聚美优品
item-based CF协同过滤
IUF及归一化改进

赵海臣

IUF(Inverse User Frequency)背景

- 假设有一个用户，是开化妆品店的，并且买了聚美80%的商品准备来自己卖，如果按照之前商品相似度的算法，那么这个用户的行为将导致80%的商品间两两产生相似度。
- 这个用户虽然活跃，但是他的购物行为并非出于自身的兴趣，所以这个用户对于他所购买的化妆品的两两相似度的贡献应该远小于一个只买了十几种化妆品的女孩。

IUF(Inverse User Frequency)

- IUF，即用户活跃度对数的倒数的参数，是假设活跃用户对物品的相似度的贡献应该小于不活跃的用户：

$$w_{ij} = \frac{\sum_{u \in N(i) \cap N(j)} \frac{1}{\log(1+|N(u)|)}}{\sqrt{|N(i)||N(j)|}}$$

  - 与标准余弦相似度相比，分子部分的共同清单计数换成了共同清单对应用户的活跃度对数的倒数的参数。
    - *由于我们使用”userId”, “data_date”联合作为划分粒度，所以其中N(i)同样为”userId”, “data_date”的联合划分粒度。

性能提升

- 与itemCF相比，itemCF-IUF
  - 在准确率和召回率上与标准算法相近
  - 明显提高了结果的覆盖率并且降低了商品的流行度

itemCF-IUF确实改进了itemCF的性能。

表2-9 MovieLens数据集中ItemCF算法和ItemCF-IUF算法的对比

	准确率	召回率	覆盖率	流行度
ItemCF	22.28%	10.76%	18.84%	7.254526
ItemCF-IUF	22.29%	10.77%	19.70%	7.217326

CF+点击模型下的IUF处理

- 对指向商品A自身的IUF值乘以固定加权系数clickConcernValue (<= 1.0)：
  - IUF(A – A) = IUF(A – A) * clickConcernValue。

物品相似度的归一化

- 将itemCF的相似度矩阵按最大值归一化，可以提高推荐的准确率
  - 将已经得到的物品相似度矩阵w进行归一化：

$$w'_{ij} = \frac{w_{ij}}{\max\limits_{j} w_{ij}}$$

性能提升

- 与itemCF相比，itemCF-Norm
  - 提升了准确率、召回率、覆盖率以及降低了流行度

各方面都有提升。

表2-10 MovieLens数据集中ItemCF算法和ItemCF-Norm算法的对比

	准确率	召回率	覆盖率	流行度
ItemCF	22.28%	10.76%	18.84%	7.254526
ItemCF-Norm	22.73%	10.98%	23.73%	7.157385

实际测试

- *不能完全与之前的“聚美优品CF”结果数据进行对比，因为点击模型、复购策略都有改动。

有IUF、有归一化

- **********推荐[_CF_test]**********
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.443241][0.010180][7970234.685989][0.403239][0.808043][1.000000][11.994880][19.747940][0.006200],[0.023870] **********
- **********mainpageTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.096125][0.002110][2059454.962743][0.007006][0.013871][22.000000][30.000000] **********
- **********allTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.169021][0.003670][3073151.585620][0.007006][0.014011][20.000000][23.000000] **********
- **********sevenDaysTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.182029][0.003950][3430441.477159][0.007006][0.014011][18.000000][21.000000] **********
- **********推荐[_CF_test] **********
- lowRank = 0, highRank = 19
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.336066][0.007278][6016595.294541][0.742093][1.000000][1.000000][6.125010][6.125010][0.137420],[0.365800] **********
- **********推荐[_CF_test] **********
- lowRank = 20, highRank = 39
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.046591][0.001286][858004.350678][0.750586][1.000000][1.000000][7.574259][7.574259][0.067208],[0.222947] **********
- **********推荐[_CF_test] **********
- lowRank = 40, highRank = 59
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.026246][0.000709][481491.870405][0.754821][1.000000][1.000000][8.344018][8.344018][0.044758],[0.163488] **********
- **********推荐[_CF_test] **********
- lowRank = 60, highRank = 79
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- [illegible][1.000000][1.000000][8.963544][8.963544][0.030495],[0.120050] **********
- [illegible]**
- lowRa[illegible]
- ***** 召回,准确率,[illegible]覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.014917][0.000397][illegible][1.000000][1.000000][9.419124][9.419124][0.021044],[0.091998] **********

有IUF、无归一化

- **********推荐[_CF_test]**********
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.448641][0.010302][8037816.236094][0.379174][0.777910][1.000000][12.353111][20.329210][0.005020],[0.022251] **********
- **********mainpageTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.096125][0.002110][2059454.962743][0.007006][0.013871][22.000000][30.000000] **********
- **********allTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.169021][0.003670][3073151.585620][0.007006][0.014011][20.000000][23.000000] **********
- **********sevenDaysTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.182029][0.003950][3430441.477159][0.007006][0.014011][18.000000][21.000000] **********
- **********推荐[_CF_test] **********
- lowRank = 0, highRank = 19
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.341734][0.007392][6099827.374678][0.738974][1.000000][1.000000][6.307519][6.307519][0.129724],[0.347100] **********
- **********推荐[_CF_test] **********
- lowRank = 20, highRank = 39
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.046564][0.001292][852708.420687][0.748992][1.000000][1.000000][7.716016][7.716016][0.066638],[0.214156] **********
- **********推荐[_CF_test] **********
- lowRank = 40, highRank = 59
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.025719][0.000699][472944.410309][0.754022][1.000000][1.000000][8.509231][8.509231][0.042507],[0.154425] **********
- **********推荐[_CF_test] **********
- lowRank = 60, highRank = 79
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- [redacted]1.000000][1.000000][9.144305][9.144305][0.028195],[0.112119] **********
- [redacted]**
- lowRa[redacted]
- ***** 召回,准确率,[redacted]覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.015221][0.000401][redacted][1.000000][1.000000][9.600500][9.600500][0.018984],[0.085437] **********

