

融合用户兴趣的海量 Web 多属性 数据查询算法

孙建平¹, 李志河²

(1. 北京师范大学 人工智能学院, 北京 100088; 2. 山西师范大学 信息技术与智慧教育研究院, 太原 030031)

摘要: 为了提升客户满意度, 通过融合用户兴趣, 进行了海量 Web 多属性数据查询。根据数据的属性获取 Web 多属性数据的主题向量, 并考虑遗忘因子, 动态管理用户兴趣, 以更加准确地反映用户的当前兴趣。在此基础上, 利用余弦相似度计算 Web 多属性数据特征向量与用户兴趣主题向量之间的相似度, 并利用 Borda 计数法整合初始化查询列表和个性化查询列表, 获得综合考虑用户兴趣的最终查询列表, 实现个性化数据查询体验。实验结果表明: 该算法可实现 Web 多属性数据查询, 具有处理歧义性查询的能力, 且随着查询次数的增加, 用户满意度接近 100%。说明本文算法可以根据用户的查询内容, 进一步了解用户所需, 并为用户提供更加精准的查询结果。

关键词: Web 多属性数据; 逆用户频率指数; 遗忘因子; 动态兴趣模型; 扩展机制; 相似度

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5497(2025)07-2372-06

DOI: 10.13229/j.cnki.jdxbgxb.20240603

Query algorithm for massive Web multi-attribute data that integrates user interests

SUN Jian-ping¹, LI Zhi-he²

(1. School of Artificial Intelligence, Beijing Normal University, Beijing 100088, China; 2. Research Institute of Information Technology and Intelligence, Shanxi Normal University, Taiyuan 030031, China)

Abstract: In order to improve customer satisfaction, massive Web multi-attribute data queries are conducted by integrating user interests. Based on the attributes of the data, obtain the topic vector of Web multi-attribute data, and consider the forgetting factor to dynamically manage user interests, in order to more accurately reflect the current interests of users. On this basis, cosine similarity is used to calculate the similarity between Web multi-attribute data feature vectors and user interest topic vectors, and Borda counting method is used to integrate the initialization query list and personalized query list to obtain the final query list that comprehensively considers user interests, achieving a personalized data query experience. The experimental results show that the algorithm can achieve Web multi-attribute data queries and has the ability to handle ambiguous queries. And as the number of queries increases, user satisfaction approaches 100%. This article demonstrates that the algorithm can further understand the needs of users based on their

收稿日期: 2024-05-31.

基金项目: 天津市科技计划项目(23KPMRC00060).

作者简介: 孙建平(1974-), 女, 副教授. 研究方向: 人工智能, 移动学习. E-mail: m13663575630@163.com

query content, and provide users with more accurate query results.

Key words: Web multi-attribute data; reverse user frequency index; forgetting factor; dynamic interest model; expansion mechanism; similarity

0 引 言

Web 数据在电子商务、社交媒体、在线新闻等领域呈现出爆炸式增长的态势^[1],其中多属性数据尤为丰富,涵盖了商品价格、品牌、评价等维度^[2,3]。然而,当用户在海量 Web 多属性数据中进行查询时,无法综合考虑多个属性之间的关联和组合,深入理解用户的个性化需求和兴趣^[4,5],导致用户难以迅速找到真正符合其个性化需求的信息,极大地影响了用户体验^[6]。因此,如何有效融合用户兴趣,实现对海量 Web 多属性数据的高效查询,成为当前信息检索领域亟待解决的问题。

杨矫云等^[7]利用概率近似正确理论,通过 k 个随机排序器抽取样本推断流数据的 Top- k 序列。当数据流分布不均匀或变化快速时,Top- k 结果可能会偏离真实的 Top- k 元素;高俊杰等^[8]基于群体智能的协同搜索策略,将查询任务分解为子任务并分配给粒子进行处理。粒子间通过合作与竞争实现高效查询。该算法对静态或变化较慢的数据更有效,对快速变化的 Web 数据,可能需要额外的机制来保证其适应性和鲁棒性;罗芳等^[9]采用分布式存储架构和优化的检索算法,支持 Web 数据多属性关联存储和快速查询。该算法主要关注数据的关联性和高效检索,未考虑查询是否与用户兴趣相匹配;Tung 等^[10]通过理解口语会话上下文预测用户查询意图,确保查询结果与用户兴趣匹配,但需克服口语理解和识别上的挑战,尤其是该算法未考虑用户兴趣的动态变化特性,导致其在提供全面和深入个性化搜索体验方面存在缺陷。

