

# 用于评价推荐系统的多样性指数的研究

孙琛恺, 安俊秀

(成都信息工程大学软件工程学院, 四川 成都 610225)

**摘要:**针对当今数据量的庞大导致用户获取所需信息困难以及推荐系统评价体系缺乏多样性评价指标的问题,提出基于三部图校准的 Herfindahl 多样性指数,通过该指标来量化推荐系统的多样性。首先,根据设定好的分类方式进行 URL 分类;进而设计形成“类别 URL 用户”的三部图;其次,对原本的 Herfindahl 指数进行改良,减少数量的差异对多样性的影响;最后,结合改良的 Herfindahl 多样性指数,得到推荐系统的多样性指数。多样性指数的出现有助于在评价推荐系统时,不仅关注推荐的准确与否,而且考虑推荐信息是否全面。实验表明,基于此实验提出的方法所得的改良后的 Herfindahl 指数可以对推荐系统类别受众多样性进行准确的量化。

**关键词:**计算机软件与理论;推荐系统;Herfindahl 指数;多样性指数;URL;三部图

**中图分类号:**TP393.027.2

**文献标志码:**A

**doi:**10.16836/j.cnki.jcuit.2021.03.002

## 0 引言

随着互联网的飞速发展以及用户的迅速增长,导致数据量的指数级增长。2019年,中国产生的数据量已达到9.8 ZB,而且还保持着快速增长,预计在2025年的数据量将达到48.6 ZB。互联网的海量信息可以方便用户的日常生活,但也带来信息过载和信息迷航的问题。如何从数据中提取到有价值的信息并更好地对用户提供服务,成为现在企业界以及学术界研究的重点。

目前,常见的推荐系统主要应用于搜索引擎(例如百度,Google等)以及平台的推荐机制(例如今日头条,抖音等)。Gema Bello-Orgaz等<sup>[1]</sup>从推特上获取数据,用多种聚类方法对酒庄的推特进行分析,最终发现采用精准营销策略的酒庄销量更好。陈兴喆等<sup>[2]</sup>利用Web日志挖掘技术,掌握客户的行为模式,最终实现针对用户个人的精准推荐,避免了“千人一面”的问题,有效提高了访问量。Yiqun Liu等<sup>[3]</sup>构建点击行为模型,使得到的文档不总是与用户的查询相关,但是用户点击的部分最有可能满足用户的需求,从点击率的角度实现了精准推荐。由此可见,合理地使用推荐算法对平台进行优化可以创造更大的价值。

推荐系统的核心是算法。其中,搜索引擎主要依赖于用户输入的关键字、语音识别以及拍照识别等方式输入信息,不同的用户输入相同的信息所得到的内容是一致的,没有办法实现个性化的推荐。现在用在平台上的推荐算法更多的是根据用户访问的内容、标

签以及其对应的点击、收藏、点赞、评论等行为构建模型,针对不同的用户向其推荐个性化内容。虽然这种方式让用户更加便捷地获取信息,但同时也影响到用户获取信息的机会。对于推荐系统的评价,更多的研究者将目光放在准确性的提升方面,但科学界对如何衡量推荐算法的多样性还没有达成共识。文献[4]对如何避免推荐算法的偏见性,同时又可以保证用户访问的个性化问题进行了探索,通过多重过滤算法对网页内容进行筛选。文献[5]对研究推荐系统多样性的困境进行了探索,提出在推荐系统中通过引入多样性指数的方式来量化推荐系统中的多样性。但在以上文献中,均没有实现推荐系统多样性的量化。

## 1 相关概念

### 1.1 部图

在大多数论文和资料中,出现较多的是二部图。二部图又称为二分图,设 $G=(V,E)$ 是一个无向图,如果顶点 $V$ 可分割为两个互不相交的子集 $A$ 和 $B$ ,并且图中的每条边 $(i,j)$ 所关联的两个顶点 $i$ 和 $j$ 分别属于这两个不同的顶点集( $i \in A, j \in B$ ),则称图 $G$ 为一个二部图,见图1。

