Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2020, 9(4), 232-236
Published Online November 2020 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2020.94026

Hans汉斯

污染源信息推荐的用户喜好模型研究

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11日；录用日期：2020年10月19日；发布日期：2020年10月26日

摘　要

由于污染给社会生活带来的诸多困扰和污染源的固有特性，作为污染源信息需求者的环境保护机构和个人，如何从大量污染源信息中找到自己关注的信息；同时，对于污染源信息提供者，怎样使自己的信息为广大用户所关注，是环保领域比较突出的矛盾和问题。本文通过建立基于年龄和职业的用户喜好模型，利用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，根据搜索出的与目标用户相似度最高的N位用户的喜好对目标用户进行污染源信息推荐。

关键词

协同过滤算法，UFTB，用户喜好模型，污染源信息

Study on User Preference Model about Recommendation of Pollution Source Information

Lina Wang

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

Received: Sep. 11th, 2020; accepted: Oct. 19th, 2020; published: Oct. 26th, 2020

Abstract

Because of the many problems brought by pollution to social life and the inherent characteristics of pollution sources, as the environmental protection institutions and individuals who demand information from pollution sources, how to find their own information from a large number of pollution

文章引用: 王丽娜. 污染源信息推荐的用户喜好模型研究[J]. 人工智能与机器人研究, 2020, 9(4): 232-236.
DOI: 10.12677/airr.2020.94026

source information, and how to make their own information for the vast number of users concerned about, are the more prominent contradictions and problems in the field of environmental protection. By establishing an age-based and occupation-based user preference model, 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, and recommend the pollution source information to the target user according to the preferences of the n-bit users with the highest similarity to the target user.

Keywords

Co-Filtering Algorithm, UFTB, Model of User Preferences, Pollution Source Information

1. 引言

污染给社会生活带来了非常多的困扰，同时由于污染源的固有特性，作为污染源信息需求者的环境保护机构和个人，如何从大量污染源信息(见本页脚注 [1])中找到自己关注的信息；同时，对于污染源信息提供者，如何让自己的信息为广大用户所关注，是环保领域比较棘手的矛盾和问题。协同过滤推荐(Collaborative Filtering recommendation)系统就是解决这一突出矛盾的有效工具[1] [2] [3] [4]。Hou *et al.* (2009)给出了压缩稀疏用户评分矩阵的协作过滤算法；Wang (2011)建立了基于用户首选项类型的协作筛选推荐算法；王丽娜等(2015)探讨了基于协同过滤算法的智能推荐系统。总的来看，现有的研究集中于各种算法的研发，对协同过滤算法尤其是应用到污染源推荐方面的用户喜好模型尚未涉猎。

本文通过建立基于年龄和职业的用户喜好模型，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，根据搜索出的与目标用户相似度最高的 N 位用户的喜好对目标用户进行污染源信息推荐。

2. 用户喜好模型的建模思路

用户喜好模型的基本假设、符号说明及建模思路如下表 1：

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。

[1]污染源：1) 大气污染：烟尘、二氧化硫；2) 水污染：生活污水和其它耗氧废物、传染病菌和病毒、植物营养剂-如氮和磷、有机化学合成剂-杀虫剂、除锈剂和合成洗涤剂、工、矿、农业操作的其他矿物质和化学物质、土地侵蚀的沉淀物、放射性物质、X.热污染；3) 土壤污染：化肥、农药、有机和无机污染物、大气、水的污染物质迁移转化进入土壤的污染物质、自然界或矿床周围元素富集形成的污染；4) 其他污染源：光污染、噪声、电磁辐射、其他资料来源：
http://mip.findlaw.cn/shpc/teshuqinquanjiufen/pcjf/1416533.htm

Table 1. Symbol description of user preference model
表 1. 用户喜好模型符号说明

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

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，$sim_c = \cos(\boldsymbol{R}_i, \boldsymbol{R}_j) = \dfrac{\boldsymbol{R}_i \cdot \boldsymbol{R}_j}{\|\boldsymbol{R}_i\|\|\boldsymbol{R}_j\|}$。读取 R_{ij} 矩阵，设对项 i 和项 j 两者共同评分过的用户集合用 U_{ij} 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[5]，则项 i 与项 j 之间的相似性 sim_{ij} 表达式如下，$sim_{ij} = \dfrac{\sum_{c\in U_{ij}}\left(R_{c,i}-\overline{R_i}\right)\left(R_{c,j}-\overline{R_j}\right)}{\sqrt{\sum_{c\in U_i}\left(R_{c,i}-\overline{R_i}\right)^2}\sqrt{\sum_{c\in U_j}\left(R_{c,j}-\overline{R_j}\right)^2}}$。$R_{c,i}$ 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[6]，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，公式如下：$P_{u,TI} = \overline{R_{TI}} + \dfrac{\sum_{n\in NN_{TI}} sim(TI,n)*\left(R_{u,n}-\overline{R_n}\right)}{\sum_{n\in NN_{TI}}\left(\left|sim(TI,n)\right|\right)}$。$sim(TI,n)$ 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。