为了提高 Web 数据查询效率和准确性,以满足用户个性化需求,本文提出融合用户兴趣的海量 Web 多属性数据查询算法,通过融合用户兴趣并考虑其动态变化,对海量 Web 多属性数据进行个性化查询。

1 海量 Web 多属性数据查询

1.1 Web 多属性数据主题向量获取

Web 上包含海量的多属性数据^[11],这些数据

不仅数量庞大,而且结构复杂,包括文本、图像、视频等形式。这些信息的内容和主题各不相同。通过获取主题向量,可以将这些数据按照主题进行分类和组织,使得信息的管理和检索更加高效。

设定 U 为 Web 多属性数据查询用户集合, R 为待查询的 Web 多属性数据集^[12], T 为标识数据属性的标签集。由此,Web 多属性数据主题向量可以表示为:

$$P = \{(T_1, w_1), (T_2, w_2), \dots, (T_n, w_n)\} \quad (1)$$

式中: P 为 n 维向量; T_i 为第 i 个 Web 多属性数据查询标签主题; w_i 为用户对其兴趣度,该参数反映 T_i 的权重,二者构成了特征项 (T_i, w_i) 。

标签主题 T_i 的表达式为:

$$T_i = \{(t_{i,1}, w_{i,1}), (t_{i,2}, w_{i,2}), \dots, (t_{i,k}, w_{i,k})\} \quad (2)$$

式中: $t_{i,k}$ 为用户查询的第 i 个主题下的编号为 k 的标签, $t_{i,k} \in T_i$, $T_i \subseteq T$; $w_{i,k}$ 为用户对其的兴趣度,即标签权重。 $w_{i,j}$ 的表达式为:

$$w_{i,k} = tf_{i,k} \times \log\left(\frac{N}{nt_{i,k}}\right) \quad (3)$$

式中: $tf_{i,k}$ 为用户使用标签 $t_{i,k}$ 的频率; $\log\left(\frac{N}{nt_{i,k}}\right)$ 为逆用户频率指数; N 为查询 Web 多属性数据的用户数量; $nt_{i,k}$ 为通过标签 $t_{i,k}$ 查询 Web 多属性数据的用户个数。

由于 Web 数据海量且复杂,通过获取 Web 多属性数据的主题向量可以先将数据按照主题进行聚类或索引,然后在相关主题的数据集中进行搜索,从而提高查询效率。

1.2 用户兴趣动态管理

Web 多属性数据的主题向量是对数据内容深层语义信息的提取和表示。当用户查询或浏览 Web 数据时,这些数据的主题向量可以反映出用户的兴趣所在。通过分析用户与数据的交互行为,可以推断出用户对哪些主题的数据更感兴趣。然而用户兴趣具有时变性^[13,14],为了使用户获得更准确的 Web 多属性数据查询结果,考虑遗忘因子,结合 Web 多属性数据的主题向量,更加准确地反映用户的当前兴趣。

遗忘因子通过下式确定:

$$F(P) = \log^2(e_c - e_t) / hl \quad (4)$$

式中: e_c 为当下时间; e_t 为最后一次使用该标签进行 Web 多属性数据查询的时间; hl 为用户对某一特定 Web 多属性数据的兴趣衰减到一半所需时间。

本文利用遗忘因子描述用户兴趣逐渐遗忘的过程,在实时更新 Web 多属性数据查询标签、主题权重的过程中,删除用户已不感兴趣的标签或主题,达到动态管理用户兴趣的目的,具体通过以下几步实现:

(1) 通过式(4)对 Web 多属性数据主题向量 P 的遗忘因子进行运算,然后根据 $w'_{i,k} = w_{i,k} \times F(P)$ 对其权重进行修正。

(2) 设 t_1 为用户查询 Web 多属性数据时使用的新标签,当 t_1 已存在于兴趣主题中,则对其权重以及对应主题权重进行更新,并对 e_t 做对应更改;当 t_1 可作为一个新标签添加到兴趣主题中时,则先对 $F(t_1)$ 进行确定,用当下时间更新 e_t ,计算与该标签对应的主题权重;当 t_1 不属于任何已有主题时,将之视为新主题加入用户兴趣模型中,并对其遗忘因子进行确定,用当下时间更新 e_t ,计算其权重。

