基于近邻传播聚类的多源异构数据信息个性化推送方法

谢梦怡 [1]
XIE Mengyi

摘 要 在处理多源异构数据、大规模数据时，传统的推送方法倾向于根据整体数据趋势和热门程度进行推送，未能根据具体数据特征和用户行为进行细致区分，因此推送内容可能过于泛泛，缺乏针对性。为此，提出一种基于近邻传播聚类的多源异构数据信息个性化推送方法。使用网络爬虫技术，获取用户在网络上的信息浏览行为数据，对这些数据进行分析，构建用户偏好模型。为了提升数据质量，采用空间滤波法对多源异构数据信息进行去噪处理。通过对数据进行标准化处理，消除数据的多源异构特性，使其具有可比性。在此基础上，运用主成分分析法进一步提取多源异构数据信息的核心特征。利用邻近传播聚类算法，对数据特征和用户偏好进行聚类分析，将具有相似性的数据或用户偏好归为一类。基于这些聚类结果，实现个性化的内容推送。经过实验验证，所设计的信息推送方法展现出极高的适配度和点击率，均超过 95%。这一成果充分证明了所提出的方法能够实现对多源异构数据信息的个性化精准推送，为用户提供更加精准、贴心的信息服务。

关键词 近邻传播聚类；多源异构数据信息；个性化推送；空间滤波法；主成分分析法

doi：10.3969/j.issn.1672-9528.2024.07.035

0 引言

随着信息技术的迅猛发展和大数据时代的来临，多源异构数据信息呈现迅猛增长的趋势。这些数据不仅源自传统的数据库、文件系统等结构化数据源，还涵盖了社交媒体、物联网、移动设备等生成的半结构化或非结构化数据。多源异构数据的融合与利用，对于提升信息处理的效率和准确性，以及满足用户个性化的信息需求，具有极其重要的意义。在处理多源异构数据信息的过程中，个性化推送技术发挥着举足轻重的作用。个性化推送技术旨在根据用户的兴趣、偏好和行为模式，从海量的数据中筛选出对用户有价值的信息，并精准地推送给用户。这不仅能够提升用户的信息获取效率，还能优化用户的使用体验，提高用户满意度。然而，多源异构数据的复杂性和多样性给个性化推送带来了诸多挑战。如何有效地整合不同来源、不同格式的数据，如何准确地捕捉用户的兴趣和需求，以及如何设计高效的推送算法，都是当前亟待解决的问题。

在这种情况下，王南提出了基于云计算的信息个性化推送方法，通过云计算提取数据特征，向用户推送数据信息 [1]；刘世雄等人提出了基于兴趣图谱的信息个性化推送方法，该方法是以兴趣图谱为技术支撑，依据用户兴趣推送数据信息 [2]；王金威提出了基于大数据技术的信息个性化推送方法，利用大数据对数据处理分析，为用户提供信息个性化推送服务 [3]；张莉提出了基于用户画像的信息个性化推送思路，根据用户喜好生成用户画像，推送数据信息 [4]。这是目前信息个性化推送实践中广泛应用的方法。然而，现行方法应用于多源异构数据信息个性化推荐中存在一些问题，即未能充分根据具体数据特征和用户行为进行细致区分，仍然倾向于根据整体数据趋势和热门程度进行推送。因此，推送内容可能过于泛泛，缺乏针对性和个性化，无法达到预期的个性化推送效果。为此，本研究提出一种基于近邻传播聚类的多源异构数据信息个性化推送方法。

1 用户行为数据获取及偏好模型建立

通过爬取用户在网络上的信息浏览行为数据，可以全面、细致地了解用户的浏览习惯、兴趣点以及信息需求。这些数据反映了用户在各个平台、各个时间段的浏览轨迹，是分析用户行为的重要依据。基于获取到的用户浏览行为数据，通过数据分析技术，可以构建出用户的偏好模型。这个模型是对用户兴趣、需求和偏好的抽象化表达。有了用户偏好模型，可以更加精准地定位目标用户群体，为后续的个性化服务提供基础。详细过程描述如下。