考虑到缺省值预测算法的过优化问题，针对评分表，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。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，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，采用了较小的 N 值，即只取预测值最高的前几名作为推荐，目的是确保 Null 值在处理后保证用户最近邻计算的可信度。

经查阅可知，阈值取 0.24 时效果比较好。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_{ii} 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。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 $F_u(i)=F(i|x\&y)F(x\&y)$。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。$F(i|x\&y)$ 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 $\overline{x}$，且 y 大于 $\overline{y}$。其中 $\overline{x}$ 表示用户对所有污染源信息类型的平均评分值。$\overline{y}$ 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 $R_{\cdot j}$ 均为零。

3. 基于年龄和职业的用户喜好模型

事实上，许多新用户只有年龄与职业信息，因此只需要根据年龄和职业建立分析用户喜好的数学模型，分析与用户年龄和职业相似的其他用户的喜好即可。由假设可知，年龄差在相同的情况下，年龄越大，两个用户的相似度越高。对此，通过以下步骤进行推荐：一是读取用户 a 的信息，将现有用户按职业分类，并取出与其相同的职业；二是计算其与其他同职业用户的年龄差，用 $1-\frac{\text{年龄差}}{\text{该用户年龄}}$ 作为用户间的年龄相似度，选出与该用户相似度 Top5 的用户；三是找出 top5 用户看过的污染源信息，并用这些用户的打分记录预测用户 a 对这些污染源信息的打分，找出其中预测

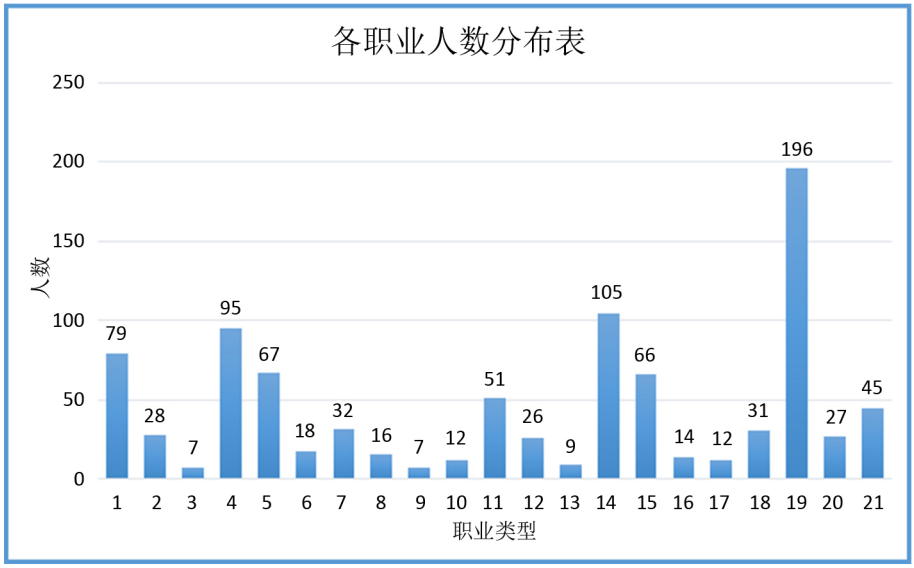

Figure 1. The distribution of the number of occupations
图 1. 各职业人数分布

得分 TOP5 的污染源信息，作为推荐。读取用户 a 年龄与职业信息，将所有用户按职业分类，取出与此用户相同职业的所有人。建立线性函数模型 $f(x)=1-\frac{\text{年龄差}}{\text{该用户年龄}}$。计算出该用户与其职业相同用户的年龄相似度，并选出相似度中的 TOP5 用户。找出 top5 看过的污染源信息，并用这些用户的打分记录预测用户 a 对这些污染源信息的打分，找出其中预测得分 top5 的污染源信息即 $R_{i,a}=\overline{R_i}+\frac{\sum_{j\in N_i\cap R_{j,a}\neq null} sim(i,j)*\left(R_{j,a}-\overline{R_j}\right)}{\sum_{j\in N_i\cap R_{j,a}\neq null}\left(\left|sim(i,j)\right|\right)}$，$\overline{R_i}$ 为用户 i 对所有污染源信息的平均评分值。模型结果如下图 1~2：

	A	B	C	D	E
1	25	35	30	50	22
2	engineer	administrator	programmer	educator	salesman
3	114	48	8	603	1012
4	919	465	306	8	329
5	408	924	91	217	90
6	270	947	169	408	312
7	50	792	175	686	186

Figure 2. Forecast of recommended pollution source information based on age occupation
图 2. 基于年龄职业的推荐污染源信息预测

4. 结论

本文模型建立基于协同过滤算法，经查阅，基于关联规则挖掘[5]也是推荐系统的一个重要规则。关联规则是指数据库中各项之间存在潜在关系的规则[6]。而类似的规则还有很多，通过理论与实践应用，发现各规则之间重合度不高，各具特点和适用性，因此在实际应用中可以设计基于多种推荐方法的组合推荐系统来为污染源信息选取提供借鉴和参考。

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] (2014) 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.

[6] 王丽娜, 张学恒, 王伟晨. 基于协同过滤算法的智能推荐系统研究[J]. 辽宁工业大学学报: 社会科学版, 2015(17): 26.