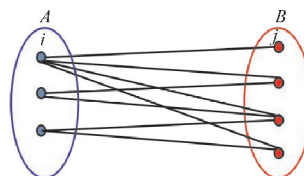


图1 二部图

图1的二部图可以记为  $G=(A,B,E)$ ,  $A$  和  $B$  为两个不相交的子集,  $E$  为  $A$  和  $B$  两个子集中边的集合。三部图则是在二部图的基础上增加了  $C$  集合, 并且通过某种联系, 将三个集合联系在一起, 假设在  $A$  集合与  $B$  集合之间找到中间联系  $C$  集合, 则三部图如图2所示。

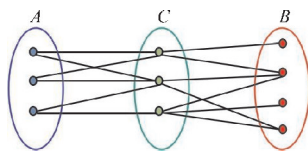


图2 三部图

现在的事物发展往往受到多方面的影响, 为更深入地研究多方面的特性, 研究者们做了以下尝试:

为改良图网络结构推荐算法的精确度, 王睿<sup>[6]</sup>建立了“用户项目内容”的三部图结构, 通过这种三部图能够有效地衡量3个因素之间的关系。A E Iordan<sup>[7]</sup>运用三部图解决了 Guarini 难题, 通过三部图的构造, 可以更加细化类与类之间的关系。戴瑾如<sup>[8]</sup>提出基于加权三部图模型的推荐算法, 通过资源整合综合考虑物质扩散算法和热量传导算法得到的资源值, 平衡推荐系统的精确性和多样性。

## 1.2 多样性指数

多样性对于确保复杂系统长期生存具有重要作用, 通常比较常见的多样性存在于生物领域、政治领域、科学领域等。为能够解释及量化这种多样性, 近百年来许多学者定义了各种各样的指标。例如王泓等<sup>[9]</sup>将信息熵指数用于语言风格分析。意大利统计与社会学家 Corrado Gini 在 1912 年提出基尼系数, 该指数作为经济学指标以衡量一个国家和地区的居民收入差距。赫芬达尔-赫希曼指数<sup>[10]</sup>通常用于测量产业市场集中度。常用的多样性指数还包括 Richness 指数(通常用于测量物种丰富度), Berger-Parker 指数(通常用于测量种族优势度)以及 1948 年提出的 Shannon 指数(衡量种群的多样性)。

多样性指数同样应用于经济方面, 而推荐平台的广泛使用(今日头条, 抖音等)让研究者对如何量化用户行为的多样性产生思考。对于如何衡量推荐算法的多样性, 现在还没有完全统一的标准, 大多数方法还是通过余弦相似度、欧氏距离以及逆皮尔逊系数来衡量。

## 2 URL 三部图

二部图是一个具有两个不相交的节点集形成的图, 文献[11]为了量化用户活动的多样性, 对在线音

乐平台的数据进行三部图的构造, 形成“类别音乐用户”的三元组。这种构造三部图的方式可以很好地反映顶部节点与底部节点间的联系, 从而为研究用户行为以及分析类别受众提供帮助。

本文将一个独立集合上的点与另一个独立集合的点联系起来, 定义为三元组  $B=(T, \perp, E)$ , 并根据搜狗搜索引擎日志中的信息构造三部图。凭借三部图的优势, 挖掘类别与用户的关系。首先, 将  $T$  定义为 URL 类别的集合, 定义为用户的集合,  $E \subseteq T \times \perp$  是 URL 类别与用户联系关系的集合。对于每个节点  $v \in T$ , 定义其相邻点集合  $N(v) = \{u \in \perp \mid (v, u) \in E\}$  并且通过类似的方式定义节点  $u \in \perp$ , 相邻点集合  $N(u) = \{v \in T \mid (u, v) \in E\}$ 。这个相邻点的集合大小称为度:  $d(u) = |N(u)|$ 。同样, 可以定义一个二部图来表示 URL 与类别之间的关系。在此基础上, 将两个二部图合并起来以分析用户活动的完整结构形成一个三部图  $T=(T, X, \perp, E_1, E_2)$ 。其中,  $T$  是 URL 的类型的集合,  $X$  是 URL 的集合,  $\perp$  是用户的集合,  $E_1 \subseteq T \times X$  是 URL 类型与 URL 关系的集合,  $E_2 \subseteq X \times \perp$  是 URL 与用户关系的集合。URL 三部图如图3所示。