采用网络爬虫技术对用户网络多源异构数据信息浏览行

1. 泉州信息工程学院 福建泉州 362000

为信息进行爬取，爬取数据用公式表示为：

$$x=\sum_{e\in R}(a-z)/e(v) \tag{1}$$

式中：x 表示网络爬虫爬取的用户行为数据；e 表示用户行为数据源编号；R 表示网络爬虫爬取次数；a 表示用户行为数据属性；z 表示用户行为数据类别；v 表示爬取的用户行为数据量 [5]。按照用户 ID、多源异构数据信息 ID、类目数据等，将网络爬虫爬取的用户行为数据进行分类，生成用户行为数据表并录入数据库，具体如表 1 所示。

表 1 用户行为数据信息表

字段	说明	类型	是否为空
用户 ID	User_id	Char（10）	N
多源异构数据信息	Data_information	Int（10）	N
用户浏览时长	Browsing_duration	Char（20）	N
信息点击次数	Click_count	Char（10）	Y
多源异构数据信息类目	Data_information categories	Char（30）	N
用户浏览深度	Browsing_Depth	Int（10）	Y
其他	other	Char（10）	N

根据多源异构数据信息网站运行更新用户行为数据信息表中的数据，其用公式表示为：

$$V_i=\frac{1}{\|A_i\|}\sum_{k\in r}(x) \tag{2}$$

式中：V_i 表示更新后的第 i 个用户行为数据；A_i 表示第 i 个用户行为数据归类；k 表示更新数据；r 表示用户行为数据权重 [6]。根据用户行为数据统计结果，建立用户行为模型，用公式表示为：

$$C(x)=\frac{\sum d\times b(V_i)}{\sqrt{d^2}} \tag{3}$$

式中：$C(x)$ 表示用户行为模型；d 表示用户行为数据稀疏度；b 表示用户行为逻辑因子 [7]。根据用户行为，计算用户对某一类多源异构数据信息的偏好程度，建立用户偏好模型：

$$y=w+\sum C(x)\upsilon \tag{4}$$

式中：y 表示用户偏好模型，即用户偏好度；w 表示多源异构数据信息的线性回归表达式；υ 表示多源异构数据信息评估分值。

2 多源异构数据去噪与特征提取

当用户行为数据获取及偏好模型建立后，如果直接根据用户偏好推送数据信息而忽略数据信息的异构性问题，可能会导致推送内容与用户实际需求和兴趣不匹配，降低推送的准确性。因此，有必要对多源异构数据信息进行去噪处理。通过标准化数据处理，消除数据的异构特性，使其可比性增强。在这一基础上，应用主成分分析法来进一步提取多源异构数据信息的核心特征 [8]。

由于多源数据可能存在噪声，可采用空间滤波法进行去噪处理，统一采集数据库中的多源异构数据信息，建立数据空间并对数据进行排列，统计数据空间内数据特征点的密度。计算公式为：

$$\kappa=y\sum\exp\left(1-\left\|\frac{o}{2}\right\|\right)^2 \tag{5}$$

式中：κ 表示多源异构数据信息特征点密度；O 表示多源异构数据信息在空间中的维度 [9]。根据数据在空间中的特征点密度，设定一个阈值，以阈值作为界限，将多源异构数据划分为两个数据群，其用公式表示为：

$$B=\begin{cases}Y,\kappa\cdot\varepsilon\le\varpi\\J,\kappa\cdot\varepsilon\ge\varpi\end{cases} \tag{6}$$

式中：B 表示多源异构数据空间划分结果；Y 表示有效数据；J 表示噪声数据；ε 表示多源异构数据特征点二维曲线；ϖ 表示阈值 [10]。将分类结果中的噪声数据去除，保留有效数据，实现对多源异构数据信息滤波。多源异构数据的异构性导致数据之间量纲不同，因此采用归一化法对多源异构数据标准化处理，用公式表示为：

