Software Engineering and Applications 软件工程与应用, 2020, 9(5), 345-351
Published Online October 2020 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2020.95039

Hans汉斯

污染源信息推荐的协同过滤算法应用模型

王丽娜

海南师范大学经济与管理学院，海南 海口
Email: lina1976113@126.com

收稿日期：2020年9月22日；录用日期：2020年10月5日；发布日期：2020年10月12日

摘　要

在世界范围内，先进的信息系统被应用到各行各业。在环境保护领域，由于污染给社会生活带来了非常多的困扰，以及污染源的固有特性，作为污染源信息需求者的环境保护机构和个人，从大量污染源信息中找到自己关注的信息往往不是一件容易的事情。推荐系统就是解决这一矛盾的主要工具。通过建立分析用户喜好模型，采用UFTB算法从用户看过的污染源信息及其信息类型入手，对用户看过的污染源信息类型与评分数据进行分析。在建立分析污染源信息推荐模型中，采用协同过滤算法计算修正后的余弦相似度，对缺省值进行预测以优化算法。为防止过度优化，采取剔除用户非喜好类型污染源信息，得到优化缺省值预测矩阵，将相似度数据带入推荐公式，得出数值并使用排序，找出与目标用户相似度最高的*N*个用户，根据它们的喜好对目标用户进行污染源信息推荐。

关键词

协同过滤推荐算法，相似度，污染源信息

Application Model of Collaborative Filtering Algorithm Recommended for Pollution Source Information

Lina Wang

School of Economics and Management, Hainan Normal University,
Haikou Hainan
Email: lina1976113@126.com

Received: Sep. 22nd, 2020; accepted: Oct. 5th, 2020; published: Oct. 12th, 2020

文章引用: 王丽娜. 污染源信息推荐的协同过滤算法应用模型[J]. 软件工程与应用, 2020, 9(5): 345-351.
DOI: 10.12677/sea.2020.95039

Abstract

In the world, advanced information systems are applied to all walks of life. In the field of environmental protection, because pollution has brought a lot of trouble to social life, as well as because of the inherent characteristics of pollution sources, as the source of pollution information needs, environmental protection institutions and individuals are often difficult to find their own concern of the information from a large number of pollution source information. The recommendation system is the main tool to solve this contradiction. By establishing the model of analyzing user preferences, UFTB algorithm is used to analyze the type of pollution source information and scoring data that users have seen. In establishing the recommendation model for analyzing pollution source information, the modified cosine similarity is calculated by using the co-filter algorithm, and the default value is predicted to optimize the algorithm. In order to prevent over-optimization, we should take the information of eliminating the user's non-preferred type of pollution source, get the optimization default prediction matrix, bring the similarity data into the recommended formula to get the value and use the sort, find the *N* user with the highest similarity to the target user, and recommend the target user the pollution source information according to their preferences.

Keywords

Co-Filtering Recommendation Algorithm, Similarity, Pollution Source Information

1. 引言

近年来，在世界范围内，先进的信息系统被应用到各行各业。在环境保护领域，由于污染给社会生活带来了非常多的困扰，以及污染源的固有特性，作为污染源信息需求者的环境保护机构和个人，如何从污染源信息中找到有用数据是一个经久不衰的课题[1] [2] [3]。本文利用协同过滤推荐系统的理念和方法，采用 UFTB 算法从用户看过的污染源信息及其信息类型入手，对用户看过的污染源信息类型与评分数据进行分析。在建立分析污染源信息推荐模型中，为防止过度优化，采取剔除用户非喜好类型污染源信息，得到优化缺省值预测矩阵，将相似度数据带入推荐公式，得出相应的结果。

2. 建模思路

模型的基本假设如下：

用户对污染源信息的评分不受已有评分影响；用户在短时间的兴趣是不会改变的；用户感兴趣的污染源信息类型仅与用户评分高的污染源信息类型相同；年龄相似，职业相仿的人兴趣相同；年龄对观看污染源信息类型的影响度大于职业；年龄差相同的情况下，年龄越大，两个用户的相似度越高。