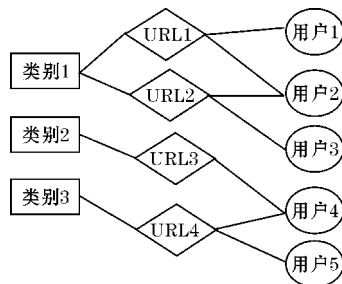


图3 URL 三部图

另外, 可以对相关信息设置权重函数: 用户及 URL 的访问次数的权重函数  $w_{E_2}: E_2 \mapsto \mathbb{R}$ 。将加权重度定义为

$$d_w(v) = \sum_{u \in N(v)} w(v, u)$$

通过三部图, 可以分析  $T$  的双向投影以对类别与用户活动的关系进行分析。本文将这种投影定义为  $Pr(T) = (T, \perp, E_{pr(T)})$ 。

其中  $E_{pr(T)} = \{(v, u) \in T \times \perp \mid \exists z \in X \text{ s. t. } (v, z) \in E_1, (u, z) \in E_2\}$ , 如图4所示。

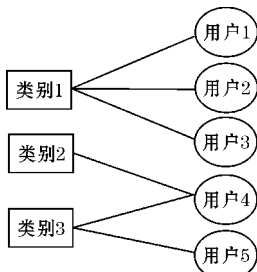


图4 用户与类别的双向投影

如果三部图是标有权重的,则投影会得出加权的函数  $w_{E_{Pr(T)}}: E_{Pr(T)} \mapsto R$ , 将其定义为

$$w_{E_{Pr(T)}}(v, u) = \sum_{z \in N(u) \cap N(v)} w_{E_2}(u, z) \quad (1)$$

三部图设置完成后,在分析用户与类别的关系时,可以舍弃常用的距离量化,而依赖三部图中的随机游走,计算从用户  $u$  访问不同 URL 的概率分布。在此定义任意节点  $v \in T, z \in X$ , 定义从  $z$  到  $v$  的概率为

$$P_{z \rightarrow v} = \frac{w(z, v)}{d_w(z)} \quad (2)$$

同理,对任意节点  $z \in X, u \in \perp$ , 可定义概率为

$$P_{u \rightarrow z} = \frac{w(u, z)}{d_w(u)} \quad (3)$$

由此得到从  $u$  到  $v$  的概率是

$$P_{u \rightarrow v} = \sum_{z \in X} P_{u \rightarrow z} P_{z \rightarrow v} \quad (4)$$

综上,可以将用户访问的随机性定义为从  $\perp$  到  $T$  的概率分布,从而使用多样性指数来进行量化。

### 3 多样性指数

#### 3.1 赫芬达尔多样性指数

不同于传统的赫芬达尔指数<sup>[12]</sup>,本文第2部分将形式上的随机游走用概率分布来表示,即可以定义  $T$  (类别)和(用户)中的节点  $u$  的 Herfindahl 指数为

$$\text{hd}(T, u) = \left( \sum_{v \in T} p_{u \rightarrow v}^2 \right)^{-1} \quad (5)$$

由此可知,当 Herfindahl 指数高时,表明类别更趋于均匀分布;当 Herfindahl 指数低时,表明该类别受众更集中。Herfindahl 指数的值是以类别数为限制的,当分布均匀时就达到了这个上限。在三部图中,可以对类别1和类别3进行分析。

类别1和类别3均对应至少有两个用户访问 URL 的三部图如图5所示。

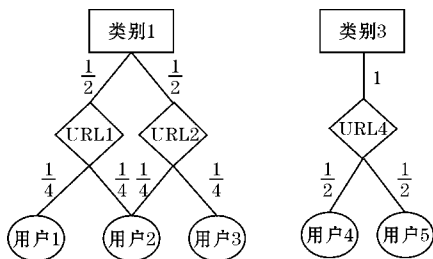


图5 类别1与类别3的三部图

类别1 ( $C_1$ ) 其赫芬达尔系数通过计算可知  $\text{hd}(C_1) = \frac{8}{3}$ ; 类别3 ( $C_3$ ) 的赫芬达尔系数  $\text{hd}(C_3) = 2$ ,