$$\overline{Y}=\frac{Y-\eta(Y)}{B\xi(Y)} \tag{7}$$

式中：$\overline{Y}$ 表示标准化处理后的多源异构数据；$\eta(Y)$ 表示多源异构数据特征量均值；$\xi(Y)$ 表示多源异构数据特征量标准差 [11]。通过对数据信息标准化处理，将其转换为同源类数据，在此基础上对数据主成分进行分析，提取数据信息特征，假设每个多源异构数据信息主要成分值为 $\overline{Y}_j\left(j=1,2,...,l\right)$，将其填入矩阵中并乘以置换矩阵后，求和再平均就可以得到数据的主特征值，用来表征多源异构数据信息的整体特征，其用公式表示为：

$$g=\frac{1}{N}\sum_{j=1}\left\{\left|\overline{Y}_j\cdot S-u\right|\left|\overline{Y}_j\cdot S-u\right|^T\right\} \tag{8}$$

式中：g 表示多源异构数据信息主特征值；N 表示多源异构数据信息样本数量；S 表示成分矩阵；u 表示根据评估得到的特征权重值；T 表示置换矩阵 [12]。通过以上计算，得到数据库中所有多源异构数据信息主特征值，为后续基于邻近传播距离的信息个性化推送奠定基础。

3 基于邻近传播聚类的信息个性化推送

在多源异构数据去噪及特征提取后，利用邻近传播聚类算法对数据特征和用户偏好进行深入的聚类分析。该算法能够将具有相似性的数据或用户偏好有效地归为一类，从而为

后续的个性化内容推送提供坚实的基础。基于这些精准的聚类结果，能够精确地实现个性化的内容推送服务。具体来说，采用邻近传播聚类算法对用户偏好 y 与多源异构数据信息特征 g 进行精确匹配。通过这一匹配过程，能够从庞大的数据集中筛选出那些真正满足用户偏好的多源异构数据信息。随后，根据这些筛选结果生成推送列表，并进行个性化的内容推送。

邻近传播聚类算法是根据两个向量之间的相似度进行聚类分析，具体聚类过程如下。

步骤 1：通过用户多源异构数据信息浏览偏好与多源异构数据信息特征的欧式距离计算，获取聚类中心与多源异构数据信息之间的基本相似度，计算公式为：

$$\delta(y,g)=-\|y-g\|^2 \tag{9}$$

式中：$\delta(y,g)$ 表示用户浏览行为与多源异构数据信息特征的基本相似度；$-\|y-g\|^2$ 表示用户浏览行为与多源异构数据信息特征的欧式距离。基本相似度越高，则用户偏好作为邻近传播聚类中心的可能性越高，因此根据基本相似度值确定邻近传播聚类中心。

步骤 2：基于闵可夫斯基度量，计算聚类中心所对应的用户偏好与多源异构数据信息的相似度，其计算公式为：

$$d(y,g)=\max\left\|\delta(y,g)-g\right\| \tag{10}$$

式中：$d(y,g)$ 表示聚类中心所对应的用户偏好与多源异构数据信息的相似度。

步骤 3：响应度计算。利用响应度，将多源异构数据信息从一点传递到另一点。

步骤 4：可用性计算。并不是所有相似度越高，就表示该多源异构数据信息越符合用户需求，因此传播到邻近点后对多源异构数据信息的可用性进行计算，其用公式表示为：

$$f=\max\left\{0,H_{d(y,g)}\right\} \tag{11}$$

式中：f 表示多源异构数据信息可用性；$H_{d(y,g)}$ 表示传播到邻近点上的多源异构数据信息自我响应度。可用性越高，则表示多源异构数据信息与用户偏好的匹配度越高，将其可用性最高的信息传播到下一个点上。

步骤 5：聚类收敛检验。检验当前是否达到收敛条件，如果满足，则输出传播到当前点的多源异构数据信息，将其作为推送信息推送给用户。如果没有满足收敛条件，则重复步骤 2 ～步骤 4，直至满足收敛条件，以此实现基于邻近传播聚类的多源异构数据信息个性化推送。