模型的符号说明如下表 1：

要建立分析用户喜好的数学模型，根据协同过滤算法模型，首先分析 u_item 表与 R_{ij} 矩阵，通过计算修正后的余弦相似度，对缺省值进行预测，得到经缺省值预测补全的 R_{ij} 矩阵[2]，综合预测污染源信息间相似度。为防止过度优化，再根据用户喜好类型分析结果，将用户非喜好污染源信息类型从经缺省值预

测补全的 R_{li} 矩阵中剔除(即相应元素置零)，得到优化后的 R_{li} 矩阵。再使用修正后的余弦相似度算法获得用户间相似度矩阵，使用 TOP-*N* 算法，获取与目标用户相似度最高的 *N* 个用户，利用该 *N* 个用户的打分记录，获得目标用户喜好的 TOP-*N* 污染源信息编号，即推荐完毕。对此本文从四个步骤进行回答：

Table 1. Model symbol description
表 1. 模型符号说明

R_{ij}	用户 *i* 对项 *j* 的评分
sim_c	两类污染源信息间类型相似度
sim_{ij}	两类污染源信息评分相似度
$R_{c,i}$	用户 *c* 对污染源信息 *i* 的评分
$sim(TI,n)$	目标项 *TI* 与其最近邻居 *n* 之间的相似度
$\overline{R_i}$	用户 *i* 对所有污染源信息的平均打分
$sim(i,j)$	用户 *i* 和 *j* 的相似度
$F_u(i)$	基于 UFTB 算法对用户 *u* 的第 *i* 个污染源信息的评分
$F(i\|x\&y)$	未评分污染源信息 *i* 所获评测分值(用户喜好的污染源信息类型中)
$P_{u,TI}$	用户对项 *TI* 的预测评分

步骤一：读取 u_item 表，对任意两行数据向量化，并求两向量夹角余弦值，结果记为 sim_c， sim_c 值越高，代表两部污染源信息相似度越大；对用户打分矩阵 R_{ij} 利用修正后的余弦相似度算法，计算污染源信息间相似度，结果记为 sim_{li}， sim_{li} 越大代表两部污染源信息相似度越大。

步骤二：分析 sim_i 与 sim_c，将两种相似度加权平均得到预测相似度 *sim*，然后对缺省值进行预测，得到经缺省值预测补全的 R_{ij} 矩阵。

步骤三：为防止过度优化，根据用户喜好类型分析结果，将用户非喜好污染源信息类型从经缺省值预测补全的 R_{li} 矩阵中剔除(即相应元素置零)，得到优化后的 R_{li} 矩阵。

步骤四：使用修正的余弦相似度算法对 R_{li} 矩阵计算获得用户间的相似度矩阵，使用 TOP-*N* 算法，获取与目标用户相似度最高的 *N* 个用户，利用该 *N* 个用户的打分记录，获得目标用户喜好的 TOP-*N* 污染源信息编号。

3. 污染源信息协同过滤算法模型的建立

污染源信息相似度的计算如下：

读取 u_item 表，将 1682 行数据任取两行数据并向量化得 $\boldsymbol{R}_i$ 与 $\boldsymbol{R}_j$，则两向量代表两部污染源信息所属类型，根据协同过滤算法，两个污染源信息类型间的相似度 sim_c 表达式如下，

$$sim_c = \cos\left(\boldsymbol{R}_i, \boldsymbol{R}_j\right) = \frac{\boldsymbol{R}_i \cdot \boldsymbol{R}_j}{\left\|\boldsymbol{R}_i\right\|\left\|\boldsymbol{R}_j\right\|}$$

