**Computer Science and Application** 计算机科学与应用, 2022, 12(4), 1016-1022
Published Online April 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.124104

Hans汉斯

# 基于改进隐语义模型算法的研究

**李帅阳，孙　杰**

天津工业大学，天津

收稿日期：2022年3月14日；录用日期：2022年4月14日；发布日期：2022年4月21日

## 摘　要

**本文融合了逾期因子，改进了传统的隐语义模型，隐语义模型是推荐算法中最常见的一个算法。传统的推荐算法大部分都是根据用户的反馈数据进行训练、建模，然而随着网络时代即将转为数据时代，当面对海量数据时，传统的推荐算法，可能将要面对训练时间长、速度慢、误差大的问题；传统的隐语义模型采用矩阵分解的方法来实现，这种方法最大的优点就是在无需了解分解矩阵因子特征的同时，还能尽可能的提高推荐准确度，但是这种方法需要不断地迭代训练来优化特征向量，训练一次可能需要更大的训练维度和更高的复杂度，以上问题给推荐算法和隐语义模型保留了很大的提升空间。在实际生活中，人们对事物的兴趣很可能会跟随时间的推移而出现变化，当不考虑时间信息的时候，很可能对推荐结果产生影响，推荐的准确率就不一定满足人们的实际需求了。为了提升隐语义模型的效率，本文融合了逾期因子，根据对数函数和反比例函数的特性，完成了对隐语义模型进行改进。通过使用MovieLens数据集进行实验，利用平均绝对误差、均方根误差和损失函数值作为评价指标，改进的隐语义模型对比传统隐语义模型算法的实验结果显示，改进的算法降低了训练维度，提升了训练速度，降低了训练误差，同时也提高了推荐的准确性，有效的改进了传统的隐语义模型算法。**

## 关键词

**推荐算法，隐语义模型，逾期因子**

# Research on Improved Latent Semantic Model Algorithm

**Shuaiyang Li, Jie Sun**

Tiangong University, Tianjin

Received: Mar. 14th, 2022; accepted: Apr. 14th, 2022; published: Apr. 21st, 2022

## Abstract

**This paper integrates the overdue factor and improves the traditional latent semantic model,**

文章引用: 李帅阳, 孙杰. 基于改进隐语义模型算法的研究[J]. 计算机科学与应用, 2022, 12(4): 1016-1022.
DOI: 10.12677/csa.2022.124104

**which is the most common algorithm in recommendation algorithms. Most of the traditional recommendation algorithms are trained and modeled based on user feedback data. However, as the network era is about to turn into the data era, when faced with massive data, traditional recommendation algorithms may face a long training time the problem of slow speed and large error; the traditional implicit semantic model is implemented by the method of matrix decomposition. The biggest advantage of this method is that it can improve the recommendation accuracy as much as possible without knowing the characteristics of the decomposition matrix factor, but this method requires continuous iterative training to optimize the feature vector. Training once may require larger training dimensions and higher complexity. The above problems leave a lot of room for improvement for recommendation algorithms and latent semantic models. In real life, people's interest in things is likely to change with the passage of time. When time information is not considered, it is likely to have an impact on the recommendation results, and the accuracy of the recommendation may not meet people's actual needs. In order to improve the efficiency of the latent semantic model, this paper integrates the overdue factor, and completes the improvement of the latent semantic model according to the characteristics of the logarithmic function and the inverse proportional function. By using the MovieLens data set to conduct experiments, using the mean absolute error, root mean square error and loss function value as evaluation indicators, the experimental results of the improved latent semantic model compared with the traditional latent semantic model algorithm show that the improved algorithm reduces the training dimension and improves the training speed, reduces the training error, improves the accuracy of recommendation, and effectively improves the traditional latent semantic model algorithm.**

## Keywords

**Recommendation Algorithm, LFM, Overdue Factor**



## 1. 引言

随着科技时代的发展与变迁，当今的世界，已经从网络时代迈向了数据时代。众所周知，互联网打破了人们获取信息的传统方式，使得人们能够从网上轻松便利的获取信息；然而当走进数据时代时，面对大量的数据，人们的需求和喜好在某些方面不明确时，人们大多是漫无目的的在接受或寻求信息，这不仅没有节省人们的时间，反而会消耗更多的精力和时间去过滤这些超载的信息，这种现象称为“信息过载”。面对信息过载[1]，消费者如何更高效的找到对自己有用的信息呢？生产信息的平台如何推出更加适合用户需求的信息呢？