(3) 去除用户不感兴趣的标签^[15]。按照标签权重计算结果排序,并与设定阈值 δ 比较,保留权重值高于阈值的标签,将余下标签予以去除后,完成主题权重的修正。

(4) 去除用户不感兴趣的主体。按照主题权重计算结果排序,并与设定阈值 ϵ 比较,保留权重值高于阈值的主题 P_i ,将余下主题予以去除。

由此,完成用户兴趣动态管理,确保查询结果、推荐内容等与用户当前的兴趣和需求更加匹配,从而提升用户的满意度和体验。

1.3 Web 多属性数据查询

Web 多属性数据包含了丰富的信息,为用户提供更加个性化、精准的服务,在完成用户兴趣动态管理后,当用户发送 Web 多属性数据查询请求时,通过余弦相似度计算每一条 Web 多属性数据特征向量与用户兴趣主题向量之间的相似度,并根据相似度的高低来决定数据在查询结果中的排序,相似度高的数据排名更靠前,以便为用户提供个性化数据查询体验。余弦相似度度量表达式为:

$$\text{Sim}(Q_i, P_i) = \cos(Q_i, P_i) = \frac{\sum_{j=1}^n g_j \times w_j}{\left(\sum_{j=1}^n g_j^2\right)^{\frac{1}{2}} \left(\sum_{j=1}^n w_j^2\right)^{\frac{1}{2}}} \quad (5)$$

式中: Q_i 为 Web 多属性数据特征参数; P_i 为用户兴趣主题向量; w_j 为用户兴趣主题权重; g_j 为待查询 Web 多属性数据主题权重。

Web 多属性数据查询流程为:

第 1 步 搜索引擎接收用户提交的 Web 多属性数据查询请求后,生成一个初始查询结果集合,用 A 表示,返还给用户的 Web 多属性数据初始化查询列表表示为 τ 。

第 2 步 初始化一个循环计数器,并从 0 开始计数计算循环次数。

第 3 步 对于数据集 R 中的第 i 个 Web 多属性数据,通过式 $Q_i = \{(q_1, g_1), (q_2, g_2), \dots, (q_o, g_o)\}$ 确定其特征向量。

第 4 步 通过式(5)求得动态用户兴趣模型主题向量与第 i 个 Web 多属性数据特征向量之间的相似度。

第 5 步 当第 i 个 Web 多属性数据是数据集 R 中最后一个数据,则执行下一步;反之,令 $i = i + 1$,退回到第三步。

第 6 步 根据相似度的高低对数据集 R 中的 Web 多属性数据进行排列,产生个性化查询列表 τ' 。

第 7 步 根据 Borda 计数法将初始化查询列表 τ 、个性化查询列表 τ' 中的排名信息转化为 Borda 分数,并将这些分数进行汇总,从而生成一个综合考虑用户兴趣的最终查询列表 τ'' 。

第 8 步 将最终查询列表 τ'' 提供给用户,完成 Web 多属性数据查询。

2 实验分析

2.1 实验设置

以互联网为实验对象,从该网站获取海量 Web 数据,每条新闻记录均由标题、内容、发布时间、所属分类、标签等属性构成,收集与 Web 数据相关的用户行为数据,包括用户对新闻的浏览量、点赞数、评论、分享次数等,以及用户 ID、地理位置、浏览历史等个人信息,完成实验样本数据集的

构建。其中用户总数为 1 000,Web 多属性数据为 300 000 个。为模拟真实世界中的分布式数据处理场景,搭建一个包含 10 个节点的集群,每个节点 CPU 均为 Intel Xeon Gold 处理器,其核心数为 20,主频为 2.10 GHz,三级缓存为 27.5 MB;随机存储器内存为 128 G,固态硬盘容量为 256 G。查询机制如图 1 所示。