读取 R_{li} 矩阵，设对项 *i* 和项 *j* 两者共同评分过的用户集合用 U_{ii} 表示，U_i 和 U_j 分别表示对项 *i* 和项 *j* 评分过的用户集合[4]，则项 *i* 与项 *j* 之间的相似性 sim_{ij} 表达式如下，

$$sim_{ij} = \frac{\sum_{c\in U_{ij}}\left(R_{c,i}-\overline{R_i}\right)\left(R_{c,j}-\overline{R_j}\right)}{\sqrt{\sum_{c\in U_i}\left(R_{c,i}-\overline{R_i}\right)^2}\sqrt{\sum_{c\in U_j}\left(R_{c,j}-\overline{R_j}\right)^2}}$$

$R_{c,i}$ 表示用户 *c* 给予项 *i* 的评分，$R_{c,j}$ 表示用户 *c* 给予项 *j* 的评分，$\overline{R_c}$ 表示用户 *c* 给予所有项的平均

评分值。

综合分析 sim_c 与 sim_{ij} 可知， sim_{ij} 对 sim 值影响因子 a 较大，而 sim_c 对 sim 值影响因子 b 较小，经分析取 $a=0.8$， $b=0.2$。即 $sim=0.8sim_{ij}+0.2sim_c$。

设目标项 TI 的最近用户集合用 $NN_{TI}=\{NN_1,NN_2,\cdots,NN_K\}$ 表示[5]，则用户对项 TI 的预测评分 $P_{u,TI}$ 可以借助用户 u 对最近邻居集合 NN_{TI} 中项的评分得到[2]，公式如下：

$$P_{u,TI}=\overline{R_{TI}}+\frac{\sum_{n\in NN_{TI}} sim(TI,n)*\left(R_{u,n}-\overline{R_n}\right)}{\sum_{n\in NN_{TI}}\left(\left|sim(TI,n)\right|\right)}$$

$sim(TI,n)$ 表示目标项 TI 与最近用户 n 之间的相似性， $R_{u,n}$ 表示用户 u 对项 n 的评分。 $\overline{R_{TI}}$ 和 $\overline{R_n}$ 分别表示对项 TI 及项 n 的平均评分值。

运用预测后的 sim 值补全原 R_{ij} 矩阵，得经缺省值预测补全的 R_{ij} 矩阵。图 1 为原 R_{ij} 矩阵，图 2 为经缺省值预测补全的 R_{ij} 矩阵。

Variables - history

rate × aaa × history ×

history <943×1682 double>

	1	2	3	4	5	6	7	8	9	10
1	5	3	4	3	3	5	4	1	5	3
2	4	0	0	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	4	3	0	0	0	0	0	0	0	0
6	4	0	0	0	0	0	2	4	4	0
7	0	0	0	5	0	0	5	5	5	4
8	0	0	0	0	0	0	3	0	0	0
9	0	0	0	0	0	5	4	0	0	0
10	4	0	0	4	0	0	4	0	4	0

Figure 1. Original R_{ij} matrix (part)

图 1. 原始 R_{ij} 矩阵(部分)

Variables - bbb

bbb ×

bbb <943×1682 double>

	1	2	3	4	5	6	7	8	9	10
1	4.3159	2.6770	3.3084	3.6593	2.8212	4.3213	4.3385	3.8903	4.4870	3.9010
2	4.1303	2.9607	21.0154	3.7308	3.3080	3.7363	3.8801	4.2450	4.0908	3.3264
3	2.6817	2.6931	2.3062	2.9322	2.7326	3.1157	3.1715	3.2856	3.3498	3.1345
4	4.4100	3.8631	4.0786	4.2687	4.0710	4.2858	4.3592	4.7739	4.1463	5.0804
5	3.8520	2.9489	1.0581	3.2093	2.5398	2.3301	4.0494	3.4690	1.9608	3.4458
6	3.6605	4.2499	3.1272	3.3805	2.0532	3.3987	3.5210	3.9166	3.7959	3.8187
7	4.2631	3.7209	2.9203	4.2358	3.7606	4.0254	4.3556	4.5802	4.4562	4.2869
8	4.1724	3.2005	4.0872	3.8054	3.1434	3.9762	4.1424	4.3127	4.3941	4.0101
9	4.2669	5.2896	3.6052	4.2020	2.4583	4.3335	3.9419	4.1542	4.1676	3.9590
10	4.1593	3.1664	3.3865	3.8978	4.3609	3.9921	4.1363	4.3638	4.2405	4.1893