4 实验结果与分析

4.1 实验数据与环境

为了验证本文提出的基于近邻传播聚类的多源异构数据信息个性化推送方法的服务效果，采集了某网络平台的多源异构数据信息作为实验测试样本，这些数据包含文本数据、图片数据、视频等多种类型的数据，共计 12 642 条数据信息。在进行实验前，搭建了本次实验的实验环境：操作系统为 Windows 10，配置了 32 GB 内存的 Intel Core i9 处理器。实验涉及了 50 名用户的数据信息，收集这些用户在网络平台上的浏览行为数据，包括点击和浏览记录。利用这些用户行为数据，建立了用户偏好模型，并据此设计了个性化推送算法，结合多源异构数据信息进行推荐。

4.2 实验指标

在评估多源异构数据信息个性化推送方法的性能时，选择适配度和推送信息点击率作为评估指标。适配度指推送信息与用户浏览需求的匹配程度，在信息个性化推送中，实现高准确性意味着方法具有良好的适配度。适配度的评估基于用户对信息个性化推送服务的评分进行计算，其计算公式为：

$$E=\frac{1}{n}\sum_{n=1}\rho_n \tag{12}$$

式中：E 表示信息个性化推送适配度；n 表示用户数量；ρ_n 表示第 n 个用户对信息个性化推送服务评分。适配度越高，则表示信息推送准确性越高。推荐信息用户点击率也是评价推送方法、模型或者算法精度的重要指标，其计算公式为：

$$P=\frac{m}{n}\times 100\% \tag{13}$$

式中：P 表示推送信息用户点击率；m 表示推送信息后用户点击浏览次数。只有在推荐信息符合用户需求时，用户才会点击进入浏览，因此点击率越高，说明信息推送精度越高。

4.3 实验结果与分析

实验以文献 [1] 方法和文献 [2] 方法与本文方法对比，按照公式（12）、（13）计算三种方法信息个性化推荐适配度和用户点击率。统计结果如表 2、表 3 所示。

表 2 多源异构数据信息个性化推送适配度 /%

推送次数 / 次	本文方法	文献 [1] 方法	文献 [2] 方法
1	98.65	75.61	80.36
2	97.15	72.62	80.25
3	99.56	73.45	80.42
4	98.52	75.42	80.62
5	98.47	71.06	80.24
6	98.26	71.25	80.12
7	98.35	71.62	80.36
8	99.95	71.52	80.24
9	99.36	71.46	80.06
10	99.14	71.03	80.07

表 3 推送信息用户点击率 /%

实验组别	本文方法	文献 [1] 方法	文献 [2] 方法
第一组	98.62	80.16	85.62
第二组	97.56	82.46	86.56
第三组	96.25	82.64	84.62
第四组	95.46	83.65	86.59
第五组	98.46	85.61	85.75
第六组	97.16	82.46	85.52
第七组	98.62	80.16	85.16
第八组	96.55	80.52	85.26
第九组	95.64	80.24	85.55
第十组	96.53	80.13	82.16

从表 2 中数据可以看出，以上 3 种方法的信息个性化推送与用户需求之间的适配度保持在 70% 以上。具体分析可知，本文提出的方法所实现的适配度高达 95% 以上，最高时甚至达到了 98.62%。与文献 [1] 和文献 [2] 中的方法相比，本文提出的方法适配度分别高出将近 14% 和 13%。此外，从表 3 中的数据也可以观察到，个性化推送信息的用户点击率同样保持在 90% 以上的高水平，这进一步表明了本文方法推送的信息与用户需求之间的高度匹配。本文方法所实现的用户点击率不仅超过了 95%，而且显著高于文献 [1] 和文献 [2] 中的方法。具体分析可知，本文方法的优势在于通过使用网络爬虫技术获取用户在网络上的信息浏览行为数据，并进一步分析这些数据以构建用户偏好模型，可以更好地了解用户的兴趣和需求，从而实现更加精确的个性化内容推送。通过用户偏好模型的建立，能够更精准地匹配推送内容与用户需求，因此提高了推送信息的适配度和用户点击率。

