，能够指导人们将行为决策模式从经验主义为主转变为数据驱动为主。然而，人们在享受大数据带来便利的同时，难以从大数据中提取有价值的信息，由此引发了“信息过载”的问题。因此，如何从大数据中根据用户的需求和兴趣挖掘出有效信息是至关重要的。

[0003] 推荐方法是解决互联网等平台中信息过载问题的有效解决方案，包括基于协同过滤的推荐方法、基于内容的推荐方法和混合推荐方法。此外，近年来随着深度学习成为互联网大数据和人工智能的研究热点，涌现了一类基于深度学习的新型混合推荐方法。虽然基于深度学习的混合推荐方法能够自动提取特征，但特征提取的精度仍需进一步提升。为了提高神经网络特征提取的精度，当前多使用注意力机制对神经网络进行拓展，其中神经网络主要包括卷积神经网络CNN和循环神经网络RNN。卷积神经网络的汇聚操作会遗失一些词汇的位置信息，也无法考虑权重高的历史词汇的影响度，降低了其在自然语言处理中提取特征的精度；循环神经网络虽能考虑动态信息并在自然语言的特征提取有较好效果，但是与CNN相比，其对静态数据的特征表达效果较差且运算速度过慢。

[0004] 综上所述，在大数据环境下进行推荐仍然面临三个挑战：一是如何提高从异构多源的数据中提取特征的精度；二是如何把传统推荐方法中特征提取的方式由人工提取转向自动提取；三是如何将深度学习与传统推荐方法融合成一个可高度协作的混合模型。

发明内容

[0005] 为了解决上述问题，本公开提供了一种推荐方法，包括如下步骤：

[0006] S100：使用词嵌入单元将关于用户和项目的自然语言形式的文本转化为数值型的训练数据；

[0007] S200：采用注意力机制在所述训练数据中增加用户与项目的相互影响力；

[0008] S300：采用基于注意力机制的卷积神经网络模型提取训练数据的局部特征和核心特征，最终得出可以表达全局的特征的隐藏特征；

[0009] S400：使用因子分解机对上述隐藏特征进行分析，得到用户与项目的关联，据所述关联完成用户对项目的评分预测，最终完成向用户推荐项目。

[0010] 上述技术方案提出了一种基于自然语言处理的注意力机制。该机制在保持卷积神经网络具有较好特征提取性能的同时，对卷积神经网络进行了拓展，增加了多源异构的文本中每个词和目标特征的影响度。该机制解决了卷积神经网络在自然语言处理当中丢失记忆的问题，使得卷积神经网络在自然语言处理环境中的特征提取精度得到了明显提升。

[0011] 本方法引入基于注意力机制的卷积神经网络模型。该模型具有深层次和非线性网络结构的特点，可以将多源异构数据映射到一个相同的隐空间，例如从基于自然语言形式的评论信息中自动提取用户和项目的隐藏特征。自动提取特征的方式解决了基于内容的推荐方法过度依赖人工提取特征的问题，使得推荐领域的特征提取的方式能够适应大数据的环境，拓展了推荐系统的适用范围。

[0012] 本方法提出一种融合了基于注意力机制的卷积神经网络模型和因子分解机的混合推荐方法。该方法首先自动提取隐藏特征，其次通过考虑多个隐藏特征的组合得到更多有效的辅助信息，最后利用这些有效的辅助信息建立了关联关系，完成了较高精度的推荐。该方法改进了传统推荐模型存在的数据稀疏、冷启动难、推荐性能差、可解释性差和适用性差等缺陷，实质性提升了推荐的效率以及推荐精确度。

附图说明

[0013] 图1是本公开一个实施例中所提供的一种基于注意力机制的卷积神经网络和因子分解机的混合推荐方法的流程图；

[0014] 图2是本公开一个实施例中所提供的一种基于注意力机制的卷积神经网络和因子分解机的混合推荐方法的框架图；

[0015] 图3是本公开一个实施例中注意力机制原理图；

[0016] 图4是本公开一个实施例中数据集详细统计信息；

[0017] 图5是本公开一个实施例中不同的卷积核数量下的RMSE指标；

[0018] 图6是本公开一个实施例中不同的隐因子数量下的RMSE指标；

[0019] 图7(a)是本公开一个实施例中全部数据集的RMSE指标对比结果图；

[0020] 图7(b)是本公开一个实施例中全部数据集的MAE指标对比结果图；

[0021] 图8(a)是本公开一个实施例中冷启动用户数据集的RMSE指标对比结果图；

[0022] 图8(b)是本公开一个实施例中冷启动用户数据集的MAE指标对比结果图；

[0023] 图9(a)是本公开一个实施例中长尾项目数据集的RMSE指标对比结果图；

[0024] 图9(b)是本公开一个实施例中长尾项目数据集的MAE指标对比结果图；

[0025] 图10(a)是本公开一个实施例中新用户的RMSE指标对比结果图；

[0026] 图10(b)是本公开一个实施例中新用户的MAE指标对比结果图；

[0027] 图11(a)至图11(d)是本公开一个实施例中在不同字符串长度下的评论数量分布图(使用Log-Log Scaling)：图11(a)Books数据集；图11(b)Movies and TV数据集；图11(c)Home and Kitchen数据集；图11(d)Tools and Home Improvement数据集；

[0028] 图12是本公开一个实施例中评论信息的字符串长度分析图；

[0029] 图13(a)是本公开一个实施例中不同字符串长度用户组的Books数据集对比结果图；

[0030] 图13(b)是本公开一个实施例中不同字符串长度用户组的Movies and TV数据集对比结果图；

[0031] 图13(c)是本公开一个实施例中不同字符串长度用户组的Home and Kitchen数据集对比结果图；

[0032] 图13(d)是本公开一个实施例中不同字符串长度用户组的Tools and Home

Improvement数据集对比结果图；

[0033] 图14是本公开一个实施例中性能对比结果图。

具体实施方式

[0034] 参看图1，在一个实施例中，公开了提供了一种推荐方法，包括如下步骤：

[0035] S100：使用词嵌入单元将关于用户和项目的自然语言形式的文本转化为数值型的训练数据；

[0036] S200：采用注意力机制在所述训练数据中增加用户与项目的相互影响力；

[0037] S300：采用基于注意力机制的卷积神经网络模型提取训练数据的局部特征和核心特征，最终得出可以表达全局的特征的隐藏特征；

[0038] S400：使用因子分解机对上述隐藏特征进行分析，得到用户与项目的关联，据所述关联完成用户对项目的评分预测，最终完成向用户推荐项目。

[0039] 该方法提出拓展卷积神经网络的注意力机制，在保证CNN在静态特征提取优点的同时，参考RNN的运行机理，增强CNN在自然语言处理中的记忆能力；其次使用基于注意力的卷积神经网络进行特征提取，提升数据利用率；最后研究基于注意力的卷积神经网络与因子分解机的融合完成高精度推荐。

[0040] 在另一个实施例中，采用评论信息作为训练数据，基于注意力机制的卷积神经网络和因子分解机的混合推荐方法(Attention-based Convolutional Neural Network with Factorization Machines，ACNN-FM)的总体框架如图2所示，主要包括以下几个功能模块：

[0041] (1)词嵌入模型：利用多维的分布向量(n-dimensional distributed vectors)，将基于自然语言的评论信息转化为CNN模型能够识别的数值型数据。

[0042] (2)注意力机制：我们提出的注意力机制，增加了用户评论以及项目评论之间的关联度，能够改进CNN模型在NLP领域缺乏记忆的缺陷。

[0043] (3)CNN模型：卷积神经网络能够利用评论文本提取用户和项目的隐藏特征，主要操作包括卷积、池化和全连接。

[0044] (4)因子分解机模型：主要利用用户和项目的隐藏特征构建用户与项目的关联，完成推荐。

[0045] 其中，用户对项目的评价信息用四元组P＝(u,m,c,r)来表示，u为用户，m为项目，c为用户u对项目m的评论文本，r为用户u对项目m的评分。同一个用户对不同项目的评论用集合$c^u=\{c_1^u,c_2^u,...c_m^u,...d_k^u\}$来表示，不同用户对同一个项目的评论用集合$c_m=\{c_m^1,c_m^2,...c_m^u,...c_m^N\}$来表示。其中，$c_m^u$表示用户u对项目m的评论，k为用户u评论过的项目数量，N为评论过项目m的用户数量。用$\hat{y}\leftarrow R(H_u,H_m)$来表示根据用户的隐藏特征和项目的隐藏特征构建关联，并预测出用户u对项目m的评分。其中，Hu表示用户u从c^u中学习到的隐藏特征，Hm表示项目m从c_m中学习到的隐藏特征；使用二元组R＝(H_u,H_m)表达用户u与项目m的关联；$\hat{y}$表示根据R完成用户u对项目m的预测评分。

[0046] 在另一个实施例中，所述向用户推荐项目包括向信息的使用者推荐商品、知识、电

影和音乐。

[0047] 在另一个实施例中，构建词嵌入模型。深度学习方法只能处理数值型数据，无法直接处理自然语言等文本形式的数据，因此，我们改进了词嵌入模型，以数值化自然语言。词嵌入包括规格化和数值化两个步骤，即先对自然语言进行分词、去除停用词和无用词等规格化操作，再使用多维分布向量对评论信息进行数值化操作。

[0048] 首先，已知c_m^u为用户u对项目m的评论，假设有$c_m^u=\{s_{m,1}^u,s_{m,2}^u,\ldots,s_{m,i}^u,\ldots,s_{m,n}^u\}$，其中$s_{m,i}^u$表示用户u对项目m评论的第i个句子；再假设句子$s_{m,i}^u=\{w_{i1},w_{i2},\ldots,w_{ij},\ldots,w_{in}\}$，其中，$W_{ij}$表示第i个句子的第j个单词，n为每个句子的单词数；为了建立单词与数值的对应关系，建立映射函数Φ(w_{ij})：$w_{ij}\rightarrow Z$，$Z\in N^+$，该函数表示从单词W_{ij}到数值Z的映射关系；在此基础上，构建出以数值表达的用户u评论的多维分布向量V^u：