Figure 2. Default value prediction completion matrix (partial)

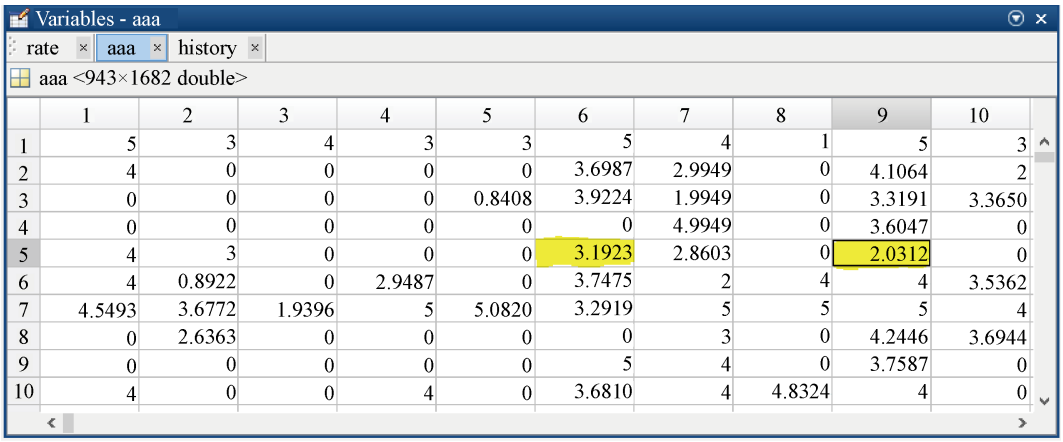
图 2. 缺省值预测补全矩阵(部分)

考虑到这种缺省值预测算法的过优化问题。即如果对每个未评分的污染源信息进行缺省值预测的话，用户评分表的稀疏程度将会是100%。针对这样的评分表，基于用户协同过滤推荐算法得到的最近邻将会和基于原有用户评分表的计算结果有着十分大的差别，甚至是完全相反的。我们可以假设用户原有的污染源信息评分表为用户喜好的真实情况，而这时计算得到的最近邻将会产生较大的反差，即过优化问题。因此，在对污染源信息进行缺省值处理时，我们应对污染源信息的相似度设置较高的阈值。只有高于设定阈值的相似度近邻才可被认可。选用Top N方法时，我们也采用了较小的N值。即只取预测分最高前几名作为推荐。目的是确保Null值处理后保证用户最近邻计算的可信度。经查阅文献可知，阈值取0.24时效果比较好。

图3为优化缺省值预测补全矩阵。使用修正后的余弦相似度算法对R_{ij}进行计算获得用户间的相似度矩阵，设用户i与用户j共同评分的污染源信息集合用U_{ij}来表示。U_i，U_j分别表示用户i与j评过分的污染源信息的集合。则用户i与用户j的相似度$sim(i,j)$为

$$sim(i,j)=\frac{\sum_{c\in U_{ij}}\left(R_{c,i}-\overline{R_i}\right)\left(R_{c,j}-\overline{R_j}\right)}{\sqrt{\sum_{c\in U_i}\left(R_{c,i}-\overline{R_i}\right)^2}\sqrt{\sum_{c\in U_j}\left(R_{c,j}-\overline{R_j}\right)^2}}$$