推荐算法由此而生，推荐算法就是利用用户与物品的交互数据，通过一些数学算法，推测出用户可能喜欢的东西。推荐算法主要分三种，分别是：基于内容、协同过滤和混合推荐算法，其中研究最多的是协同过滤，协同过滤中的算法非常丰富，本文主要研究的是隐语义模型算法，隐语义模型是推荐算法中最常见的一个算法[2]，该算法使用的是用户的历史评分和用户的交互信息，挖掘出数据中用户及物品隐含的特征，然后推荐出用户个性化的数据。

在传统的隐语义模型中，都是根据用户的反馈数据进行建模，训练一次可能需要更大的训练维度和更高的复杂度[3]，并且上下文信息使用率很低，尤其是时间信息；在实际生活中，人们对事物的兴趣很可能会跟随时间的推移而出现变化，当不考虑时间信息的时候，很可能对推荐结果产生影响，推荐的准确率就

不一定满足人们的实际需求了。本文结合了时间信息，将时间差定义为逾期因子，改进了传统的隐语义模型，降低了训练维度，降低了训练复杂度，提升了训练速度，减小了误差，同时也提高了推荐的准确性。

## 2. 基于隐语义模型的推荐算法

隐语义模型又名为隐因子模型(latent factor model，简写 LFM) [4]，属于协同过滤算法的一种。它先基于矩阵分解算法建立潜在因子模型，再依据机器学习和优化理论处理评分矩阵，从而获取用户的潜在特征并预测用户对未评分物品的评分[5]。

假设存在 $n$ 个用户和 $m$ 个物品，先获取每个用户对每个物品的评分，构建出一个评分矩阵 $R_{n\times m}$。而 LFM 是在设置特征的维度 $K$ 后，寻找两个低维矩阵 $P_{n\times k}$ 和 $Q_{m\times k}$，分别将其作为用户和物品的特征矩阵，再通过对特征矩阵相乘得到预测的评分矩阵 $R'_{n\times m}$。在为用户推荐时，根据用户对每个物品的预测得分进行降序排序，选出前 $N$ 个当前用户未评分的物品推荐给用户[6]。

为了使预测的结果更加精确，需要不断迭代改变维度 $K$ 和矩阵 $P_{n\times k}$ 和 $Q_{m\times k}$ 的值。本文使用平方损失函数 $cost$ 来量化预测评分矩阵与实际评分矩阵的差别，其计算公式如下：

$$cost = \sum_{(u,i)\in R}\left(R_{ui} - P_u Q_i^{\mathrm{T}}\right)^2 \tag{1}$$

为了防止过拟合的情况出现，本文在损失函数的基础上增加了正则化项，修正后的损失函数如下：

$$cost = \sum_{(u,i)\in R}\left(R_{ui} - P_u Q_i^{\mathrm{T}}\right)^2 + \lambda\left(\sum_u \|P_u\|^2 + \sum_i \|Q_i\|^2\right) \tag{2}$$

其中 $R_{ui}$ 表示第 $u$ 个用户对第 $i$ 个电影的评分，$P_u$ 表示第 $u$ 个用户的特征向量，$Q_i^{\mathrm{T}}$ 表示第 $i$ 个物品特征向量的转置，$\lambda$ 表示正则化系数。

为使损失值降低，推荐结果更加精准，本文采用梯度下降法进行迭代，不断迭代用户特征矩阵 $P_{n\times k}$ 和物品特征矩阵 $Q_{m\times k}$，本文实验中的迭代次数为 10,000 次，根据实验结果得知，预测矩阵 $R'_{n\times m}$ 已经取得最优值。用梯度下降法进行迭代的公式如下：

$$P := P - \alpha\frac{\sigma cost}{\sigma P} = P + \alpha\left(\sum_i 2\left(R_{ui} - P_u Q_i^{\mathrm{T}}\right)Q_i + 2\lambda P_u\right) \tag{3}$$

$$Q := Q - \alpha\frac{\sigma cost}{\sigma Q} = Q + \alpha\left(\sum_i 2\left(R_{ui} - P_u Q_i^{\mathrm{T}}\right)P_u + 2\lambda Q_i\right) \tag{4}$$

其中 $\alpha$ 表示学习率，即梯度下降的步长；$\frac{\sigma cost}{\sigma P}$ 表示损失函数 $cost$ 对用户特征矩阵 $P$ 求偏导；$\frac{\sigma cost}{\sigma Q}$ 表示损失函数 $cost$ 对物品特征矩阵 $Q$ 求偏导。