[0049] $$V^u=\begin{bmatrix}\phi(w_{11}) & \phi(w_{12}) & \ldots & \phi(w_{1j}) & \ldots & \phi(w_{1n})\\ \phi(w_{21}) & \phi(w_{22}) & \ldots & \phi(w_{2j}) & \ldots & \phi(w_{2n})\\ \ldots & & & & & \\ \phi(w_{i1}) & \phi(w_{i2}) & \ldots & \phi(w_{ij}) & \ldots & \phi(w_{in})\\ \ldots & & & & & \\ \phi(w_{q1}) & \phi(w_{q2}) & \ldots & \phi(w_{qj}) & \ldots & \phi(w_{qn})\end{bmatrix}, \quad (1)$$

[0050] 其中，V^u中的任一个元素V_{ij}^u表示词嵌入单元对用户u评论中单词W_{ij}进行处理之后的数值化结果，q表示用户u所有评论的句子数量；同理构建出以数值表达的项目m评论的多维分布向量V_m，且$V_m=V^u$，以及同理得到V_m中任一个元素$V_{m,kq}$为项目m评论单词W_{ij}(第i个句子第j个单词)对应的数值化值(第k个句子第q个单词所对应的数值化值)；

[0051] 其次，假设词嵌入单元表示某个训练批次数据的表达特征向量为$F_{i,r}\in R^{d\times n}$，$i\in\{0,1\}$，其中i等于0为用户的表达特征向量，i等于1为项目的表达特征向量，n为前面已定义的句子长度，d为每一个训练批次的数据维度；如果用户评论的句子数量q大于批次d，则需要划分多个批次，如果q小于批次d，则需要增加(d-q)行值为零的数据；即有用户特征表达向量$F_{0,r}$：

[0052] $$F_{0,r}=\begin{cases}\begin{pmatrix}V_{1:}^u & V_{2:}^u & \ldots & V_{k:}^u & \ldots & V_q^u\end{pmatrix}^T, q\geq d,\\ \begin{pmatrix}V_{1:}^u & V_{2:}^u & \ldots & V_{k:}^u & \ldots & V_q^u & \ldots & 0\end{pmatrix}^T, q<d,\end{cases} \quad (2)$$

[0053] 与公式(1)和(2)同理，可得项目特征表达向量$F_{1,r}$：

[0054] $$F_{1,r}=\begin{cases}\left(V_{m,1:}\ V_{m,2:}\ \cdots\ V_{m,k:}\ \cdots\ V_{m,q:}\right)^{T}, q\geq d,\\ \left(V_{m,1:}\ V_{m,2:}\ \cdots\ V_{m,k:}\ \cdots\ V_{m,q:}\ \cdots\ 0\right)^{T}, q<d.\end{cases} \quad (3)$$

[0055] 至此，经过上述步骤的处理，构建了使用多维分布向量数值化评论信息的词嵌入模型，最终得到用户的表达特征向量$F_{0,r}$和项目的表达特征向量$F_{1,r}$。

[0056] 在另一个实施例中，构建注意力机制模型。提出了一种注意力机制对CNN进行了拓展，该机制主要在训练数据中增加用户和项目评论信息之间的影响程度，以提升CNN模型对历史词汇的记忆能力。

[0057] 具体来说，注意力机制在上文定义的用户和项目的表达特征向量$F_{0,r}$以及$F_{1,r}$的基础上，假设有注意力矩阵$A\in R^{d\times d}$表示集合c^{u}的句子与集合c_{m}的句子之间单词的相互影响程度。如图3为迭代训练过程中某一次注意力机制的运算细节，从中可以看出注意力机制包括以下三个阶段：

[0058] 阶段一：采用欧几里得距离公式来计算注意力矩阵A，$A\in R^{d\times d}$，对任一元素$a_{kj}\in A$，$a_{kj}=\left\|v_{k:}^{u}-v_{m,j:}\right\|_{2}^{2}$，其中$\|\cdot\|_{2}^{2}$为欧式距离范式，$v_{k:}^{u}$为用户u的词向量，k：指k行的所有值，$v_{m,j}$：为项目m的词向量，j：指第j行的所有值，$R^{d\times d}$为维度为d×d的实数；

[0059] 阶段二：我们假设用户注意力特征向量为$F_{0,a}\in R^{d\times n}$，项目注意力特征向量为$F_{1,a}\in R^{d\times n}$，均表示从评论信息挖掘到的用户与项目之间的关联关系，其中$R^{d\times n}$为维度为d×n的实数，d为批次，n为句子长度(包含的单词数)。我们利用机器学习相关理论中的权重矩阵来构建该关联关系，即先假设用户的注意力权重为$W_{0}\in R^{d\times n}$以及项目的注意力权重为$W_{1}\in R^{d\times n}$，接着在训练初期随机初始化以W0以及W1，再在后续的迭代训练中，根据预测结果与训练数据结果的误差情况，反向微调W0以及W1，在多次迭代之后得到W0以及W1的最优值。该最优值与注意力矩阵A可以快速构建出最佳的$F_{0,a}$和最佳的$F_{1,a}$，详细的计算过程如下所示：

[0060] $$(F_{i,a})_{d\times n}=(W_{i})_{d\times n}\cdot(A)_{d\times d},\ i\in\{0,1\}, \quad (5)$$

[0061] 阶段三：构造融合了用户与项目之间相互影响度的新表达特征向量。已知用户的表达特征向量$F_{0,r}$(或者项目的表现特征向量$F_{1,r}$)与用户的注意力特征向量$F_{0,a}$(或者项目的注意力特征向量$F_{1,a}$)维度均为d×n，通过直接拼接构建出融合用户注意力的新表达特征向量$F_{0,nr}$以及融合项目注意力的新表达特征向量$F_{1,nr}$：

[0062] $$F_{i,nr}=\mathrm{concat}(F_{i,r},F_{i,a}),\ i\in\{0,1\}. \quad (6)$$

[0063] 总体来说，构建注意力机制模型，主要有两个核心策略，一是先利用评论信息构建用户与项目的相互影响力，即注意力特征向量F0,a和F1,a。二是构建新的用户与项目相互影响力的表达特征向量$F_{0,nr}$和$F_{0,nr}$。

[0064] 在另一个实施例中，构建卷积神经网络模型。卷积神经网络主要通过卷积层和池化层来提取训练数据的局部特征和核心特征。卷积神经网络包括卷积、池化以及全连接操作，以下分别描述：

[0065] 首先，CNN以用户与项目新表达特征向量$F_{0,nr}$和$F_{1,nr}$作为输入进行卷积操作，包括特征提取以及特征选择。特征提取指利用卷积核以固定步长扫描输入数据，提取代表局部数据的低纬度特征值。假设卷积核为$x_{j}\in R^{d\times n}$，低纬度特征值为$C=[c_{1},c_{2},\ldots,c_{j},\ldots,c_{n}]$，

$C \in R^n$,则有以下提取过程:

[0066] $c_j = f(F_{0,nr} * x_j + b_j)$,(7)

[0067] 其中,运算符*表示卷积操作,b_j为偏置变量($b_j \in R$),f为激活函数。本文的f采用ReLU激活函数,该函数与其他激活函数相比,运算量更少和执行速度更快。接着从通过卷积核x_j扫描到的数据中聚合出核心特征:

[0068] $f(c_j) = maX\{0, c_j\}$,(8)

[0069] 其次,在池化层对卷积操作之后得到的特征向量图进行下采样操作,选取局部中具有代表意义的特征信息。本文使用ReLU激活函数,在各个局部数据当中选取最大值,组合成表达重要特征信息:

[0070] $o_j = maX\{c_1, c_2, \ldots, c_i, \ldots, c_{n-t+1}\}$,(9)

[0071] 公式(9)描述了从一个卷积核中提取核心特征的过程,为了加快运算速度,我们以并行的方式使用多个卷积核对数据进行特征提取,得到总特征$O \in R$的值为:

[0072] $o = \{o_1, o_2, \ldots, o_j, \ldots, o_{nl}\}$ (10)

[0073] 其中n_1为卷积核的数量,t为卷积核的步长。

[0074] 最后,我们在全连接操作中对总特征集合O重新组合,得到可以表达全局的特征的用户特征H_u:

[0075] $H_u = f(W \cdot O + g)$. (11)

[0076] 整体来说,CNN模型能够从用户的新表达特征向量$F_{0,nr}$中提取到表达用户的隐藏特征H_u;其中g为卷积神经网路模型中全连接层的偏置变量,w为全连接层的权重矩阵。同理,与之并列运行和具有相同结构的CNN模型从项目的新表达特征向量$F_{1,nr}$中提取到项目的隐藏特征H_m。

[0077] 在另一个实施例中,构建因子分解机模型。因子分解机模型(Factorization Machines,FM)是一种由线性回归(linear regression)和奇异值分解(Singular value decomposition,SVD)扩展而来的通用预测器,通过分析由特征工程得到的实值特征向量的内在联系,预测出与用户关联性很高的项目。

[0078] 已知在ACNN-FM方法当中主要有两个部分:一是使用两个并列的基于注意力机制的卷积神经网络,同步从评论信息中提取用户隐藏特征H_u和项目隐藏特征H_m;二是使用因子分解机对H_u和H_m进行分析,得到用户与项目的关联模型$\hat{x} = (H_u, H_m)$,并根据$\hat{x}$预测出用户对项目的评分$\hat{y}$,具体计算过程如下:

[0079] $$\hat{y} = w_0 + \sum_{i=0}^{n} w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j \qquad (12)$$

[0080] 其中,$w_0 \in R$,$w = \{w_1, w_2, \cdots, w_n\}$,$w \in R^n$,$v \in R^{n \times k}$为模型参数,n为用户R为实数,$R^n$为n维实数,$R^{nxk}$为nxk维实数,k属于正整数,是一个超参数,和评论的特征维度,$n << p$表示因子分解机的维度,v_i表示矩阵v的向量($v \in R^{n,k}$),w_0为全局偏量,W_i为第i个变量的权重。$<V_i, V_j>$表示H_u和H_m向量之间的点积,表示如下:

[0081] $$< v_i, v_j >= \sum_{f=1}^{k} v_{i,f} \bullet v_{j,f} \qquad (13)$$

[0082] 其中，$V_{i,f}$和$V_{j,f}$是点积$<V_i, V_j>$拆分的2个矩阵，是为了利用更多的数据来解决数据稀疏问题。

[0083] 经过公式(12)以及(13)的运算，利用用户的隐藏特征H_u以及项目的隐藏特征H_m构建出了用户与项目的关联程度，最终完成用户对项目的评分预测。

[0084] 下面实施例结合附图4至图14进行阐述。

[0085] 以下所用的4组数据均来自于亚马逊不同行业产生的评价信息，其中，在4组数据中用户对项目的评论信息真实地反映了用户对项目的喜好程度。Books数据集是用户关于书本的评价信息，Movies&TV数据集是用户关于电影以及电视节目的评价信息，Home & Kitchen数据集是用户关于居家用品以及厨房用品的评价信息，Tools & Home Improvement数据集是用户关于工具以及家装用品的评价信息。

[0086] 4组数据集的相关统计信息如图4所示，不仅包括不同的数据量级，还包括不同的领域以及不同的稀疏程度，整体而言数据集的选取能够全面地和客观地模拟信息过载场景。其中，数据集的统计参数中，有用户数量为N_{user}，有项目数量为N_{item}，评价数量为N_{review}，数据的稀疏程度为$d=1-(N_{review}/(N_{user} \bullet N_{item}))$，单个用户的平均评论数为$\overline{N_{user}}$，有单个项目的平均评论数$\overline{N_{item}}$，有总单词数为$N_{word}$。

[0087] 在推荐模型的训练与测试的过程中，采用能够较好解决过拟合问题的Holdout验证方法，该方法将每一个数据集随机拆分为训练集(training set)、验证集(validation set)和测试集(test set)三个部分，且三个部分所占总数量的比例为80%、10%和10%。其中，使用训练集(training set)训练各个推荐模型的参数，使用验证集(validation set)评估经训练后得到的推荐模型参数的性能，最终在测试集(test set)上评估模型的泛化误差。

[0088] 在ACNN-FM方法的评测过程中，需要针对模型的特点选取合适的评价指标、具有代表意义的对比算法以及经过反复试验确定各种对比模型影响较大的运行参数，以下分别描述：

[0089] 首先，结合模型的特征提取精度以及推荐效果这两个核心关注点，采用2种常用的评价指标：平均绝对误差(mean absolute error，MAE)以及均方根误差(root mean squared error，RMSE)。两种指标通过计算真实评分与预测评分间的误差来衡量推荐结果的准确性，其值越小，推荐精度越高。

[0090] $$MAE = \frac{1}{N}\sum_{i}\left|y_i - \hat{y}_i\right| \qquad (14)$$

[0091] $$RMSE = \sqrt{\frac{1}{N}\sum_{i}\left(y_i - \hat{y}_i\right)^2} \qquad (15)$$

[0092] 其中,N表示测试集评分的总记录数,y_i表示真实评分值,$\hat{y}_i$表示预测评分值。

[0093] 其次,选取了以下三种对比方法:

[0094] (1)非负矩阵分解(Non-negative Matrix Factorization,NMF)是一种所有矩阵元素均符合非负约束的矩阵分解方法,其参照人类思维中“局部构成整体”的概念,将原始非均匀的复杂数据矩阵拆分为两个简化的非负子矩阵,再以简单的迭代方法分别求解拆解后的子矩阵。求解方法具有收敛速度快、左右非负矩阵存储空间小的特点,适合处理较大规模数据。例如对“用户-项目”评分矩阵进行分解,得到一个用户隐向量矩阵和一个项目隐向量矩阵来构建隐语义的模型,最终得到更好的推荐效果。

[0095] (2)SVD++是引入隐式反馈,使用用户的历史浏览数据、历史评分数据、电影的历史浏览数据、电影的历史评分数据等作为新的参数改进SVD的一种算法。可以根据已有的评分情况,分析出评分者对各个因子的喜好程度以及电影包含各个因子的程度,再反向分析数据得出预测结果。

[0096] (3)DeepCoNN是一种先进的基于深度学习的混合推荐方法。该方法主要分为两个部分,第一部分通过深度学习的方法利用评分以外的信息构建用户和项目的特征,第二部分利用基于内容的推荐模型融合特征完成推荐,是当前效果非常好的一种深度学习与传统推荐模型结合的混合推荐方法。

[0097] 最后,通过反复测试确定各个算法的最优超参数。NMF推荐方法在隐因子个数(number of latent factors)为{25,50,100,150,200}、正则化参数为{0.001,0.01,0.1,1.0}以及学习率为{0.006,0.005,0.004,0.003,0.002,0.001}的范围中可以取得最优推荐效果。SVD++模型方法使用Topic K的方式进行验证,当隐因子的个数(number of latent factors)为{25,50,100,150,200}的范围中可以得到最优值。图5为ACNN-FM方法与DeepCoNN方法的卷积核数量取值为{10,50,100,150,200,400}的RMSE值,其中,DeepCoNN方法在卷积核数量为20的时候效果最佳,ACNN-FM方法在卷积核数量为50的时候效果最佳;图6所示是ACNN-FM方法与DeepCoNN方法的隐因子个数取值为{10,20,40,60,80,100}的RMSE值,其中两种方法均在隐因子个数为30时,取得最佳推荐效果。此外,在ACNN-FM方法以及DeepCoNN方法在迭代次数(epochs)取值为40的训练过程中,模型在测试集(test set)上的损失函数(Loss)均趋于平稳。

[0098] 推荐多样性评测

[0099] 考虑到推荐系统的应用场景存在差异性,从三个不同的角度进行评测:全部数据集、冷启动用户数据集和长尾项目数据集。其中,全部数据集是指待评测的数据为全部数据集;冷启动用户数据集是指待评测的数据是评价数量为1至5的用户数据集;长尾项目数据集是指带评测的数据为所有库存项目的数据集(按使用次数倒序,冷门项目的数量约占总数的80%)。具体结果如下所示:

[0100] 图7(a)至图7(b)为全部数据集的对比结果。结果显示ACNN-FM方法以及DeepCoNN方法的RMSE值均远低于NMF方法以及SVD++方法的RMSE值,ACNN-FM方法与DeepCoNN方法相比,在4个不同数据集的评测结果中RMSE值平均提升了12%以及MAE值平均提升了8%。此外,ACNN-FM方法在数据集Books和Movies and TV具有较高稀疏度(其稀疏度分别为99.996%和99.973%)的环境下,误差率也仍然小于DeepCoNN方法。说明ACNN-FM方法在大

数据环境下,能够在自然语言形式的评论中自动提取有效特征,提高数据利用率,最终取得了最好的推荐效果。

[0101] 图8(a)至图8(b)为冷启动用户数据集的对比结果。4个数据集Books、Movies&TV、Home&Kitchen和Tools&Home Improvement在训练集中的平均冷启动用户数为140680、31759、5768和1519,分别占总用户数的23.3%,25.6%,34.7%和27.4%。结果表明,随着用户评价数据的减少,所有方法的推荐效果均出现下滑,其中NMF方法以及SVD++方法的推荐精确度下滑严重,ACNN-FM方以及DeepCoNN方法的推荐精度轻微下滑。而ACNN-FM方法的误差率小于DeepCoNN方法的误差率。由此可得,基于深度学习的混合推荐方法解决冷启动问题的能力优于传统推荐方法,ACNN-FM方法由于具有更高的特征提取精度,与DeepCoNN方法相比能够更好缓解冷启动问题。

[0102] 图9(a)至图9(b)为长尾项目数据集的对比结果。长尾现象是幂律分布的通俗提法,表现为在电商平台中销量占主导的产品只占少数,而多数产品则被人遗忘。与冷启动不同,长尾项目可能存在评价数量较多但是销售量却很低的情况。将销售最少的70%的项目设为长尾项目的数据占比(低于长尾分布的80%,更能体现算法的性能),最后表明:与其他推荐方法相比,ACNN-FM方法在Books、Movies&TV、Home & Kitchen和Tools &Home Improvement数据集上取得了最好的推荐性能。说明ACNN-FM方法在具备较高数据利用率的情况下,在长尾现象组上取得了最好的推荐效果。

[0103] 新用户推荐效果评测

[0104] 能否为新用户推荐准确的项目是衡量推荐系统冷启动处理能力的重要性能指标之一。假设用户在初次进入系统的时候选择了某几项系统推荐的大众热门项目,针对该场景,使用项目的标题和描述作为用户对项目的评价文本,从而构造出了与元组P类似的数据集。

[0105] 图10(a)至图(b)为新用户推荐结果,结果表明ACNN-FM方法的RMSE值比DeepCoNN方法提升了7.5%,比SVD++方法提升了51%,比NMF方法提升了60.5%。评价文本由于使用标题以及描述进行构造,所以所蕴含的隐藏特征比较丰富,更有利于建立用户与项目之间的关联关系。即随着ACNN-FM方法从自然语言形式评论文本中提取有效特征的精确度的提升,相比其他3种方法,有更高的数据利用率,在解决数据稀疏问题以及处理新用户的冷启动问题中具有更高的效率。

[0106] 字符长度对推荐效果的影响

[0107] ACNN-FM方法的主要特点是通过在训练过程当中增加用户评论文本与项目评论文本之间的关联以提升卷积神经网络特征提取的精度,由此可预测,评论信息的字符串长度对该方法的推荐效果有影响。为了分析该特点,如图11(a)、图11(b)、图11(c)和图11(d)所示,针对4个不同领域的数据集,分析了不同字符串长度下的评论数量的分布,其中,使用对数对坐标值进行了处理。从图中可以看出不同字符串长度下的评论数量的分布符合长尾分布,即大部分评论信息的字符长度集中在某一个区间。

[0108] 为了进一步分析评论信息的字符长度的分布区间,针对每个字符长度下的评论数量进行统计。其中,每个字符长度下的评论数量进行按倒序的方式排序,从高至低取一定量数据的比率为数据占比,具体字符长度分布情况如图12所示。

[0109] 由图12可知,大部分评论信息的字符串长度介于58-885之间,在该区间评测不同

算法的效果最具有说服力，此外，为了尽量保证每个分组有相同的数据量，使用评论信息的字符长度作为刻度将用户划分为5组：0-500、501-1000、1001-1500、1501-2000、2001-3500以及＞3500。图13(a)至图13(d)展示了4种推荐方法在4个数据集下的评测结果，即传统的NMF方法与SVD++方法无论处于哪一个用户组，预测结果基本一致，其结果不受字符串长度的影响；ACNN-FM方法优于DeepCoNN方法，尤其在0～1000的区间ACNN-FM方法显著优于DeepCoNN方法，说明ACNN-FM方法在增加了历史词汇之间的影响度之后取得了最好的推荐效果，并且推荐的精度随着评论信息长度的增加而增加。