其中$\overline{R_i}$与$\overline{R_j}$分别表示用户i和用户j对所有污染源信息评分的平均值。



	1	2	3	4	5	6	7	8	9	10
1	5	3	4	3	3	5	4	1	5	3
2	4	0	0	0	0	3.6987	2.9949	0	4.1064	2
3	0	0	0	0	0.8408	3.9224	1.9949	0	3.3191	3.3650
4	0	0	0	0	0	0	4.9949	0	3.6047	0
5	4	3	0	0	0	3.1923	2.8603	0	2.0312	0
6	4	0.8922	0	2.9487	0	3.7475	2	4	4	3.5362
7	4.5493	3.6772	1.9396	5	5.0820	3.2919	5	5	5	4
8	0	2.6363	0	0	0	0	3	0	4.2446	3.6944
9	0	0	0	0	0	5	4	0	3.7587	0
10	4	0	0	4	0	3.6810	4	4.8324	4	0

Figure 3. Optimize the default value prediction completion matrix (partial)
图3. 优化缺省值预测补全矩阵(部分)

通过以上步骤，我们可以得到所有用户的近邻集合。设用户i的近邻集合为N_i，可以得到针对特定污染源信息a，用户i的预测评分为

$$R_{i,a}=\overline{R_i}+\frac{\sum_{j\in N_i\cap R_{j,a}\neq null}sim(i,j)*\left(R_{j,a}-\overline{R_j}\right)}{\sum_{j\in N_i\cap R_{j,a}\neq null}\left(\left|sim(i,j)\right|\right)}$$

其中$\overline{R_i}$为用户i对所有污染源信息的平均评分值。

得出用户针对未观看过污染源信息的预测评分后，再使用TOP-N算法，获得目标用户预测评分最好的N个污染源信息编号，即为目标用户喜好的TOP-N污染源信息编号。

此时考虑问题一中UFTB算法[2]，即

$$F_u(i)=F\left(i\left|x\&y\right.\right)F(x\&y)$$

其中$F_u(i)$是基于UFTB算法对用户u的第i个污染源信息的评测评分。$F\left(i\left|x\&y\right.\right)$表示在用户喜好的污

染源信息类型中，未评分污染源信息 i 所获得的评测分值。$F(x\&y)$ 中的 x 为用户对某类污染源信息的评分高低。y 表示用户对这类污染源信息的评分个数。$F(x\&y)$ 可表示为 $F(x\&y)=1$，当 x 大于 $\bar{x}$，且 y 大于 $\bar{y}$。其中 $\bar{x}$ 表示用户对所有污染源信息类型的平均评分值。$\bar{y}$ 表示用户对所有污染源信息类型的平均评分值个数。即表明如果用户不喜欢某些类型的污染源信息，则该类型的污染源信息所在列 $R_{\cdot i}$ 均为零(图 4)。

Variables - rate

rate × aaa × history ×

rate <943×1682 double>

	1	2	3	4	5	6	7	8	9	10
1	5	3	4	3	3	5	4	1	5	3
2	4	0	0	0	0	3.6987	2.9949	0	4.1064	2
3	0	0	0	0	0.8408	3.9224	1.9949	0	3.3191	3.3650
4	0	0	0	0	0	0	4.9949	0	3.6047	0
5	4	3	0	0	0	0	2.8603	0	0	0
6	4	0.8922	0	2.9487	0	3.7475	2	4	4	3.5362
7	4.5493	3.6772	1.9396	5	5.0820	3.2919	5	5	5	4
8	0	2.6363	0	0	0	0	3	0	4.2446	3.6944
9	0	0	0	0	0	5	4	0	3.7587	0
10	4	0	0	4	0	3.6810	4	4.8324	4	0

Figure 4. The optimized default value prediction completion matrix after non-user preferences are eliminated (partial)
图 4. 非用户喜好剔除后的优化缺省值预测补全矩阵(部分)