## 3. 改进的隐语义模型算法

### 3.1. 对数函数

对数函数是 6 类基本初等函数之一。一般地，函数 $y = \log_a x$ 叫做对数函数，其中 $x$ 是自变量，$a$ 为底数[7]。当 $a > 0$ 时，对数函数是单调递增函数，且其斜率逐渐减小，在趋于无穷大的时候，其斜率趋于 0，函数值将趋于一个固定值。根据这一特性，结合时间信息对梯度下降函数中步长的影响，即时间间隔越久，逾期因子越大，对推荐结果影响越小，对应步长应该增大，当逾期因子越来越大的时候，对应步长应该趋于一个固定值。

本文中底数为 e，e 的定义为：

$$\mathrm{e} = \lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x = 2.718281828\cdots$$

本文采用的对数函数是以 e 为底，逾期因子为自变量 $x$ 的对数函数，以 e 为低的对数函数的可表示为 Ln($x$)，其图像如图 1 所示：

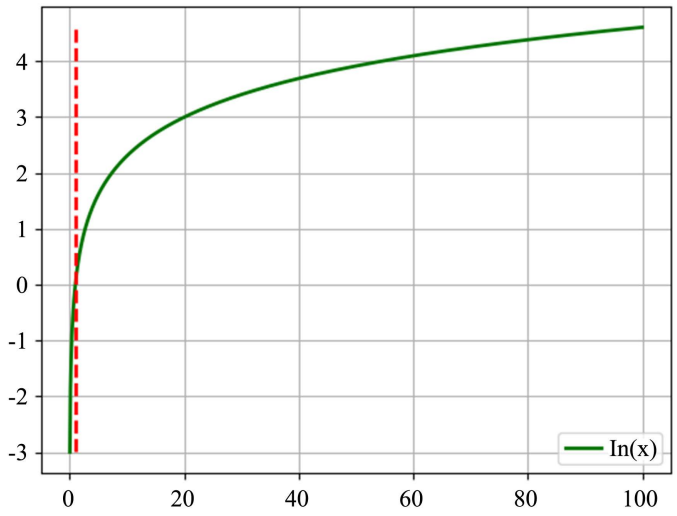


**Figure 1.** $y$ = Ln($x$) function image
**图 1.** $y$ = Ln($x$)函数图像

### 3.2. 反比例函数

一般的，如果两个变量 $x$，$y$ 之间的关系可以表示为 $y = k/x$ ($k$ 为常数，$k \neq 0$，$x \neq 0$)，其中 $k$ 叫做反比例系数，$x$ 是自变量，$y$ 是 $x$ 的函数。当 $k > 0$ 时，图像在一、三象限；当 $k < 0$ 时，图像在二、四象限。$k$ 的绝对值表示的是 $x$ 与 $y$ 的坐标形成的矩形的面积[8]。当 $k > 0$，且 $x > 0$ 的时候，反函数在第一象限是单调递减，且斜率逐渐增大，当 $x$ 趋于无穷时，斜率趋于 0，函数值趋于 0；根据这一特性，结合逾期因子对正则化系数的影响，即逾期因子越大，过拟合的影响就越小，正则化系数对应的应该减小；当逾期因子趋于一个无穷大的值时，正则化系数应该趋于 0。