类别1的值更大,可以说明类别1的受众多样性更广泛,而事实也是如此。通过这种方式实现了多样性的量化。

#### 3.2 校准的赫芬达尔指数

正文内容在统计多样性得分时,因为不同类别的数量差距过大,实验可能会受数量的影响,错误地将多样性与数量联系起来。因此,为了消除量的因素,根据文献[13]提出了校准的赫芬达尔多样性的概念。

$$\text{chd}(T, u) = \frac{\text{hd}(T, u)}{\text{hd}(\text{Rand}(T), u)} \quad (6)$$

其中,  $\text{Rand}(T)$  表示用户访问 URL 时,在  $T$  集合下的随机访问形成的三部图。在该随机三部图形成过程中,假设用户访问次数是恒定的,访问任何 URL 都是随机选择的,校准的多样性指数可以将用户访问过程中对不同量的类别所产生的 Herfindahl 指数的差异进一步缩小。

#### 3.3 推荐算法多样性指数

同样,可以对改良的赫芬达尔系数进行分析,最终将各类校准的多样性指数与其在总类别中所占的比例进行计算,得到该推荐算法的多样性指数。

$$\text{chd}(S) = \sum_{v \in T} \text{chd}(v, u) \cdot \frac{\text{count}(v)}{\text{count}(T)} \quad (7)$$

其中,  $S$  表示推荐系统的算法,  $v$  表示  $T$  集合中的某一类别,  $\text{count}(v)$  表示  $v$  类别下的 URL 总数,  $\text{count}(T)$  表示在  $T$  集合中的所有类别的 URL 总数,同样可以对其进行分析,最终将各类校准的多样性指数与其在总类别中所占的比例进行计算,得到该推荐系统的多样性指数。

#### 3.4 多样性指数计算的基本流程

步骤1 从日志中获取 URL 及用户信息,并根据分类词提前设定 URL 种类,将相关信息保存在同一文件中。

步骤2 编写程序将 URL 进行分类并初次筛选,将可分类且分类项明确的 URL 保存下来,形成“类别 URL 用户”的三部图。

步骤3 将所得的三元组中的值导入 MySQL 中,再从中筛选重复数据。第二次筛选将用户访问 URL 次数大于2次的筛选出来,再将其根据类别分别导出,得到多个类别的三部图。

步骤4 将三部图思想编入程序并进行计算,得到



Herfindahl 系数的值以及改良的 Herfindahl 系数的值, 将其可视化并得出推荐系统多样性的值, 对以上内容进行说明。

4 实验

数据来源于搜狗实验室的用户查询日志, 通过分类词筛选, 得到了 929588 个用户, 1801560 条记录和 17 个类别。数据的类别与 URL 数量的关系如图 6 所示。

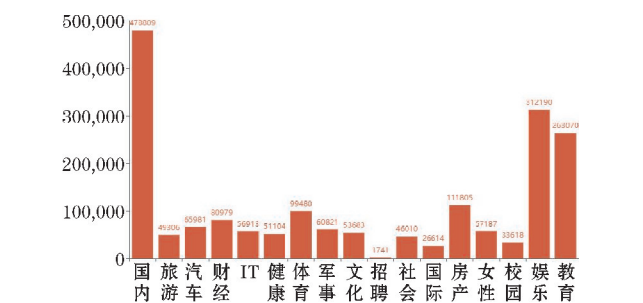
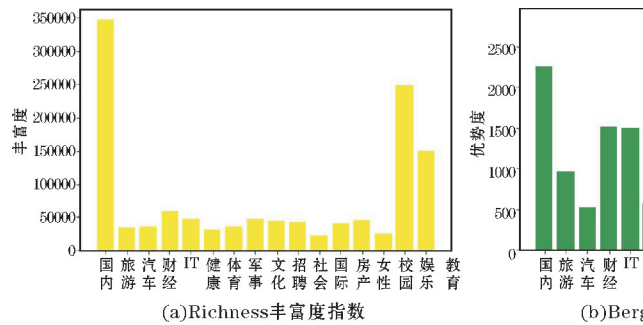
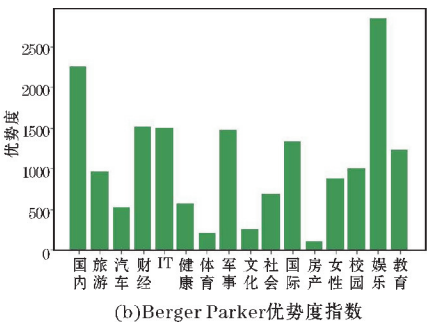


