Hans Journal of Data Mining 数据挖掘, 2020, 10(4), 277-281
Published Online October 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.104029

Hans汉斯

基于UTFB算法的污染源信息推荐

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9日；录用日期：2020年9月23日；发布日期：2020年10月10日

摘　要

由于污染给社会生活带来了诸多困扰，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；而对于污染源信息提供者，让自己的信息为广大用户所关注，也是一件非常困难的事情。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，UTFB算法，污染源信息推荐

Pollution Source Information Recommendation Based on UTFB Algorithm

Lina Wang

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

Received: Sep. 9th, 2020; accepted: Sep. 23rd, 2020; published: Oct. 10th, 2020

Abstract

Because pollution has brought a lot of trouble to social life, as well as the inherent characteristics of pollution sources, as the source of pollution information needs of environmental protection institutions and individuals, from a large number of pollution source information to find their own concern of the information is often not an easy thing. The recommendation system is the main tool to solve this contradiction. By establishing the model of analyzing user preferences, UFTB algo-

文章引用: 王丽娜. 基于 UTFB 算法的污染源信息推荐[J]. 数据挖掘, 2020, 10(4): 277-281.
DOI: 10.12677/hjdm.2020.104029

rithm 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, UTFB Algorithm, Pollution Source Information Recommendation

1. 引言

近年来，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需要借助信息系统[1] [2] [3]。本文通过采用 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数值并根据它们的喜好对目标用户进行污染源信息推荐。模型基本假设如下：

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。模型的符号说明如表 1。

Table 1. Model's signal specifications
表 1. 模型的符号说明

符号	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
$P_{u,TI}$	用户对项 TI 的预测评分

2. 基于 UFTB 算法的用户喜好模型

建立分析用户喜好的数学模型，应考虑用户看过污染源信息种类以及用户对其打分，若用户对某类污染源信息打分越高则说明用户喜欢该类污染源信息，对此本文从两个步骤进行：

步骤一：读取每位用户看过的污染源信息和所有污染源信息的分类，构造用户评分矩阵[1]。

步骤二：计算每位用户对各类型污染源信息的评分 x 以及全部类型的平均评分值 $\bar{x}$，计算用户对各个类型的评分个数 y 和全部类型的平均评分值个数 $\bar{y}$，比较后得出结果。

先建立一个 $m*n$ 的用户评分矩阵 A (m, n)，m 代表用户观看的污染源信息，n 代表用户。第 i 行第 j 列的元素 r，代表用户 i 对项 j 的评分，若 i 用户对项 j 无评分，记 $R_{ij}=0$。用户评分数据矩阵如表 2 所示。

Table 2. User scoring data matrix
表 2. 用户评分数据矩阵

	$Item_1$	……	$Item_k$	……	$Item_k$
$User_1$	R_{11}	……	R_{1k}	……	R_{1n}
……	……	……	……	……	……
$User_j$	R_{j1}	……	R_{jk}	……	R_{jn}
……	……	……	……	……	……
$User_m$	R_{m1}	……	R_{mk}	……	R_{mn}

采用 UFTB 算法[2]判断用户是否喜好某类污染源信息，即 $F(x\&y)=\begin{cases}1, & \text{当}x\text{大于}\bar{x}\text{, 且}y\text{大于}\bar{y}\\ 0, & \text{其他情况}\end{cases}$，其中 $F(x\&y)$ 中的 x 为用户对某类污染源信息的评分值，y 表示用户对此类污染源信息的评分个数，其中 $\bar{x}$ 表示用户对全部污染源信息类型的平均评分值。$\bar{y}$ 表示用户对全部污染源信息类型的平均评分值个数。若 $F(x\&y)=1$，则表明用户喜欢该类污染源信息。否则不喜欢或中立。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，将 108 号用户数据带入模型得到图 1 (用 excel 表示)，图中用户平均评分值高于总体平均线的污染源信息类型即为该用户喜好的污染源信息类型。

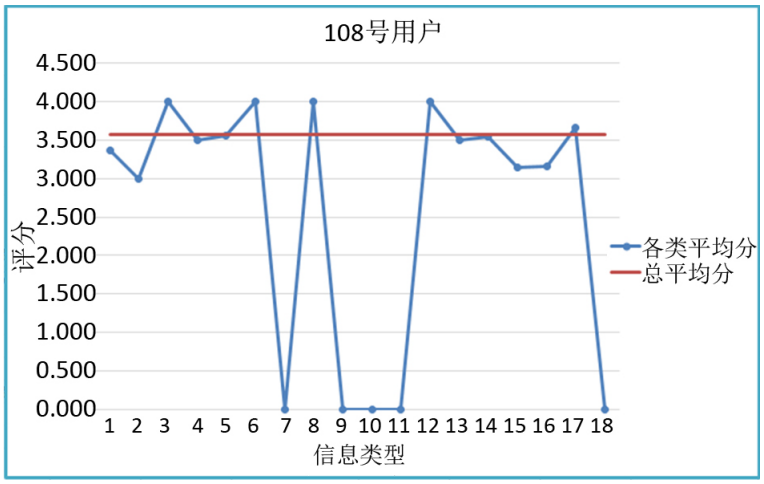

Figure 1. 108 user's average sub-line chart of various types of pollution source information
图 1. 108 号用户各类型污染源信息平均分折线图