[0110] 时间成本评测

[0111] 在实际应用当中，时间成本是衡量推荐方法可用性的重要指标。针对模型训练时间以及执行时间两个指标进行对比。如图14所示，随着数据量的增加，传统推荐算法的执行时间会越来越长，远超2秒(注：用户能承受的互联网响应时间不超2秒，最佳响应时间为1秒内)；基于机器学习的混合推荐方法模型训练时间远高于传统推荐方法，但是执行时间始终保持在1秒以内。与DeepCoNN方法相比，虽然ACNN-FM方法增加注意力机制导致训练时间变长，但是执行时间基本保持不变。相对于最终推荐效果的提升度，时间损耗在可接受范围之内。

[0112] 尽管以上结合附图对本发明的实施方案进行了描述，但本发明并不局限于上述的具体实施方案和应用领域，上述的具体实施方案仅仅是示意性的、指导性的，而不是限制性的。本领域的普通技术人员在本说明书的启示下和在不脱离本发明权利要求所保护的范围的情况下，还可以做出很多种的形式，这些均属于本发明保护之列。

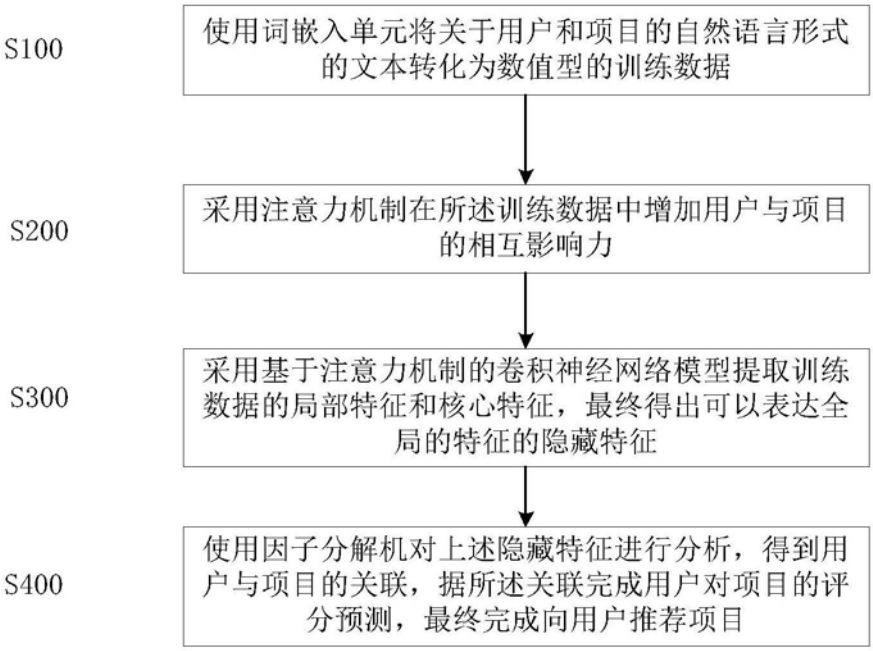

S100
使用词嵌入单元将关于用户和项目的自然语言形式的文本转化为数值型的训练数据
S200
采用注意力机制在所述训练数据中增加用户与项目的相互影响力
S300
采用基于注意力机制的卷积神经网络模型提取训练数据的局部特征和核心特征，最终得出可以表达全局的特征的隐藏特征
S400
使用因子分解机对上述隐藏特征进行分析，得到用户与项目的关联，据所述关联完成用户对项目的评分预测，最终完成向用户推荐项目

图1

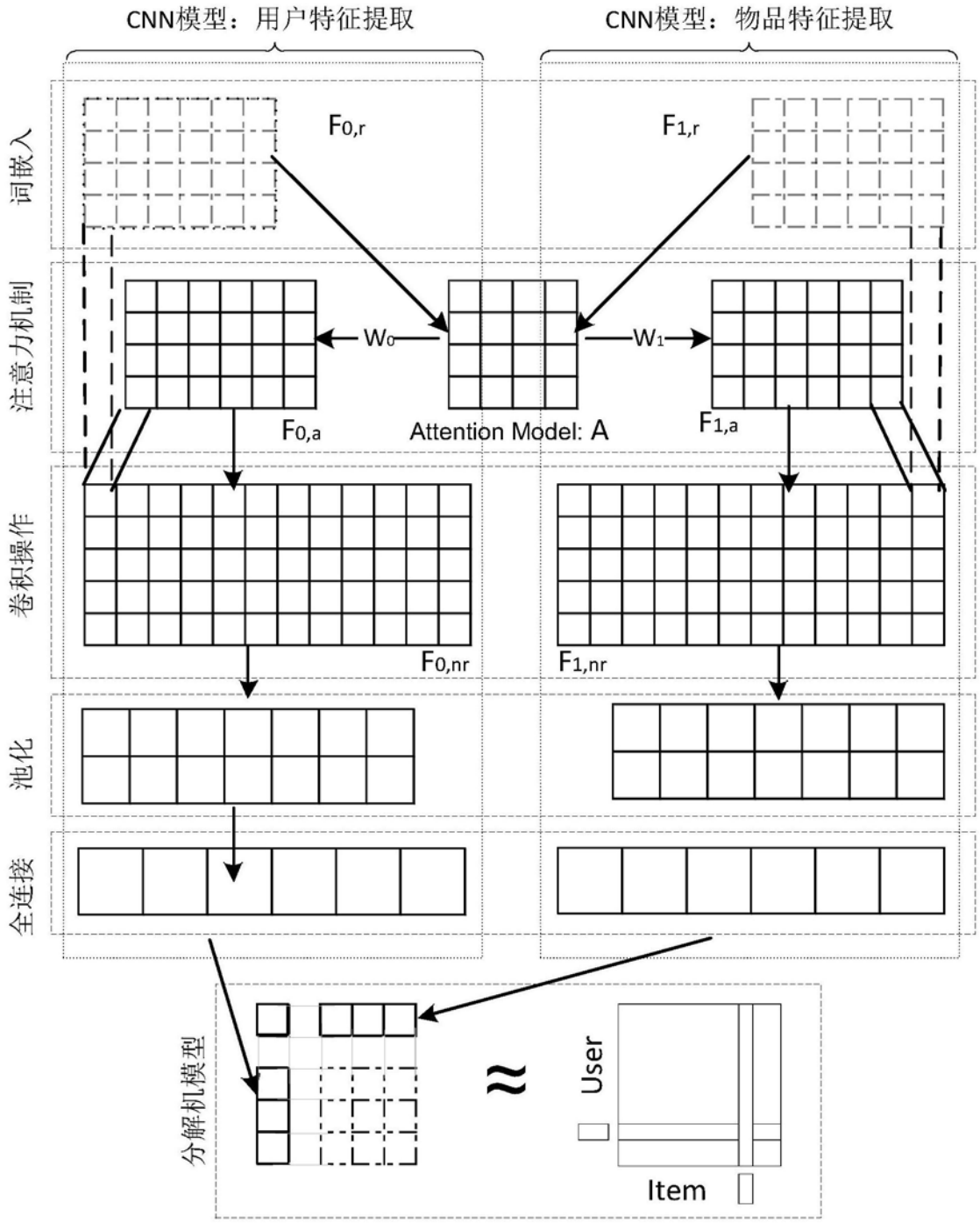

CNN模型：用户特征提取
CNN模型：物品特征提取
词嵌入
注意力机制
卷积操作
池化
全连接
分解机模型
F0,r
F1,r
W0
W1
F0,a
Attention Model: A
F1,a
F0,nr
F1,nr
≈
User
Item

图2

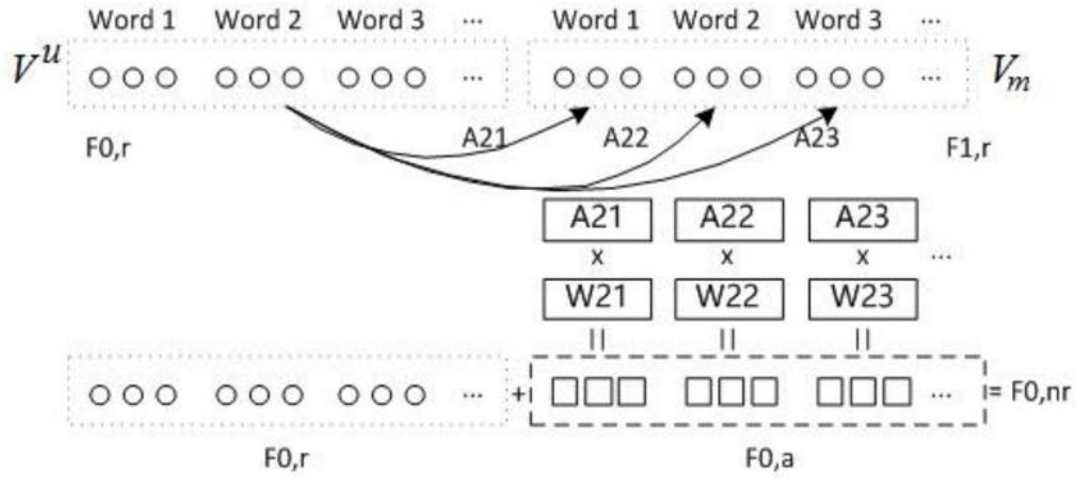


图3

数据集	Books	Movies & TV	Home& Kitchen	Tools & Home Improvement
Nuser	603668	123960	66519	16638
Nitem	367982	50052	28237	10217
Nreview	8898041	1697533	551682	134476
d (%)	99.996	99.973	99.971	99.921
$\overline{N_{user}}$	15	14	8	8
$\overline{N_{item}}$	24	34	20	13
Nword (KB)	9,236,338	1,938,890	411,435	109,742