图 6 各类别及其 URL 数量

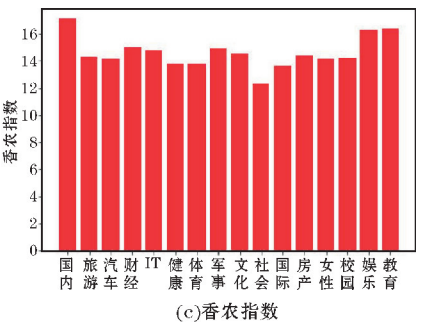
由图 6 可以看出, 关于招聘的信息查询量较少, 调



(a)Richness丰富度指数



(b)Berger Parker优势度指数



(c)香农指数

图 7 各多样性指数及类别的关系

图 7(a) 中的 Richness 多样性指数理论来源于物种丰富度, 在数量更多的类别中, 其多样性值更高。图 7(b) 中的 Berger-Parke 多样性指数会在各类别中选择概率最大的 URL, 从所得结果来看, 在访问量更多的类别中, 整体多样性指数还是更高。并且因为其计算与最大值相关, 故容易受极端值影响。由图 7(c) 可以看出, 与图 7(a)、(b) 类似, “国内”“娱乐”“教育”在多样性方面仍然占据优势, 而“社会”“国际”数量较少的类多样性值更低, 表明这 3 种多样性指数不能很好地体现多样性的概念。

4.2 Herfindanl 多样性指数

正因为现在多样性指数存在缺陷, 所以需要对 Herfindahl 多样性指数进行改进并提出新的多样性指数。各类别的 Herfindahl 多样性指数如图 8 所示。

查访问最多的类是国内信息以及娱乐方面的信息。因为招聘类提供的信息较少, 所以将招聘类去除。对数据进行统计分析, 将日志中访问网站的次数大于 2 次的用户筛选出来。最终得到了 902065 个用户, 715415 条 URL, 16 个类别。

4.1 多样性指数

在许多文献中都提到多样性这个概念, 本文提到的 Herfindahl 指数是通过随机游走来实现一种均匀分布的量化方式, 本文拟进行其他多样性指数的实验, 与 Herfindahl 多样性指数进行对比, 见表 1。

表 1 各多样性指数及多样性指数公式

| 指数名(作者名)                       | 多样性指数公式                 |
|--------------------------------|-------------------------|
| Richness( MacArthur )          | $\sum_i 1$              |
| Berger-Parker( Berger&Parker ) | $(\max(p_i))^{-1}$      |
| Shannon( Shannon )             | $-\sum_i p_i \log(p_i)$ |