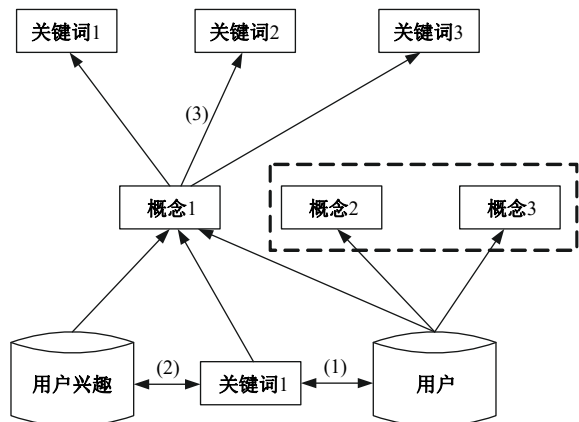


图 1 查询机制
Fig. 1 Query mechanism

在图 1 所示的查询机制中,当用户以“D”为 Web 多属性数据查询关键词时,可利用本体知识查找与“D”相关的概念,并对这些概念进行语义信息。然后通过用户兴趣动态管理确定用户查询的具体意图,删除不满足查询需求的概念。最后利用本体知识进行查询扩展。添加与“D”相关的标签,丰富 Web 多属性数据查询语义信息,提高查询的准确性和完整性。

查询过程中的 Web 属性索引存储形式表如表 1 所示。

表 1 Web 属性索引存储形式表				
Table 1 Web property index storage form table				
表名	RowKey	索引(列族)		
SOURCE	MM-DD	索引:attr1	索引:attr2	索引:attrn
YYYY		排序 Urls	排序 Urls	排序 Urls

将属性数据和其对应索引分别存储在相同表的 Url 列族和 Index 列族中,在 Web 多属性数据查询时,便可以通过各个分布式节点完成属性数据的同时检索,融合各个节点检索结果,即可完成符合查询条件数据的定位。

由此,将本文算法应用到海量 Web 多属性数据查询中,分析其数据查询性能。查询过程中的实验参数设置如表 2 所示。

在上述设置的基础上,选取体育运动爱好者

表 2 实验参数设置	
Table 2 Setting of experimental parameters	
参数名称	参数值
初始化查询列表长度	10
个性化查询列表长度	5
遗忘因子	0.7
数据属性权重	[0.2, 0.3, 0.1, 0.2, 0.2]
交互强度(点击)	1
交互强度(长时间浏览)	2

用户 A,该用户以“体育”为查询主题,以“CBA 联赛”为查询关键词,从互联网上获取有关辽宁男子篮球队 CBA 联赛赛事的多属性数据信息,应用研究算法对该数据信息进行查询,通过对查询结果进行分析,验证本文算法的数据查询性能,实验结果如图 2 所示。



图 2 查询结果分析
Fig. 2 Analysis of query results

分析图 2 可知,应用本文算法在互联网上对辽宁男子篮球队 CBA 联赛赛事的多属性数据信息进行查询是有效的,本文算法能够准确捕捉到与查询主题和关键词高度相关的数据信息,查询结果与用户需求具有高度匹配性,验证了本文算法在数据查询性能上的优异表现。这不仅体现了本文算法在复杂信息环境下的高效处理能力,也进一步证明了其在体育领域数据查询应用中的实用性和可靠性。

2.2 结果与分析

以排序计分(RS)指标衡量 Web 多属性数据查询精度,将文献[7]算法、文献[10]算法作为对比算法,通过对比 3 种算法在所有查询、最优查询以及非最优查询上的排序计分指标差异,分析本文算法的查询性能优势,实验结果如表 3 所示。

表 3 不同算法下的数据查询效果对比

Table 3 Comparison of data query performance under different algorithms

查询类型	本文算法	文献[7]算法		文献[10]算法	
		RS	相比本文算法降低/%	RS	相比本文算法降低/%
所有查询	0.698	0.567	0.188	0.352	0.496
非最优查询	0.493	0.426	0.136	0.312	0.367
最优查询	0.983	0.825	0.161	0.569	0.421

分析表 3 可知,本文算法在所有查询类型上的排序计分均显著优于其他两种算法,在最优查询中表现尤为突出,即生成的查询列表能够精准匹配用户个性化需求。这充分说明了用户兴趣模型在提高 Web 多属性数据查询结果精度方面的重要性,同时证明了本文算法在捕捉和反映用户兴趣变化上的有效性。

点击熵是一个衡量用户点击行为多样性的指标,其值越大,表明用户点击不同网页的数量越多,即查询的歧义性可能越高。本文通过对比分析不同点击熵下各算法在 Web 多属性数据查询上的 RS 指标差异,验证本文算法在处理歧义性查询上的优越性,实验结果如表 4 所示。