图4

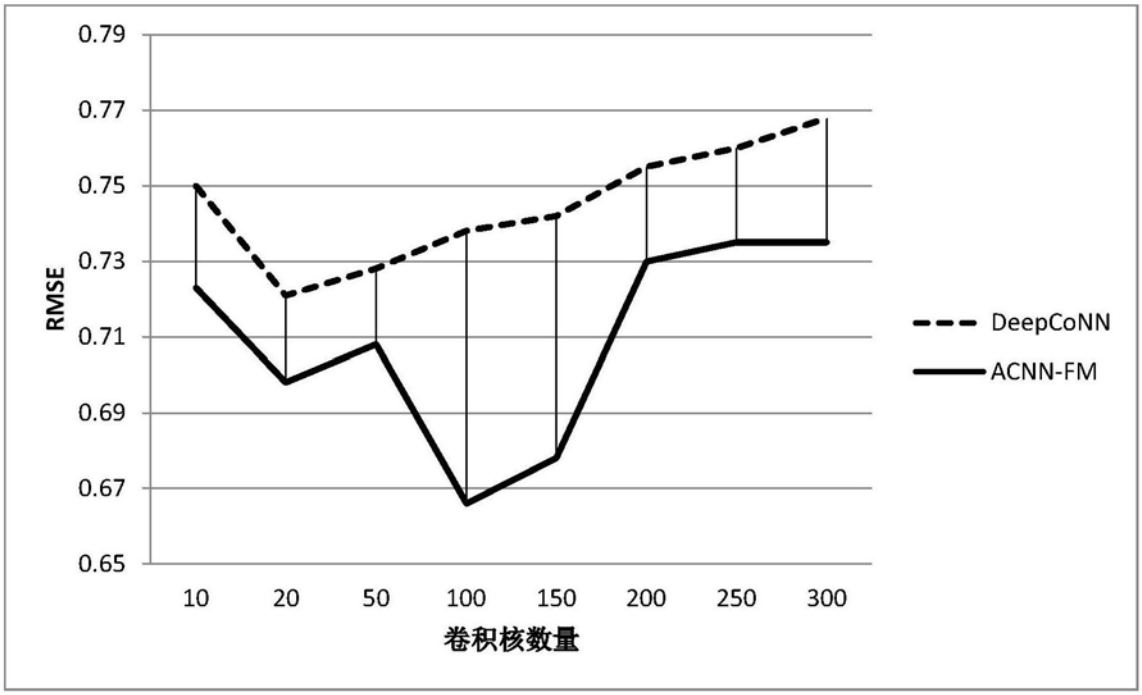

0.79
0.77
0.75
0.73
0.71
0.69
0.67
0.65
RMSE
10
20
50
100
150
200
250
300
卷积核数量
DeepCoNN
ACNN-FM

图5

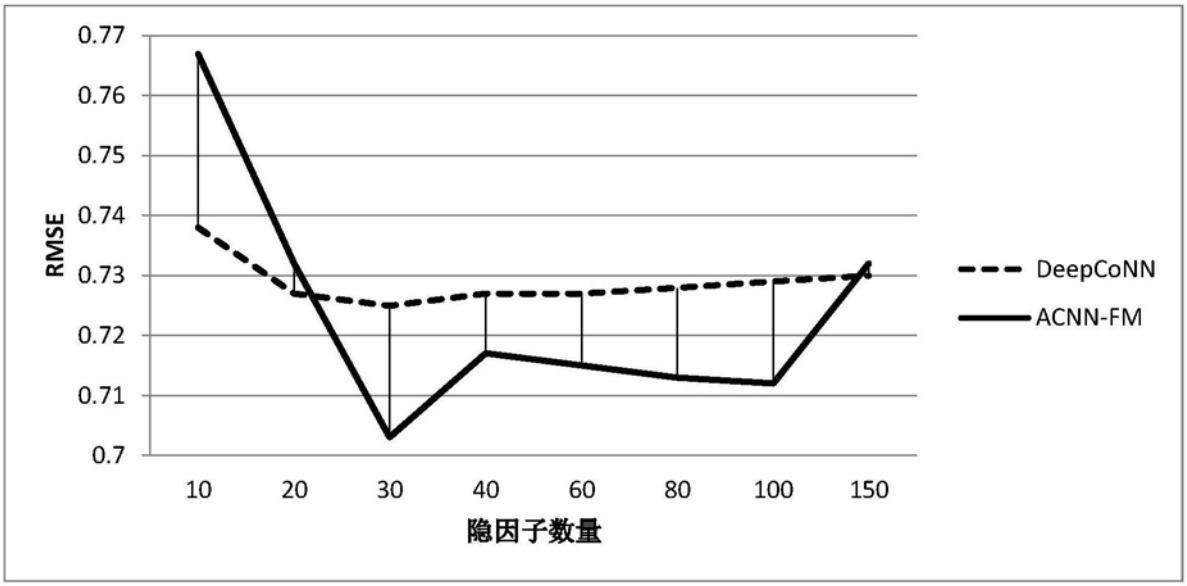

0.77
0.76
0.75
0.74
0.73
0.72
0.71
0.7
RMSE
10
20
30
40
60
80
100
150
隐因子数量
DeepCoNN
ACNN-FM

图6

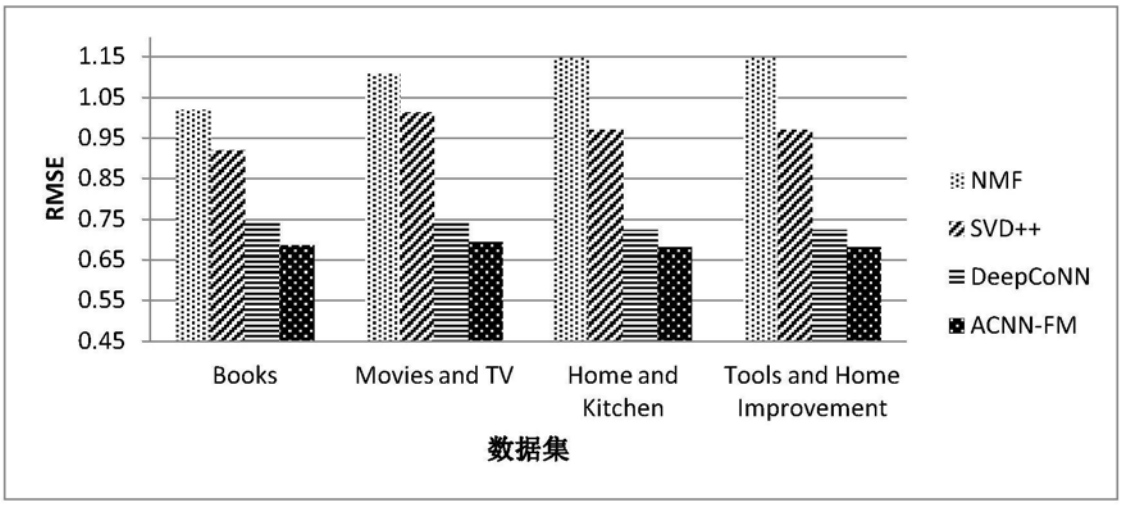

1.15
1.05
0.95
0.85
0.75
0.65
0.55
0.45
RMSE
Books
Movies and TV
Home and Kitchen
Tools and Home Improvement
数据集
NMF
SVD++
DeepCoNN
ACNN-FM

图7(a)

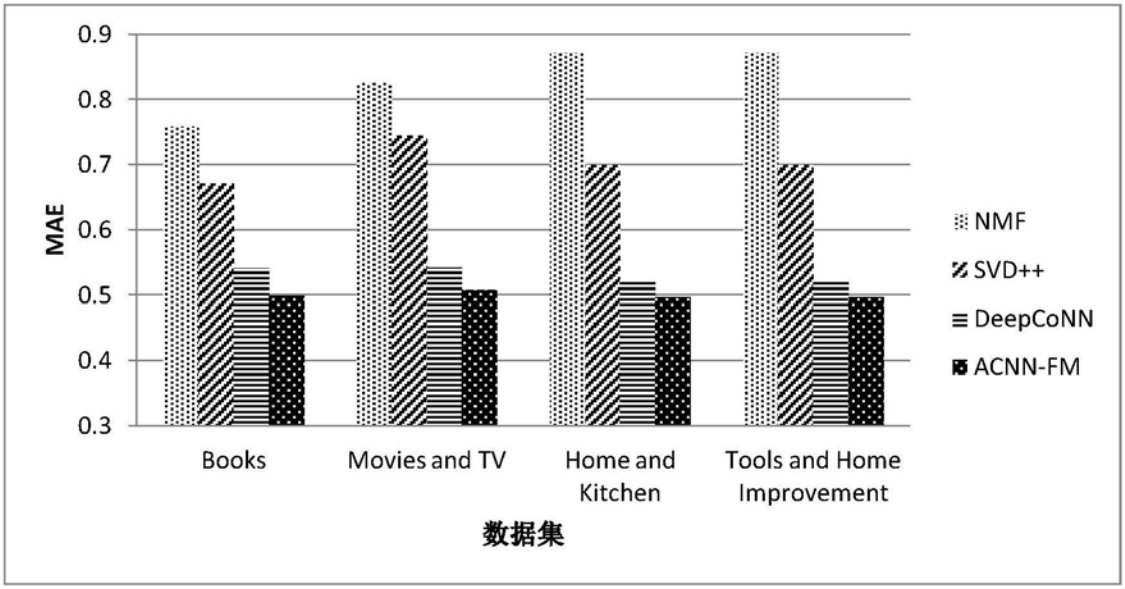

0.9
0.8
0.7
0.6
0.5
0.4
0.3
MAE
Books
Movies and TV
Home and Kitchen
Tools and Home Improvement
数据集
NMF
SVD++
DeepCoNN
ACNN-FM

图7(b)