实验结果如图 7 所示。

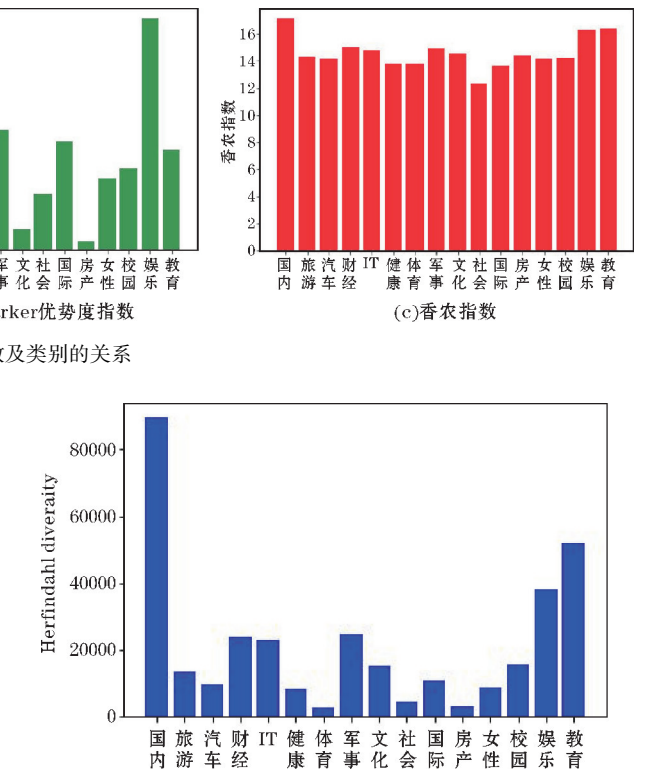


图 8 各类别的 Herfindahl 多样性指数

由图 8 可以看出, 各类别的多样性指数与 URL 数量的分布类似, 这时考虑 Herfindahl 的多样性可能与什么因素相关。结合前面, 可以从量的角度进行思考, 见图 9。

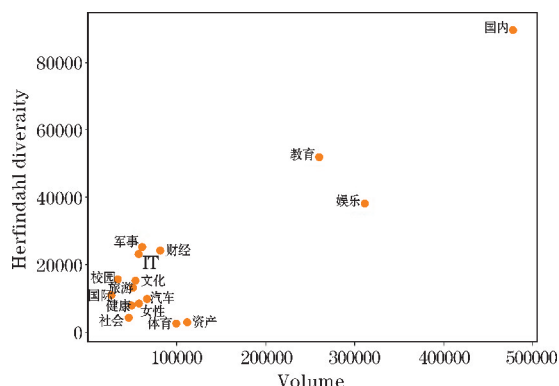


图9 各类别的 Herfindahl 多样性指数与类别数量的关系

由图9可以较为明显地发现,“国内”“教育”“娱乐”的数量相较于其他类别来说更为庞大,所以这三者的多样性指数更高,这样无法实现多样性指数引入的初衷。此外,根据类别分析,可以发现“国内”与“国际”这两个类不同于其他类,故将类别分为两组,采用不同标记表示:一组是按照区域划分的含有“国内”与“国外”2个元素,另一组是按照内容划分的其他13个元素。因此,为使多样性得分只捕捉多样性而不被数量所影响,采用了改进 Herfindahl 多样性指数的办法。校准的赫芬达尔多样性(Calibrated Herfindahl diversity)与数量的关系如图10所示。

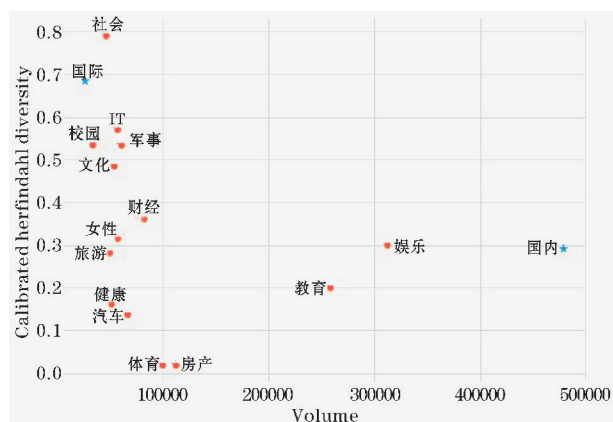


图10 校准的赫芬达尔多样性指数与类别数量的关系

由图10可知,在引入随机行走模型后,“旅游”这个数量较少的类与“国内”这个数量较大的类具有相似的多样性指数。此外,还可以看出:校准的多样性指数的区间大都在0~0.7。而“社会”类的校准值过高,这是因为这个类数量相对较少并且在用户的访问记录中多个用户访问的URL较少。由图10可以看出,这个指数的提出不能完全将量的影响剥离,一个类别的访问量更多就说明该类别的受众更多,说明该类别可以有更多的机会吸引更广泛的受众。但本文提出的改进指数能在一定程度有效减少类别的量对多样性的影响。

最后,用公式(7)进行计算,算得搜狗搜索引擎的推荐算法得分为

$$\text{chd}(\text{sogou}) = \sum_{v \in T} \text{chd}(v, u) \cdot \frac{\text{count}(v)}{\text{count}(T)} = 0.287594648$$