表 4 3 种算法在处理歧义性查询上的性能对比

Table 4 Performance comparison of three algorithms in handling ambiguous queries

点击熵	本文算法	文献[7]算法		文献[10]算法	
		RS	相比本文算法降低/%	RS	相比本文算法降低/%
低于 1.7	0.728	0.605	16.9	0.535	26.6
高于等于 1.7	0.503	0.468	6.8	0.412	18.1

分析表 4 得出,本文算法在处理歧义性查询时显示出明显的优越性。当点击熵低于 1.7 时,即用户点击行为相对集中,查询的歧义性较低时,本文算法的排序计分指标值明显高于对比方法,分别提高了 16.9% 和 26.6% 的性能。而当点击熵大于等于 1.7,即用户点击行为更加多样化,查询的歧义性较高时,本文算法依然保持较好的性能,其排序计分相比其他两种算法分别提高了 6.8% 和 18.1%。实验结果表明:本文算法能够更有效地处理具有不同歧义性的查询,为用户提供更准确、更相关的搜索结果。

将文献[7]算法、文献[10]算法作为对比算法,将本文算法、对比算法应用到 Web 多属性数据查询中,各算法下的用户满意度指标差异如图

3 所示。

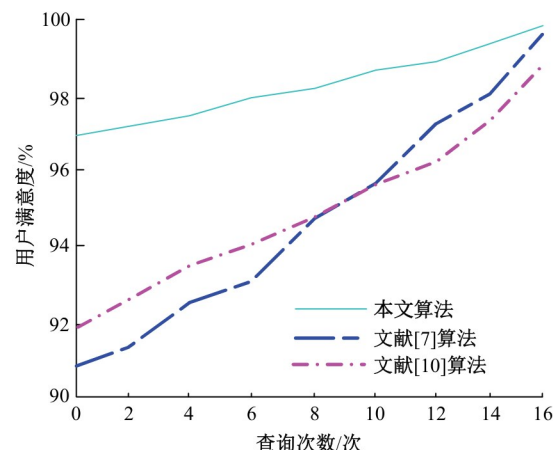


图 3 用户对 Web 多属性数据查询结果的满意度对比

Fig. 3 Comparison of user satisfaction with Web multi-attribute data query results

分析图 3 可知,随着查询次数的增加,3 种算法所得到的客户满意度均呈现上升趋势。其中,本文算法的用户满意度一直保持在 97% 左右,相对较高,这是因为本文方法考虑到了用户兴趣的时变性,可以确保每次生成的查询列表都能紧密贴合用户的当前需求和偏好,且随着查询次数的增加,更能理解用户需求,提供更加具有针对性的查询结果。由此表明,本文算法在个性化服务查询方面具有突出优势。

3 结束语

在海量 Web 多属性数据查询领域,为能够显著提升查询的个性化和准确性。提出了融合用户兴趣的海量 Web 多属性数据查询算法。实验结果表明:本文算法可以较为全面地考虑海量 Web 的属性,并实时更新用户兴趣。查询得到的结果有较高的用户满意度,能够为海量 Web 多属性数据查询提供一种高效、个性化的解决方案。

参考文献:

- [1] 赵红梅,肖明,白宇,等. 面向用户偏好的动态网页数据交互式查询算法[J]. 吉林大学学报:理学版, 2024, 62(2): 417-422.
Zhao Hong-mei, Xiao Ming, Bai Yu, et al. Interactive query algorithm for dynamic web page data based on user preference[J]. Journal of Jilin University(Science Edition), 2024, 62(2):417-422.
- [2] 孙琛琛,申德荣,肖迎元,等. 面向查询式实体解析的多属性数据索引技术[J]. 软件学报, 2022, 33(6):