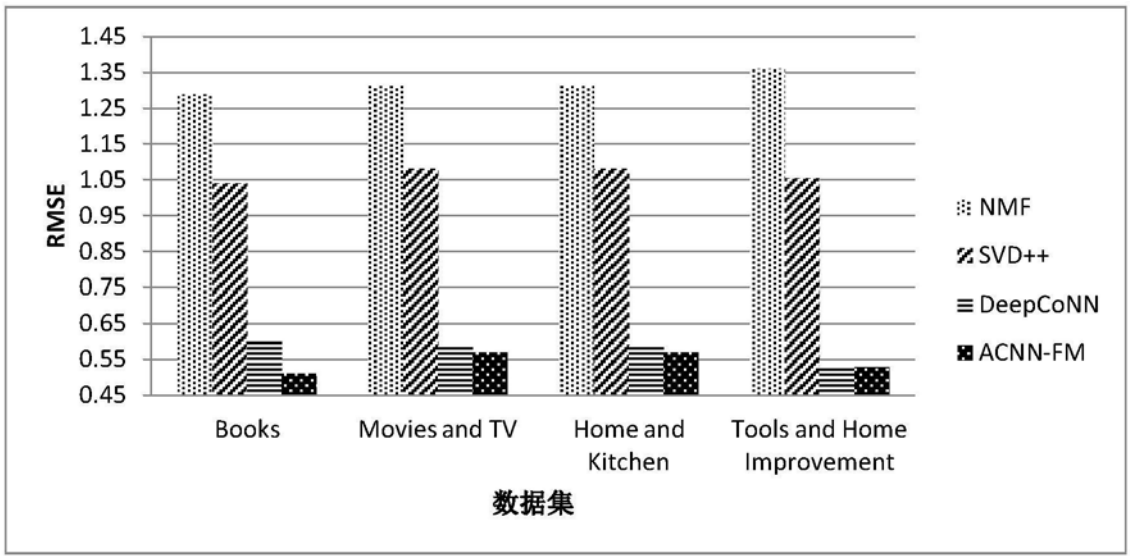

1.45
1.35
1.25
1.15
1.05
0.95
0.85
0.75
0.65
0.55
0.45
RMSE
Books
Movies and TV
Home and Kitchen
Tools and Home Improvement
数据集
NMF
SVD++
DeepCoNN
ACNN-FM

图8(a)

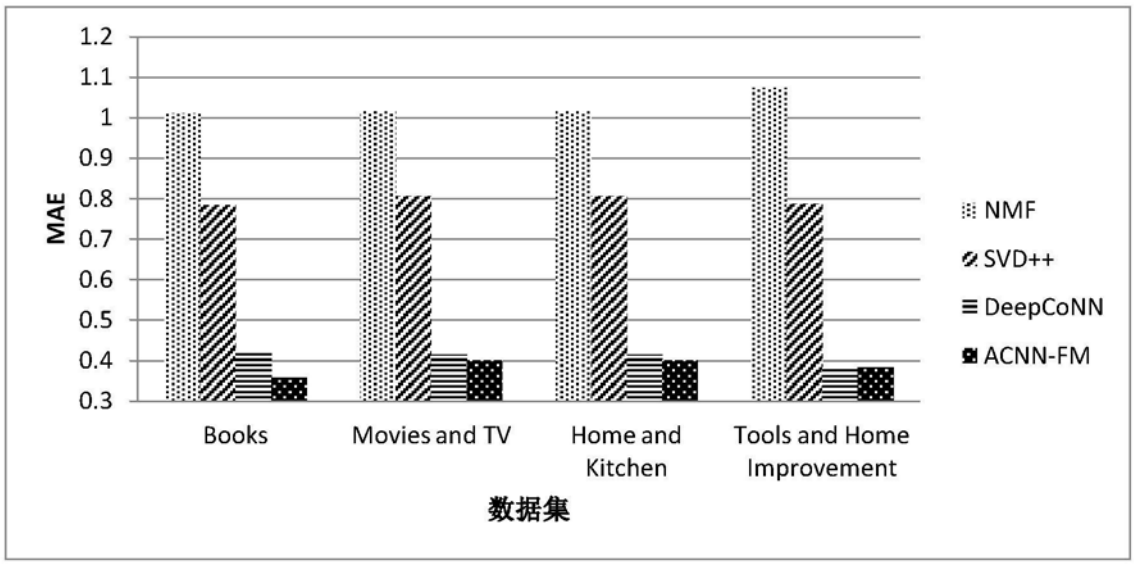

1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
MAE
Books
Movies and TV
Home and Kitchen
Tools and Home Improvement
数据集
NMF
SVD++
DeepCoNN
ACNN-FM

图8(b)

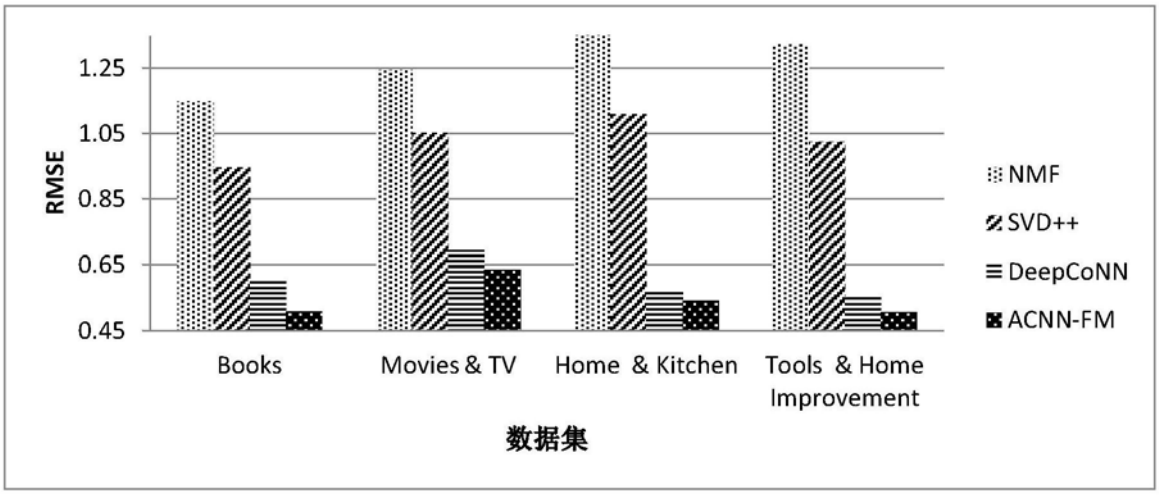

RMSE
1.25
1.05
0.85
0.65
0.45
Books
Movies & TV
Home & Kitchen
Tools & Home Improvement
数据集
NMF
SVD++
DeepCoNN
ACNN-FM

图9(a)

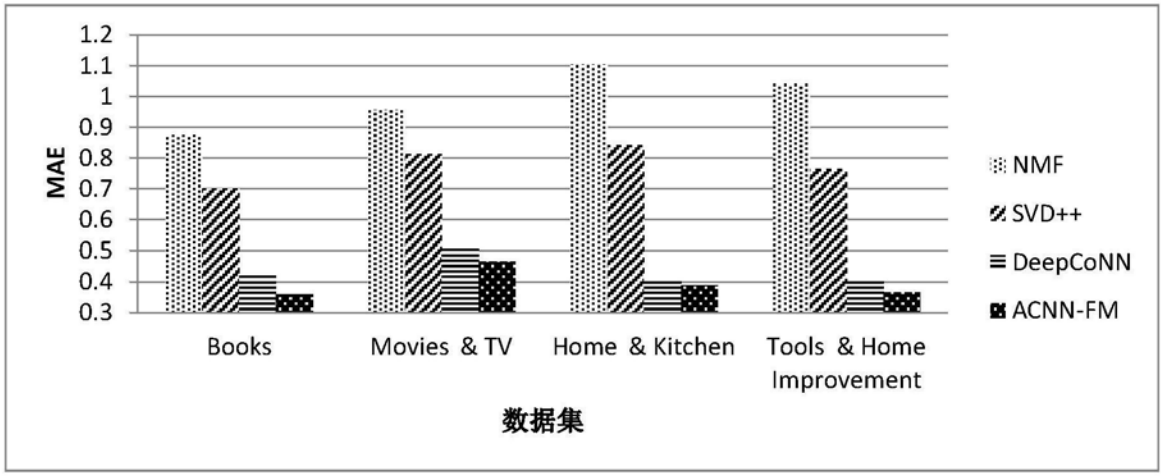

MAE
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Books
Movies & TV
Home & Kitchen
Tools & Home Improvement
数据集
NMF
SVD++
DeepCoNN
ACNN-FM

图9(b)

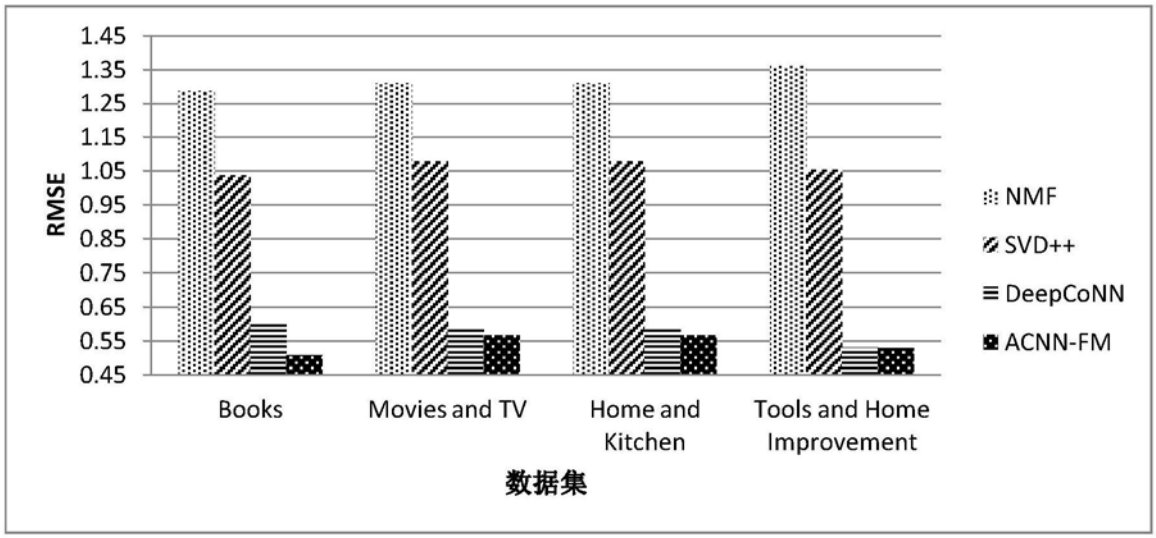

1.45
1.35
1.25
1.15
1.05
0.95
0.85
0.75
0.65
0.55
0.45
RMSE
Books
Movies and TV
Home and Kitchen
Tools and Home Improvement
数据集
NMF
SVD++
DeepCoNN
ACNN-FM

图10 (a)