为验证实验测试结果的有效性，运用不同方法对推送信息召回率进行计算，得到对应的结果如图 1 所示。

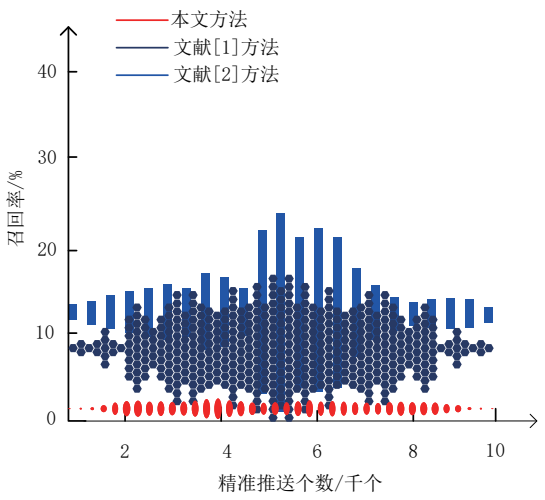

图 1 实验召回率比较

由图 1 结果可知，相比于两种文献方法，本文方法的召回率较低，说明本文的推送方法准确程度较高，能够从数据集中筛选出满足用户偏好的多源异构数据信息，生成推送列表，进行个性化推送。具体分析可知，本文方法的优势在于通过邻近传播聚类算法对数据特征和用户偏好进行聚类分析，将具有相似性的数据或用户偏好归为一类，可实现更加精准的个性化内容推荐，从而提升推送方法的准确性。

5 结语

在当今信息化社会，多源异构数据信息繁杂多样，如何高效提取对用户具有实际价值的信息并实现个性化精准推送，已成为信息科技领域亟待解决的关键问题。本研究提出的基于近邻传播聚类的多源异构数据信息个性化推送方法，在实际应用中表现出了卓越的性能和显著成效。该方法不仅能够有效整合不同来源、不同格式的数据信息，还能精准捕捉用户的个性化需求，实现高度个性化的内容推送。通过持续收集和分析用户的反馈和行为数据，不断优化推送策略，进一步提升用户体验和满意度。同时，该方法的灵活性和可扩展性也使其能够适应不同领域和场景的需求变化，展现出广阔的应用前景。然而，任何方法和技术的完善都是一个持续的过程。展望未来，将继续关注多源异构数据的发展趋势和个性化推送的新需求，不断优化和改进本文方法。同时，也将积极探索与其他先进技术的融合及应用，如深度学习、强化学习等，以期进一步提升推送的准确性和效率，为用户提供更加优质、个性化的信息服务。

参考文献：

[1] 王南 . 基于云计算的短视频媒体资源个性化推送方法 [J]. 兵工自动化 ,2024,43(2):16-22.

[2] 刘世雄 , 张俊 . 基于兴趣图谱的移动端个性化消息推送服务研究 [J]. 电脑知识与技术 ,2022,18(4):57-58.

[3] 王金成 . 基于大数据分析的高校云招聘信息个性化推送研究 [J]. 安徽电子信息职业技术学院学报 ,2022,21(4):25-31.

[4] 张莉 . 基于用户画像的常德地区休闲农业信息个性化推荐系统研究 [J]. 软件 ,2022,43(2):1-3.

[5] 李高扬 , 王宁 , 高若田 , 等 . 面向新型电力系统智能化提效的多源异构数据融合技术研究 [J/OL]. 电测与仪表 :1-8[2024-02-06].http://kns.cnki.net/kcms/detail/23.1202.TH.20240311.1147.002.html.

[6] 段昕汝 , 陈桂茸 , 姬伟峰 , 等 . 基于联邦学习的多源异构网络无数据融合方法 [J]. 空军工程大学学报 ,2024,25(1):90-97.

[7] 李阳 , 何文峰 , 黄伦春 . 一种设施普查中多源异构数据的处理方法 [J]. 城市勘测 ,2023(S1):181-184+204.
[8] 王彩霞 , 陶健 . 数据库中多源异构异常数据清洗方法 [J]. 通化师范学院学报 ,2023,44(12):54-60.
[9] 程雪婷 , 王玮茹 , 暴悦爽 , 等 . 基于联邦学习的多源异构数据安全融合方法 [J]. 通信技术 ,2023,56(10):1173-1183.
[10] 李坚 , 杨峰 , 吴佳 , 等 . 基于改进 FCM 的多源异构能源数据预处理与去噪 [J]. 微型电脑应用 ,2023,39(10):80-82+87.
[11] 杨桥桥 , 洪东彬 , 周扬 , 等 . 互联网背景下基于个性化推送的供电服务学习机制 [J]. 互联网周刊 ,2022(1):40-43.
[12] 李学威 , 孙滨 . 基于深度学习的自适应移动学习服务智能推荐研究 [J]. 信息与电脑 (理论版),2023,35(3):254-256.