无IUF、无归一化

- **********推荐[_CF_test]**********
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.438661][0.010051][7881399.056042][0.396729][0.793486][1.000000][11.615458][19.338810][0.007640],[0.029851] **********
- **********mainpageTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.096125][0.002110][2059454.962743][0.007006][0.013871][22.000000][30.000000] **********
- **********allTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.169021][0.003670][3073151.585620][0.007006][0.014011][20.000000][23.000000] **********
- **********sevenDaysTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.182029][0.003950][3430441.477159][0.007006][0.014011][18.000000][21.000000] **********
- **********推荐[_CF_test] **********
- lowRank = 0, highRank = 19
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.325036][0.006990][5826375.374403][0.745273][1.000000][1.000000][5.811284][5.811284][0.164275],[0.406682] **********
- **********推荐[_CF_test] **********
- lowRank = 20, highRank = 39
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.049106][0.001335][893761.620627][0.750973][1.000000][1.000000][7.403348][7.403348][0.081340],[0.242379] **********
- **********推荐[_CF_test] **********
- lowRank = 40, highRank = 59
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.027485][0.000746][501778.350496][0.755393][1.000000][1.000000][8.317804][8.317804][0.051399],[0.171689] **********
- **********推荐[_CF_test] **********
- lowRank = 60, highRank = 79
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- [redacted][1.000000][1.000000][8.952071][8.952071][0.033486],[0.126373] **********
- [redacted]**
- lowRa[redacted]
- ***** 召回,准确率,[redacted]覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.016433][0.000430][redacted][1.000000][1.000000][9.410868][9.410868][0.023444],[0.096378] **********

结论

- IUF+NORM方式能提高CF的召回率、覆盖率、销售额、多样性。
- IUF方式能提高CF的召回率、覆盖率、销售额，但是相对于+NORM，多样性明显降低。

对购买过的同类商品限制措施

- 通过使用item-similarity相似度矩阵对CF的用户召回阶段进行双重屏蔽处理：
  - 对用户召回前的浏览数据进行处理：
    - 根据item-similarity矩阵，对用户最近购买的商品召回最近的top N个商品，对用户的近似商品浏览记录进行屏蔽，从而达到在接下来的召回阶段失去召回源头；
  - 对用户召回后的初始推荐列表进行处理：
    - 根据item-similarity矩阵，对用户最近购买的商品召回最近的top N个商品，对用户cf输出的初始推荐列表中的商品进行屏蔽；

无IUF、无归一化+CF复购N=10近似商品限制

- **********推荐[_CF_test]**********
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.202655][0.004498][4042554.023578][0.411547][0.805251][1.000000][11.920826][19.755258][0.006344],[0.027192] **********
- **********mainpageTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.096125][0.002110][2059454.962743][0.007006][0.013871][22.000000][30.000000] **********
- **********allTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.169021][0.003670][3073151.585620][0.007006][0.014011][20.000000][23.000000] **********
- **********sevenDaysTOP召回**********
- ***** 召回,准确率,銷售額,top50覆盖率,top100覆盖率,top50多样性,top100多样性**********
- ***** [0.182029][0.003950][3430441.477159][0.007006][0.014011][18.000000][21.000000] **********
- **********推荐[_CF_test] **********
- lowRank = 0, highRank = 19
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.133668][0.002828][2730605.152334][0.752448][1.000000][1.000000][6.021488][6.021488][0.154076],[0.383819] **********
- **********推荐[_CF_test] **********
- lowRank = 20, highRank = 39
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.026896][0.000652][519894.840210][0.751650][1.000000][1.000000][7.519881][7.519881][0.077977],[0.235685] **********
- **********推荐[_CF_test] **********
- lowRank = 40, highRank = 59
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.017095][0.000412][331226.940452][0.757522][1.000000][1.000000][8.370302][8.370302][0.048245],[0.168828] **********
- **********推荐[_CF_test] **********
- lowRank = 60, highRank = 79
- ***** 召回,准确率,銷售額,top10覆盖率,top50覆盖率,top100覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- [redacted]1.000000][1.000000][8.994614][8.994614][0.032398],[0.125413] **********
- [redacted]**
- lowRa[redacted]
- ***** 召回,准确率,[redacted]覆盖率,top50多样性,top100多样性,2个以下百分比,4个以下百分比**********
- ***** [0.011616][0.0002[redacted][1.000000][1.000000][9.455421][9.455421][0.022297],[0.096701] **********

结论

- 对同类商品的限制措施将造成较大的推荐效率下降。

The end