本文使用某数据网上的 943 个用户，1682 个污染源信息的数据进行模拟。同时，18 类主要污染源信息分别为：(1) 大气污染：I. 烟尘、II. 二氧化硫；(2) 水污染：III. 生活污水和其它耗氧废物、IV. 传染病菌和病毒、V. 植物营养剂——如氮和磷、VI. 有机化学合成剂-如杀虫剂、除锈剂和合成洗涤剂、VII. 来自工、矿、农业操作的其他矿物质和化学物质、VIII. 来自土地侵蚀的沉淀物、IX. 放射性物质、X. 热污染；(3) 土壤污染：XI. 化肥、农药、XII. 有机和无机污染物、XIII. 来自大气、水的污染物质迁移转化进入土壤的污染物质、XIV. 自然界或矿床周围元素富集形成的污染；(4) 其他污染源：XV. 光污染、XVI. 噪声、XVII. 电磁辐射、XVIII. 其他资料来源：http://mip.findlaw.cn/shpc/teshuqinquanjiufen/pcjf/1416533.html。下图 5、表 2 以 108 号用户为例说明获得推荐的污染源信息的过程。

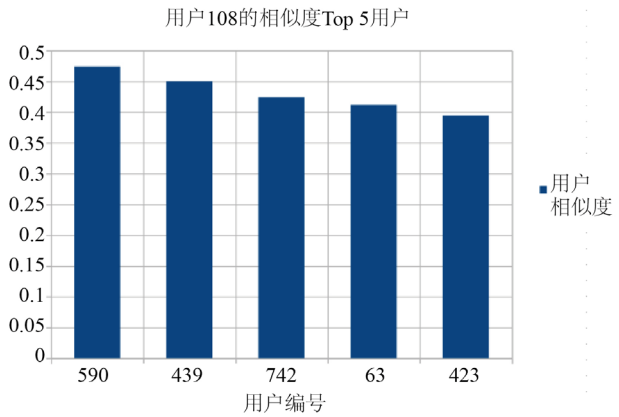


Figure 5. Top 5 users of similarity among users 108
图 5. 用户 108 的相似度 Top 5 用户

Table 2. Recommended pollution source information number table for specific users
表 2. 对特定用户的推荐污染源信息编号表

用户编号	推荐污染源信息编号(Top 1 - 5)
108	813, 316, 251, 285, 306
133	332, 1022, 1293, 302, 292
228	742, 705, 919, 318, 91
232	141, 543, 1126, 463, 1073
336	855, 653, 408, 114, 302
338	919, 12, 611, 127, 50
545	207, 408, 169, 526, 483
613	403, 470, 483, 474, 251
696	423, 200, 316, 372, 661
777	48, 169, 249, 261, 644

按照建立用户 i 对项 j 的评分矩阵 R_{ij}，预测补全缺省值，优化缺省值，剔除用户不喜欢的污染源信息类型这一顺序，我们得出了用户的近邻集合，求出近邻集合所有人对所有污染源信息的评分和，再排序取 TOP5 即为最后结果。

4. 结论

本文建立了基于协同过滤算法的数学模型，借助修正后的余弦相似度公式与 UTFB 模型，针对传统协同过滤推荐算法中的局限性，本算法解决了针对高稀疏度和低准确度问题。针对相似度度量标准这一问题，我们对用户评分表中的Null值进行了缺省值预测[2]。对于用户污染源信息喜好程度的修正与增强，在优化用户评分表的稀疏度同时，也增强了推荐系统的针对性。

参考文献

[1] Hou, C.Q., Zhu, L.C. and Zhang, W.G. (2009) A Collaborative Filtering Algorithm That Compresses Sparse User Scoring Matrix. *Xi'an University of Electronic Science and Technology Journal* (*Natural Science Edition*), **36**, 1-2.

[2] Wang, Z.W. (2011) Collaborative Filtering Recommendation Algorithm Based on user Preference Type. Master's Degree Thesis, East China Normal University, Shanghai, 21-25.

[3] Collaborative Filter Baidu Encyclopedia. http://baike.baidu.com/

[4] Wang, J. (2009) Personalized Recommendation System Design and Implementation of Library Sales Site Based on Associated Rules. Master's Degree Thesis, University of Electronic Science and Technology, Chengdu, 1-5.

[5] Zhuo, J.W. and Wei, Y.S. (2011) MATLAB Application in Mathematical Modeling. Beijing University of Aeronautics and Astronautics Press, Beijing, 104-108.