本文中令 $k = 1$，逾期因子为自变量的函数，可表示为 $y = 1/x$，其图像如图 2 所示：

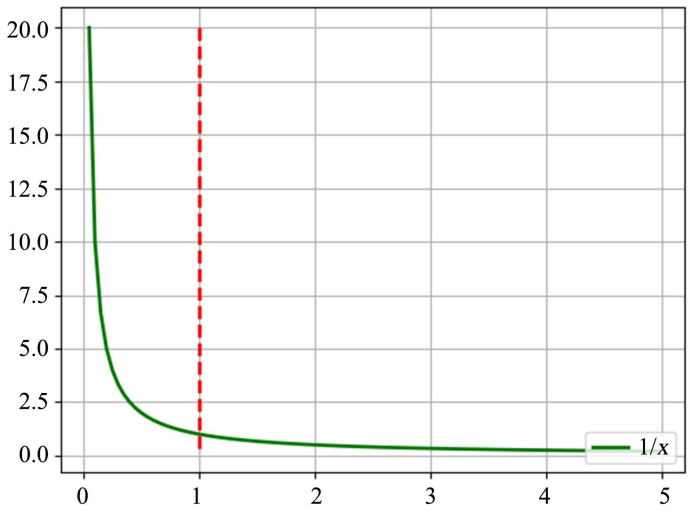


**Figure 2.** $y = 1/x$ function image
**图 2.** $y = 1/x$ 函数图像

### 3.3. 改进的隐语义模型

根据对数函数和反比例函数的函数特性，融合逾期因子，构造出改进的隐语义模型。令逾期因子为$\Delta T$，其对应的损失函数为：

$$cost = \sum_{(u,i)\in R}\left(R_{ui} - P_u Q_i^{\mathrm{T}}\right)^2 + \lambda\frac{1}{\Delta T}\left(\sum_u \|P_u\|^2 + \sum_i \|Q_i\|^2\right) \tag{5}$$

对应的梯度下降公式可更改为：

$$P := P + \alpha\ln(\Delta T)\left(\sum_i 2\left(R_{ui} - P_u Q_i^{\mathrm{T}}\right)Q_i + 2\lambda\frac{1}{\Delta T}P_u\right) \tag{6}$$

$$Q := Q + \alpha(\Delta T)\left(\sum_i 2\left(R_{ui} - P_u Q_i^{\mathrm{T}}\right)P_u + 2\lambda\frac{1}{\Delta T}Q_i\right) \tag{7}$$

## 4. 实验结果分析

### 4.1. 平均绝对误差

平均绝对误差(Mean Absolute Error，简称 MAE)，它是所有单个观测值与算术平均值的偏差的绝对值的平均[9]。平均绝对误差可以避免误差相互抵消的问题，因而可以准确反映实际预测误差的大小。平均绝对误差可表示为：

$$\mathrm{MAE} = \frac{\sum_{u(,i)\in T}\left|R_{ui} - R'_{ui}\right|}{|T|}$$

### 4.2. 均方根误差

均方根误差(Root Mean Square Error，简称 RMSE)，它是预测值与真实值偏差的平方与观测次数 $n$ 比值的平方根，在实际测量中，观测次数 $n$ 总是有限的，真值只能用最可信赖(最佳)值来代替[10]。均方根误差可表示为：

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{u(,i)\in T}\left(R_{ui} - R'_{ui}\right)^2}{|T|}}$$

### 4.3. 实验结果比较

本次实验使用 MovieLens 数据集，利用平均绝对误差、均方根误差和损失函数值作为评价指标，分别对传统的隐语义模型和改进隐语义模型算法进行训练、计算，最终统计出对比结果。如图 3~5 所示，其中绿色线代表传统隐语义模型，蓝色线代表改进的隐语义模型，对应图中横轴表示 $K$ 值大小，竖轴表示对应的结果值。

根据图 3 和图 4 所示，改进隐语义模型算法的平均绝对误差和均方根误差均比传统隐语义模型下降的快，并且能够在保证优先找到最小 K 值的前提下，对应误差值减小到最小值后保持不变，且改进的隐语义模型比相同 K 值下的传统隐语义模型误差要小。