【作者简介】

谢梦怡（1986—），女，福建泉州人，硕士，讲师，研究方向：云计算、大数据。

（收稿日期：2024-03-29）

（上接第 164 页）

5 结语

本文设计了一种基于椭圆曲线的 Diffie-Hellman 的零知识根密钥更新协议。在设备激活前，通过带有零知识性的密钥交换算法生成共享机密，再由密钥推导函数将共享机密与上一次根密钥作为原材料派生出新的根密钥，解决 LoRaWAN 的 OTAA 中根密钥不更新的问题。通过对协议的验证分析，证明该协议是安全的，能够抵御常见攻击。

参考文献：

[1] 黄长清 . 智慧武汉 [M]. 武汉 : 长江出版社 ,2012.
[2]XU Z, JIE N.A study on key LPWAN technologies[EB/OL].(2021-04-11)[2024-03-21].https://iopscience.iop.org/article/10.1088/1742-6596/1871/1/012011.
[3]LORA ALLIANCE TECHNICAL COMMITTEE.LoRaWAN-backend-interfaces-v1.0[EB/OL].(2017-10-11)[2024-03-21]. https://lora-alliance.org/wp-content/uploads/2020/11/lorawantm-backend-interfaces-v1.0.pdf.
[4]LORA ALLIANCE TECHNICAL COMMITTEE.LoRaWAN® specification v1.1[EB/OL].(2017-10-11)[2024-03-21].https://resources.lora-alliance.org/technical-specifications/lorawan-specification-v1-1.
[5]HAN J, WANG J.An enhanced key management scheme for LoRaWAN[C]//Security,Privacy,and Anonymity in Computation,Communication,and Storage.Cham: Springer, 2018: 407-416.
[6]ANATOLY K, SERGEY Z.Zero knowledge proof and ZK-SNARK for private blockchains[J].Journal of computer virology and hacking techniques,2023,19(3):443-449.
[7]SCHNORR C P.Efficient signature generation by smart cards[J]. Journal of cryptology,1991,4:161-174.
[8]ISMAIL B, NUNO P, MIKAEL G.Security risk analysis of LoRaWAN and future directions[J].Future internet, 2019, 11(1): 1-22.
[9]ILSUN Y, SOONHYUN K, GAURAV C, et al.An enhanced LoRaWAN security protocol for privacy preservation in IoT with a case study on a smart factory-enabled parking system[J]. Sensors,2018,18(6):1888.
[10] VICTOR R, RAIMIR H F, ALEX R.A secure and fault-tolerant architecture for LoRaWAN based on blockchain[C]//2019 3rd Cyber Security in Networking Conference. Piscataway: IEEE, 2019:35-41.
[11]CHEN X, LECH M, WANG L.Complete key management scheme for LoRaWAN v1.1[J].Sensors,2021,21(9):2962.
[12]TSAI K, CHEN L, LEU F, et al.Two-Stage high-efficiency encryption key update scheme for LoRaWAN based IoT environment[J]. Computers,materials& continua, 2022, 73(1): 547-562.
[13]BURROWS M, ABADI M, NEEDHAM R.A logic of authentication[J].ACM transactions on computer systems, 1990, 8(1):18-36.
[14]ARMANDO A, BASIN D, BOICHUT Y, et al.The AVISPA tool for the automated validation of internet security protocols and applications[C]//Computer Aided Verification.Berlin: Springer-Verlag, 2005:281-285.
[15]DOLEV D, YAO A.On the security of public key protocols[J]. IEEE transactions on information theory,1983,29(2):198-208.

【作者简介】

李心（1999—），女，四川眉山人，硕士，研究方向：无线传感器网络。

（收稿日期：2024-04-23）