- 2331-2347.
- Sun Chen-chen, Shen De-rong, Xiao Ying-yuan, et al. Multi-attribute data indexing for query based entity resolution[J]. Journal of Software, 2022, 33(6): 2331-2347.
- [3] 王子泓, 邵莹侠, 何吉元, 等. 基于多空间属性信息融合的序列推荐[J]. 计算机科学, 2024, 51(3): 102-108.
- Wang Zi-hong, Shao Ying-xia, He Ji-yuan, et al. Sequential recommendation based on multi-space attribute information fusion[J]. Computer Science, 2024, 51(3): 102-108.
- [4] 翁彬月, 秦永彬, 黄瑞章, 等. NEMTF: 基于多维度文本特征的新闻网页信息提取方法[J]. 计算机应用研究, 2022, 39(4): 1043-1048.
- Weng Bin-yue, Qin Yong-bin, Huang Rui-zhang, et al. NEMTF: method of news Web content extraction based on multi-dimensional text features[J]. Application Research of Computers, 2022, 39(4): 1043-1048.
- [5] 蒲岍岍, 雷航, 李贞昊, 等. 增强列表信息和用户兴趣的个性化新闻推荐算法[J]. 计算机科学, 2022, 49(6): 142-148.
- Pu Yan-yan, Lei Hang, Li Zhen-hao, et al. Personalized news recommendation algorithm with enhanced list[J]. Computer Science, 2022, 49(6): 142-148.
- [6] 聂卉, 邱以菲. 融合用户兴趣及评论效用的评论信息推荐[J]. 图书情报工作, 2021, 65(10): 68-78.
- Nie Hui, Qiu Yi-fei. Integrating user interests with review helpfulness for review recommendation[J]. Library and Information Service, 2021, 65(10): 68-78.
- [7] 杨矫云, 郭思伊, 李廉. 基于Pac算法的流数据top- k 实时查询[J]. 华中科技大学学报: 自然科学版, 2021, 49(2): 56-61.
- Yang Jiao-yun, Guo Si-yi, Li Lian. Pac based top- k real-time query algorithm for streaming data [J]. Journal of Huazhong University of Science and Technology(Nature Science Edition), 2021, 49(2): 56-61.
- [8] 高俊杰, 杨帆. 基于群体智能的半结构化数据查询优化算法[J]. 计算机仿真, 2021, 38(8): 381-385.
- Gao Jun-jie, Yang Fan. Semi-structured data query optimization algorithm based on swarm intelligence [J]. Computer Simulation, 2021, 38(8): 381-385.
- [9] 罗芳, 李春花, 周可, 等. 基于多属性的海量Web数据关联存储及检索系统[J]. 计算机工程与科学, 2022, 36(3): 404-410.
- Luo Fang, Li Chun-hua, Zhou Ke, et al. An associated storage and retrieval system of massive Web data based on multi-attributes[J]. Computer Engineering and Science, 2022, 36(3): 404-410.
- [10] Tung V, Salvatore A, Giulio J, et al. Spoken conversational context improves query auto-completion in web search[J]. ACM Transactions on Information Systems, 2021, 39(3): 1-32.
- [11] 张星, 张兴. dckdp:改进 k -prototype聚类的差分隐私混合属性数据发布方法[J]. 计算机应用研究, 2022, 39(1): 249-253.
- Zhang Xing, Zhang Xing. Differential privacy mixed attribute data publishing method for improved k -prototype clustering[J]. Application Research of Computers, 2022, 39(1): 249-253.
- [12] 虞文波, 游进国, 牛祥虞. 基于强化学习的数据库多属性索引推荐[J]. 计算机应用研究, 2023, 40(6): 1789-1793.
- Yu Wen-bo, You Jin-guo, Niu Xiang-yu. Mira: database multi-attribute index recommendation based on reinforcement learning[J]. Application Research of Computers, 2023, 40(6): 1789-1793.
- [13] 麻天, 余本国, 张静, 等. 基于混合聚类与融合用户兴趣的协同过滤推荐算法[J]. 电子技术应用, 2022, 48(4): 29-33.
- Ma Tian, Yu Ben-guo, Zhang Jing, et al. Collaborative filtering recommendation algorithm based on hybrid clustering and user preferences fusion[J]. Application of Electronic Technique, 2022, 48(4): 29-33.
- [14] 凌宇, 单志龙. 基于兴趣增强的知识概念推荐系统[J]. 计算机应用, 2023, 43(12): 3697-3702.
- Ling Yu, Shan Zhi-long. Knowledge concept recommendation system based on interest enhancement[J]. Journal of Computer Applications, 2023, 43(12): 3697-3702.
- [15] 张彬, 徐建民, 吴姣. 大数据环境下基于知识图谱的用户兴趣扩展模型研究[J]. 现代情报, 2021, 41(8): 36-44.
- Zhang Bin, Xu Jian-min, Wu Jiao. Research on user interest expansion model based on knowledge graph in big data environment[J]. Modern Information, 2021, 41(8): 36-44.