根据图 5 所示，改进隐语义模型算法的损失值比传统隐语义模型的损失值相差无几，但改进的隐语义模型能够优先找到最小 $K$ 值，并且优先达到最小损失值。

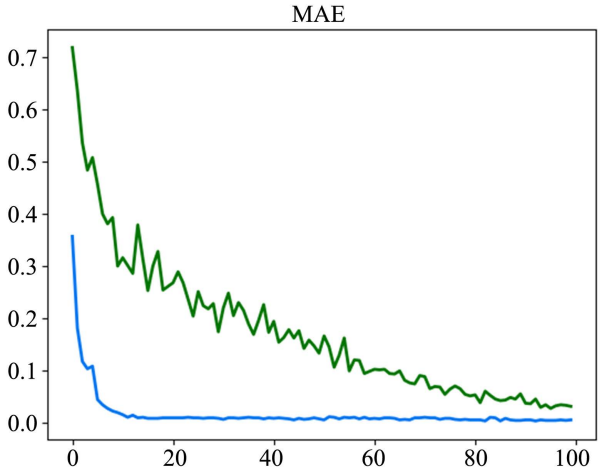


**Figure 3.** Mean absolute error comparison chart
**图 3.** 平均绝对误差对比图

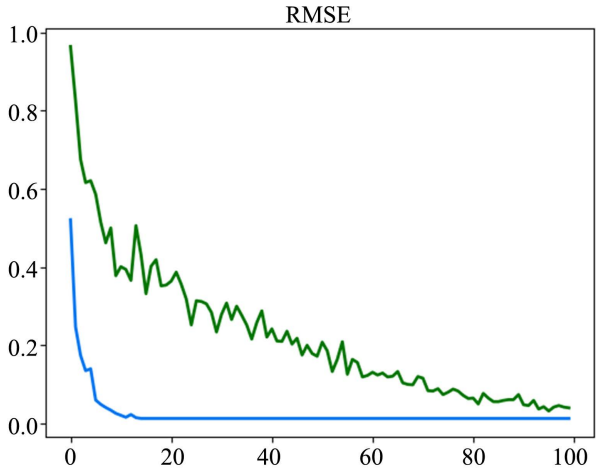


**Figure 4.** Root mean square error comparison chart
**图 4.** 均方根误差对比图

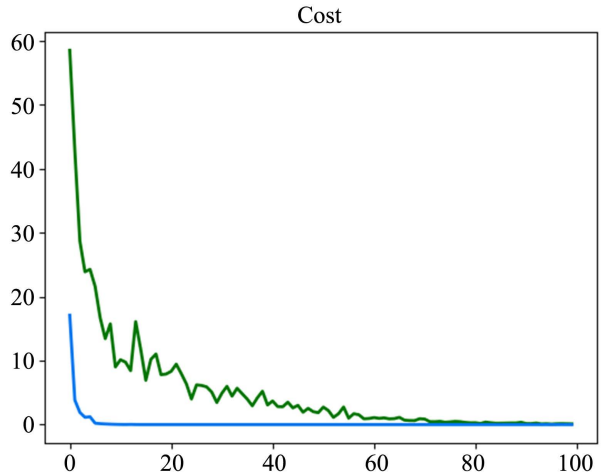


**Figure 5.** Loss function comparison chart
**图 5.** 损失函数对比图

## 5. 结论

根据上述分析，改进的隐语义模型算法，降低了训练维度，降低了训练复杂度，同时提升了训练速度，降低了训练误差。 此次研究改进隐语义模型算法有效的提高了传统的隐语义模型。

## 参考文献


[1] Kim, J., Kwon, E., Cho, Y. and Kang, S. (2011) Recommendation System of IPTV TV Program Using Ontology and K-means Clustering. In: Kim, Th., Adeli, H., Robles, R.J. and Balitanas, M., Eds., *Ubiquitous Computing and Multimedia Applications, UCMA* 2011, *Communications in Computer and Information Science*. Vol. 151, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-20998-7_16

[2] Le, N.H.N. (2022) Incorporating Textual Reviews in the Learning of Latent Factors for Recommender Systems. *Electronic Commerce Research and Applications*, 52, Article ID: 101133. https://doi.org/10.1016/j.elerap.2022.101133

[3] Shen, R. (2022) IA Recommender System Integrating Long Short-Term Memory and Latent Factor. *Arabian Journal for Science and Engineering*, 1-11. https://doi.org/10.1007/s13369-021-05933-9

[4] 杨春. 基于 RBM 模型和 LFM 模型的推荐算法研究与实现[D]: [硕士学位论文]. 重庆: 重庆邮电大学, 2020.

[5] 陈晔, 刘志强. 基于 LFM 矩阵分解的推荐算法优化研究[J]. 计算机工程与应用, 2019, 55(2): 116-120+167.

[6] 彭宇, 宁慧, 张汝波. LFM 基于改进的 LFM 算法的短视频推荐系统的研究与实现[J/OL]. 应用科技. https://kns.cnki.net/kcms/detail/23.1191.u.20220217.0943.002.html, 2022-02-17.

[7] 百度百科“对数函数”词条[EB/OL]. https://baike.baidu.com/item/%E5%AF%B9%E6%95%B0%E5%87%BD%E6%95%B0/6013318?fr=aladdin#ref_[1]_331649, 2022-01-22.

[8] 百度百科“反比例函数”词条[EB/OL]. https://baike.baidu.com/item/%E5%8F%8D%E6%AF%94%E4%BE%8B%E5%87%BD%E6%95%B0/3228967?fr%20=%20aladdin, 2021-12-13.

[9] 百度百科“平均绝对误差”词条[EB/OL]. https://baike.baidu.com/item/%E5%B9%B3%E5%9D%87%E7%BB%9D%E5%AF%B9%E8%AF%AF%E5%B7%AE/9383373?fr=aladdin, 2021-06-24.

[10] 百度百科“均方根误差”词条[EB/OL]. https://baike.baidu.com/item/均方根误差/3498959, 2021-12-13.