至此,本文实现了推荐系统多样性的量化,并得到推荐系统的多样性指数。通过此模型得到的多样性指数可以作为推荐系统的评价参数。

## 5 结束语

本文结合推荐系统的特性,提出了对其多样性进行量化的多样性指数——校准的赫芬达尔指数。通过对搜狗搜索引擎的用户日志进行分析研究,发现实验所得的校准的 Herfindahl 多样性指数可以对推荐系统的多样性进行描述。对类别受众进行分析,还有很多潜在的地方可以去挖掘:例如仿照类似方法寻找新的推荐系统评价指数<sup>[14]</sup>以及寻找系数来量化检测推荐算法中的偏差<sup>[15]</sup>,还可以将多样性这个概念作为底层算法的优化并将其度量指数纳入推荐系统<sup>[16]</sup>。

## 参考文献:

- [1] Bello-Organ G, Mesas R M, Zarco C, et al. Marketing analysis of wineries using social collective behavior from users' temporal activity on Twitter[J]. Information Processing & Management, 2020, 57(5):102220.
- [2] 陈兴喆,王强. 基于Web日志挖掘的电商平台产品个性化推荐算法研究[J]. 科技与企业, 2015(17):92-93.
- [3] Liu Y, Miao J, Zhang M, et al. How do users describe their information need: Query recommendation based on snippet click model[J]. Expert Systems with Applications, 2011, 38(11):13847-13856.
- [4] Bozdag E. Bias in algorithmic filtering and personalization[J]. Ethics and information technology, 2013, 15(3):209-227.
- [5] Zhou T, Kuscsik Z, Liu J G, et al. Solving the apparent diversity-accuracy dilemma of recommender systems[J]. Proceedings of the National Academy of Sciences, 2010, 107(10):4511-4515.
- [6] 王睿. 基于图网络结构的推荐方法研究[D]. 哈尔滨:哈尔滨理工大学, 2019.
- [7] Iordan A E. Optimal solution of the Guarini puzzle

- extension using tripartite graphs [J]. MS&E, 2019, 477(1):012046.
- [8] 戴瑾如. 基于加权三部图模型的推荐算法研究[D]. 广州:华南理工大学, 2018.
- [9] 王泓, 方艳梅, 黄方军. 基于信息熵的语言风格分析方法初探[J]. 中山大学学报(自然科学版), 2020, 59(6):113-125.
- [10] 迟景明, 任祺. 基于赫芬达尔-赫希曼指数的我国高校创新要素集聚度研究[J]. 电子测试, 2016(4):5-9.
- [11] Poulain R, Tarissan F. Quantifying the diversity in users activity: an example study on online music platforms[C]. Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE, 2018:3-10.
- [12] Hall B. A note on the bias in herfindahl-type measures based on count data [J]. Revue d'économie industrielle, 2005, 110(1):149-156.
- [13] Poulain R, Tarissan F. Investigating the lack of diversity in user behavior: The case of musical content on online platforms[J]. Information processing & management, 2020, 57(2):102169.
- [14] 何慧芳. 基于用户感知的电子商务信息推荐服务质量评价指标体系构建研究[D]. 保定:河北大学, 2020.
- [15] Courtland R. Bias detectives: the researchers striving to make algorithms fair [J]. Nature, 2018, 558(7710):357-357.
- [16] 周艳榕. 基于个性化特征的电子商务智能推荐系统[J]. 现代电子技术, 2020, 43(19):155-158.

## Research on Diversity Index for Evaluating Recommendation System

SUN Chenkai, AN Junxiu

(College of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

**Abstract:** Aiming at the problem that the huge amount of data makes it difficult for users to obtain the required information and the recommendation system evaluation system lacks diversity evaluation indicators, the Herfindahl diversity index based on the tripartite graph calibration is proposed to quantify the diversity of the recommendation system. First, URLs are classified according to the set classification method; then a three-part graph of category-URL-user can be designed and formed; Secondly, the original Herfindahl index is improved to reduce the impact of quantitative differences on diversity; Finally, Combined with the improved Herfindahl diversity index, the diversity index of the recommendation system is obtained. The emergence of diversity index helps to not only pay attention to the accuracy of the recommendation, but also consider whether the recommendation information is comprehensive when evaluating the recommendation system. Experiments show that the improved Herfindahl index based on the method proposed in this experiment can accurately quantify the audience diversity of the recommendation system category.

**Keywords:** computer software and theory; recommendation system; Herfindahl index; diversity index; URL; tripartite graph