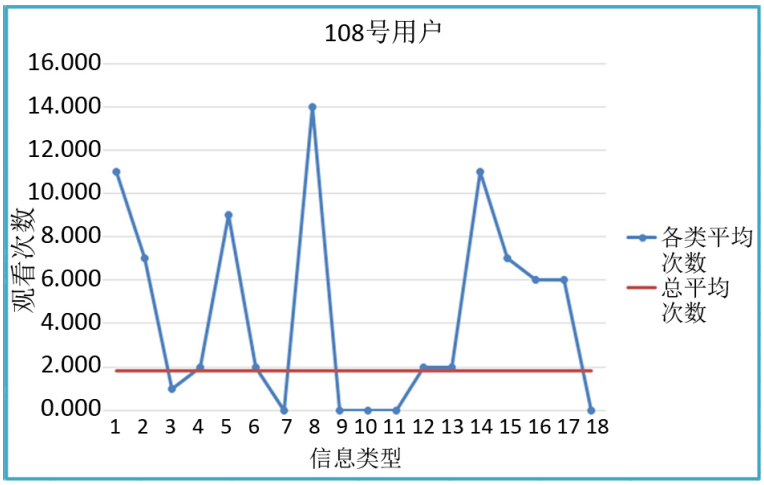

Figure 2. No. 108 user's line chart of all types of average number of viewing
图 2. 108 号用户各类型观看平均次数折线图

对图 1 与图 2 综合分析可知，108 号用户对 6，8，12，17 这四个类型的污染源信息评分高于总体平均分且观看次数高于总体平均次数，我们可以得出结论：108 号用户喜好 I、II、III、IV 这 4 种类型的污染源信息(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. 其他污染——资料来源：http://mip.findlaw.cn/shpc/teshuqinquanjiufen/pcjf/1416533.html)。类似 108 号用户的分析，对 10 位用户逐一分析，可得表 3。

Table 3. Specific users' preferences pollution source information types
表 3. 特定用户喜好污染源信息类型

用户编号	用户喜好污染源信息类型
108	I, II, III, IV
133	V, II, IV
228	X, V, II, IV
232	VI, I, II, III, VII, VIII, IX
336	X, V, III, I, VII, IX
338	II, XI, VII, IX
545	X, V, XII, VII, VIII, IX, IV
613	V, XIII, VII, VIII, IX, IV
696	V, I, II, XIV, XI, IV
777	XIII

对 UFTB 算法模型利用 MATLAB 编程进行求解，可以得到图 1、图 2。并且利用原始数据对表 2 进

行了验证，以 108 号用户为例，由模型可知用户的喜好类型为{I, II, III, IV}，查原始数据知 108 号用户所看污染源信息大部分属于此集合，且评分普遍较高。其他用户类似。由实际经验可知分析结果的可靠性。

3. 结论

本文模型的建立基于协同过滤算法[4]，基于关联规则挖掘也可以实现污染源信息的推荐，常用的实现算法为 Apriori 算法[5]。通过 Apriori 算法，得到如下结果，见表 4。

Table 4. Recommendation of pollution source information based on association rules
表 4. 基于关联规则的污染源信息推荐

用户编号	推荐污染源信息编号
108	50, 56, 174
133	56, 98, 172, 174, 181
228	79, 98, 174, 258, 300
232	1, 50, 56, 174, 181, 204
336	50, 181, 204
338	50, 100, 174, 181, 258, 294, 300
545	1, 100, 286, 313
613	50, 174, 204
696	50, 98, 100, 174, 268, 286, 300
777	50, 98, 100, 181

分析基于协同过滤与基于关联规则的推荐污染源信息编号，可以发现重合度并不高，这是因为两种推荐规则依据不同，所以最后在结果方面会有所不同，但都具有实用性。所以各环保当局和个人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实现对污染相关信息的取舍，指导具体的环境管理实践和进行相关环保政策制定。

参考文献

[1] Hou, C., Zhu, L. and Zhang, W. (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. (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 (2014). 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 of China, Chengdu, 1-5.

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