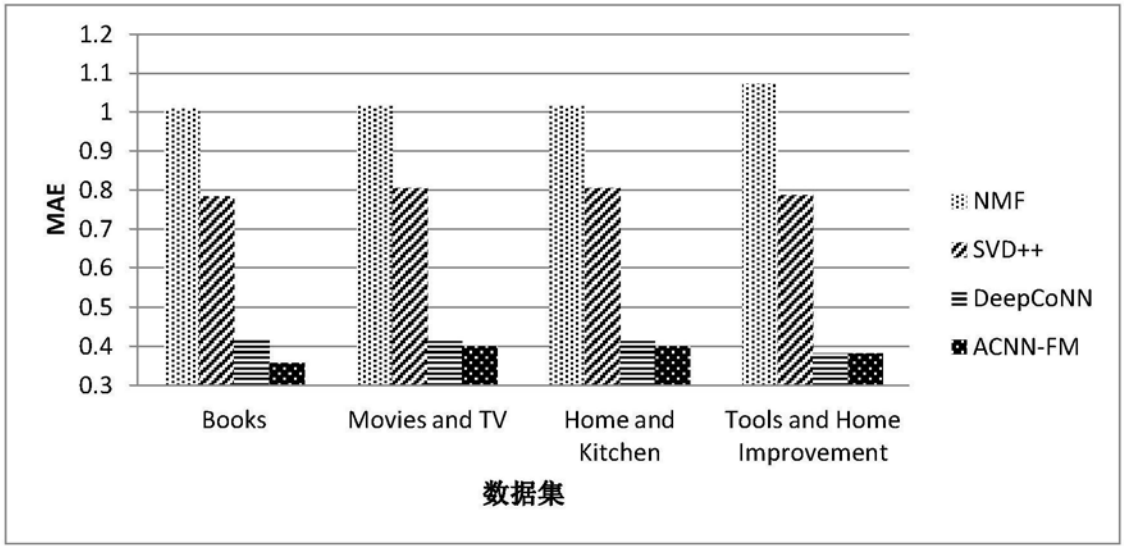

1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
MAE
Books
Movies and TV
Home and Kitchen
Tools and Home Improvement
数据集
NMF
SVD++
DeepCoNN
ACNN-FM

图10 (b)

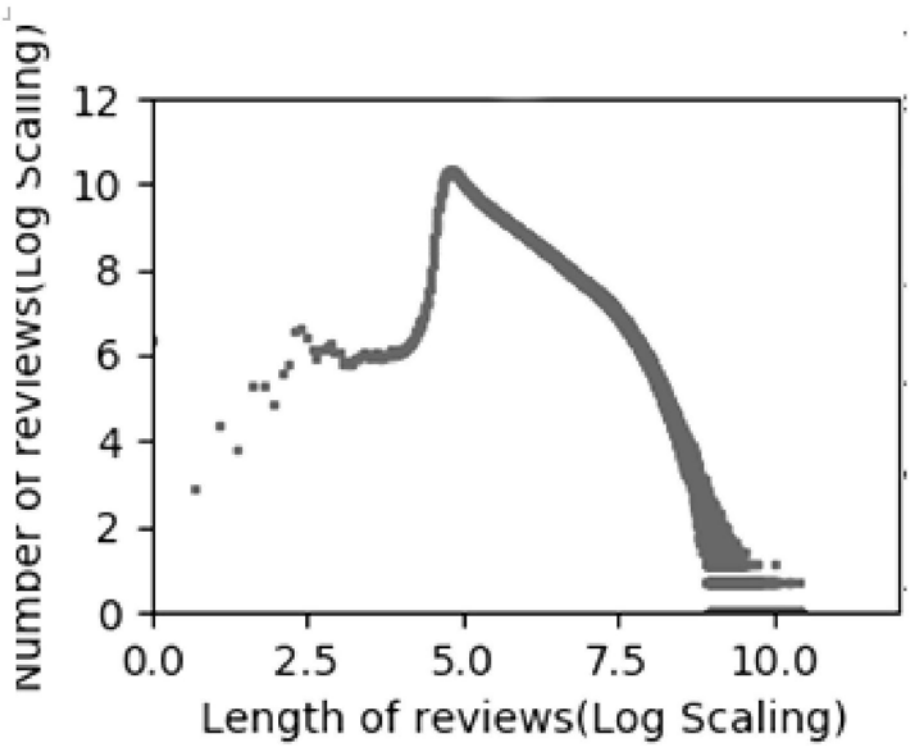

Number of reviews(Log Scaling)
12
10
8
6
4
2
0
0.0
2.5
5.0
7.5
10.0
Length of reviews(Log Scaling)

图11 (a)

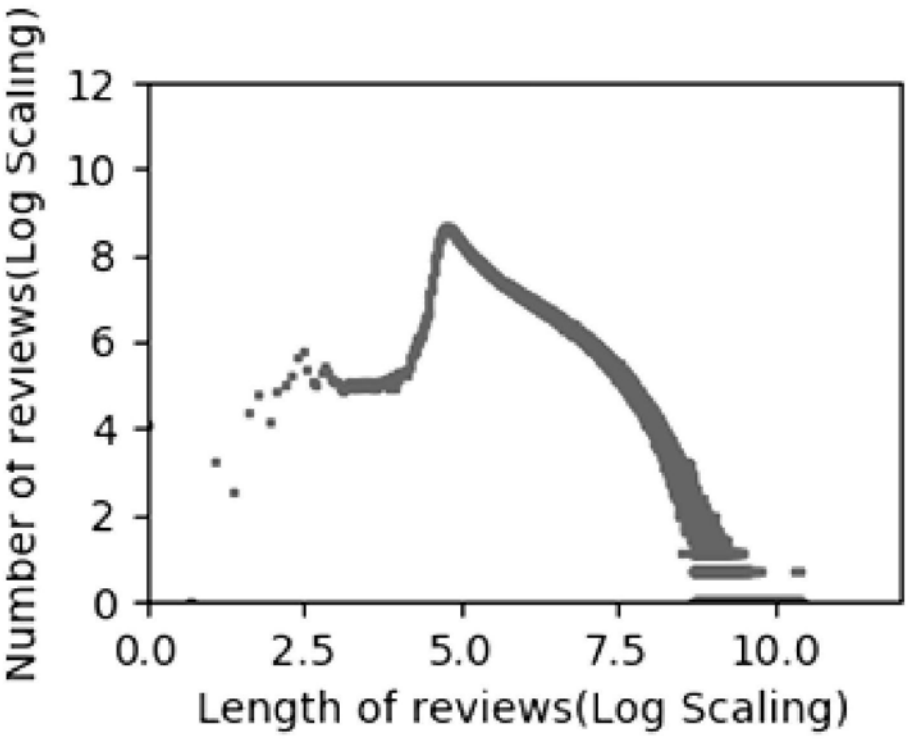

Number of reviews(Log Scaling)
12
10
8
6
4
2
0
0.0
2.5
5.0
7.5
10.0
Length of reviews(Log Scaling)

图11 (b)

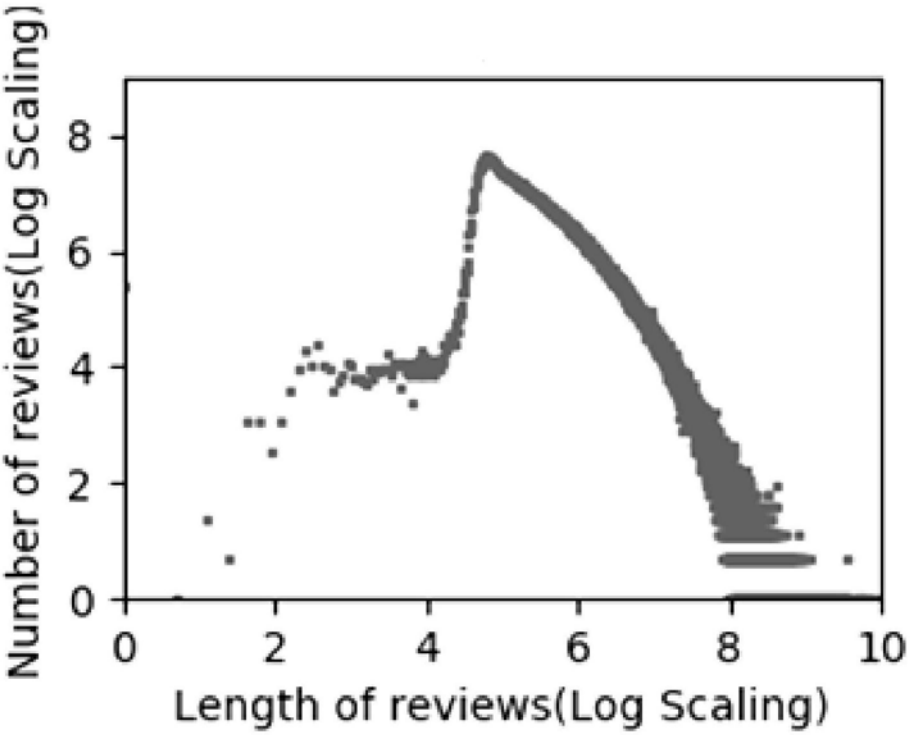

Number of reviews(Log Scaling)
0
2
4
6
8
0
2
4
6
8
10
Length of reviews(Log Scaling)

图11(c)

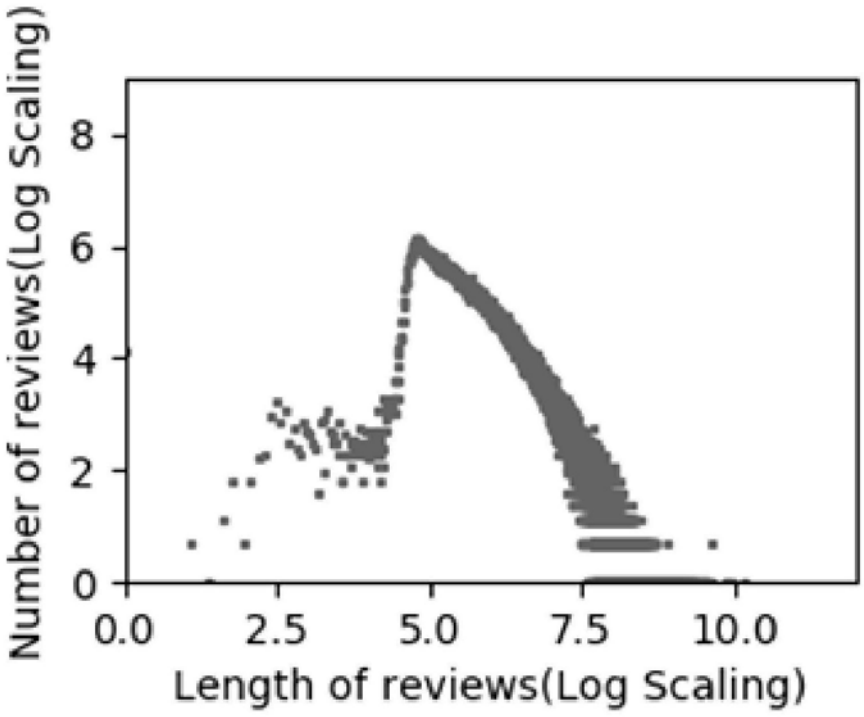

Number of reviews(Log Scaling)
0
2
4
6
8
0.0
2.5
5.0
7.5
10.0
Length of reviews(Log Scaling)

图11(d)

数据集	数据占比	最小字符长度	最长字符长度
Books	0.8	84	1351
	0.6	91	620
Movies & TV	0.8	10	1587
	0.6	84	749
Home & Kitchen	0.8	1	796
	0.6	94	435
Tools & Home Improvement	0.8	1	1014
	0.6	94	526
平均		58	885

图12

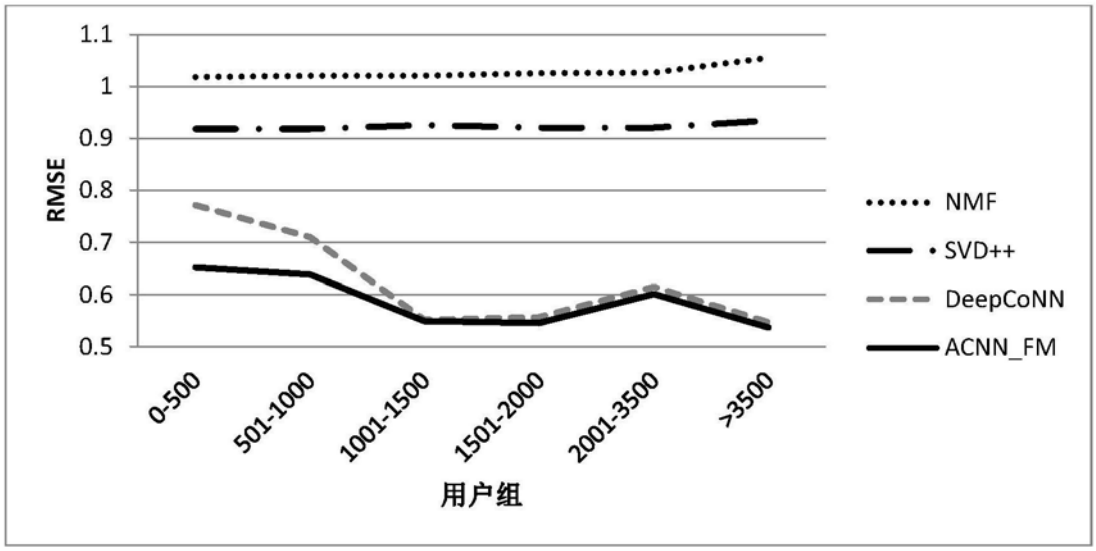


图13(a)

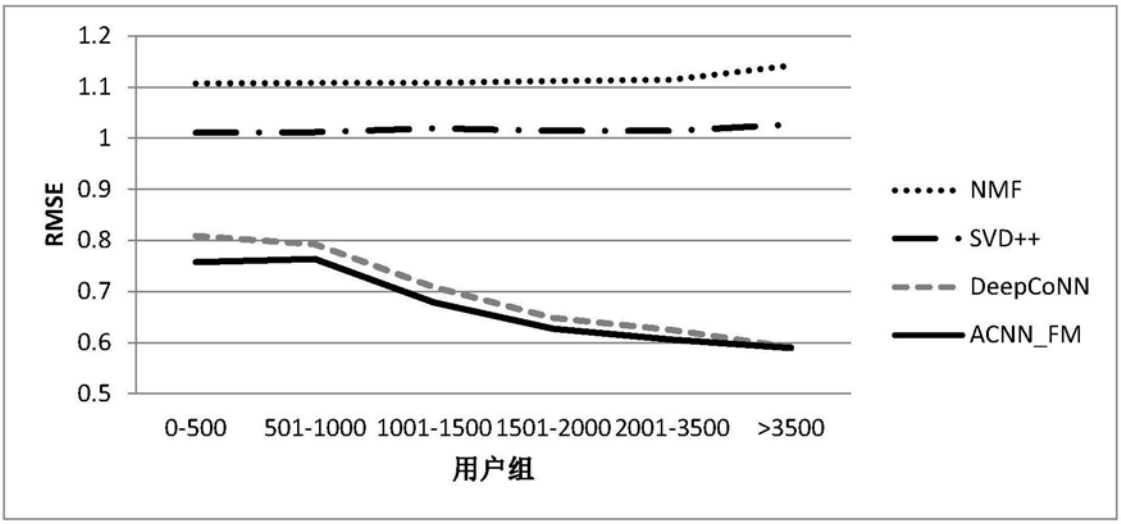

1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
RMSE
0-500
501-1000
1001-1500
1501-2000
2001-3500
>3500
用户组
NMF
SVD++
DeepCoNN
ACNN_FM

图13(b)

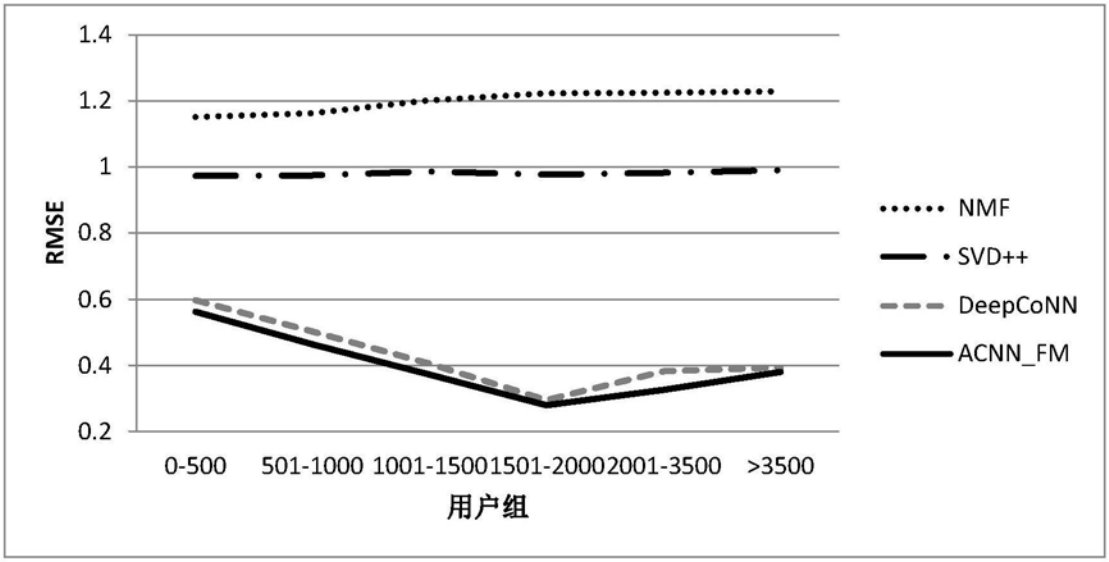

1.4
1.2
1
0.8
0.6
0.4
0.2
RMSE
0-500
501-1000
1001-1500
1501-2000
2001-3500
>3500
用户组
NMF
SVD++
DeepCoNN
ACNN_FM

图13(c)

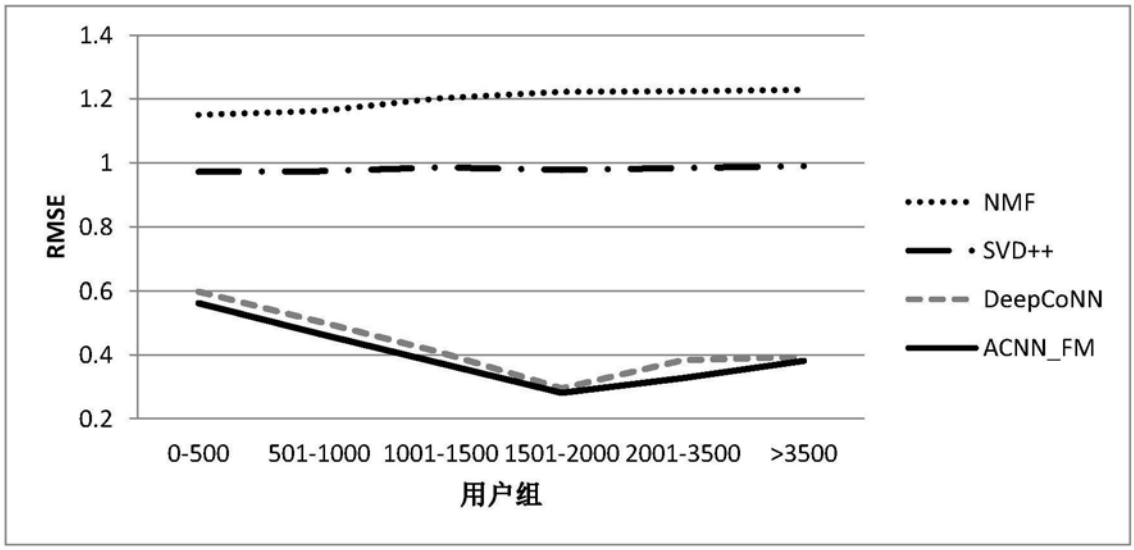


图13(d)

评测指标	数据集	评测方法			
		NMF	SVD++	DeepCoNN	ACNN-FM
Train Time(sec.)	Books	1070	21517	261605	1307026
	Movies and TV	181	3571	49908	250542
	Home and Kitchen	13.5	43.23	13454	69250
	Tools and Home Improvement	13.5	43.23	13454	69250
Test Time(sec.)	Books	36.8	372	0.15	0.55
	Movies and TV	6.79	70.8	0.11	0.51
	Home and Kitchen	0.35	1.17	0.05	0.36
	Tools and Home Improvement	0.35	1.17	0.05	0.36

图14