

# 基于历史反馈的 IETM 查询结果优化排序研究

呼凯凯,徐宗昌,倪明仿,张永强,胡春阳

(装甲兵工程学院 技术保障工程系,北京 100072)

**摘 要:** 为了满足用户对 IETM 系统在集成数据环境中对装备保障数据查询的查准率需求,构建了基于历史反馈的 IETM 查询结果优化排序算法,分析了用户在使用 IETM 进行数据查询过程中的点击浏览行为,从个人用户行为反馈和历史用户行为反馈两方面对影响查询结果排序的影响因子进行量化,并最终实现对 IETM 在集成数据环境下对数据查询结果的优化排序.实验结果表明,所提出的优化排序算法在保证装备综合保障数据查全率的基础上提升数据查询的准确率.

**关键词:** 历史反馈;数据查询;交互式电子技术手册;优化排序

**中图分类号:** TP391

**文献标识码:** A

**文章编号:** 1000-7180(2015)11-0176-04

## Research on Optimizing Sort of IETM Retrieve Results Based on Historical Feedback

HU Kai-kai, XU Zong-chang, NI Ming-fang,  
ZHANG Yong-qiang, HU Chun-yang

(Department of Technical Support Engineering, Academy of Armored Force Engineering, Beijing 100072, China)

**Abstract:** Aiming to meet the precision demand when query the equipment integrated logistic support data in the integration data environment by IETM, a optimizing sort of IETM retrieve results based on historical feedback was built. It analyses the behaviors about browsing and clicking while querying the data, quantifies the influence factor that influencing the result of retrieve from the personal behavior feedback and the historical behavior feedback, and realizes finally the optimizing sort of IETM retrieve results in the integration data environment. The experimental result shows that this arithmetic of optimizing sort can promote the precision demand in base of keeping the recalling demand of equipment support data.

**Key words:** historical feedback; Data retrieval; IETM; Optimizing Sort

### 1 引言

交互式电子技术手册 (Interactive Electronic Technical Manual, IETM) 作为装备综合保障领域一项新的研究内容,其为用户提供了装备基本原理、使用操作和维修等技术信息<sup>[1-2]</sup>.然而,随着武器装备的不断发展以及高新技术的不断应用,装备维修保障活动中所需的技术信息越来越多, IETM 内部所包含的有限的技术信息显然无法完全满足装备日常的综合保障活动,装备技术保障人员

不得不借助 IETM 内置的搜索引擎从集成数据环境去获取更多的保障数据.但是,现有的搜索引擎只是根据简单的数据相似度对所查询出的数据结果进行排序,这显然无法满足用户对查询准确率的要求.

历史用户的反馈信息主要来自于用户的搜索行为日志,其包含了用户在查询过程中的查询词组、对查询结果的点击记录等相关信息<sup>[3-4]</sup>.这些信息反映了用户在搜索过程中的行为过程以及对查询结果相关性的判定.为了提升 IETM 在集成数据环境中对相关数据查询的准确度的要求,本文对用户历史行

为进行分析,提出了一种基于历史反馈的 IETM 查询结果优化排序算法。

## 2 IETM 用户历史行为分析

用户行为分析是在用户在进行数据查询过程中,对用户与搜索引擎的交互信息进行分析的过程,目的是通过分析用户的查询请求、点击过程以及相应的浏览行为,挖掘用户的真实查询意图以及对应的查询结果与用户需求的相关性<sup>[5-6]</sup>。针对 IETM 用户使用中的行为流程,可以从两方面对其行为进行分析,分别是:①个人用户行为分析,主要是指个人用户查询之前对 IETM 内部数据信息的历史浏览与点击行为,其反映了用户在查询时的个人查询意图;②历史用户行为分析,主要是指不同历史用户在查询之后对相应结果的浏览与点击行为,其反映了用户对查询结果相关性的判定过程。因此,为了获取 IETM 用户的查询意图以及统计不同数据的相关性,可以从个人用户行为以及历史用户行为两方面进行分析,用于改进搜索结果,提高搜索质量。

### 2.1 个人用户行为分析

IETM 集成数据环境下的数据查询主要是指当 IETM 数据无法满足用户进行辅助维修保障工作时,在集成数据环境中进一步获取相关信息的过程。而用户在进行数据查询之前一定对 IETM 中与用户需求相关的信息进行浏览学习,这些数据信息必定是与用户查询意图最相关的数据,其所查询到的数据信息的相关度具有很好的修正作用。因此,通过寻找用户浏览记录中与查询相关的数据信息集,并将其用于判定本次查询结果数据的相关性。

### 2.2 历史用户行为分析

IETM 用户在每次搜索的过程中,系统都会为用户返回成百上千的结果,用户通过对数据的标题以及摘要进行判断,点击浏览其认为质量相对较高并且满足其用户需求的查询结果。如果该数据能够为用户提供装备维修保障过程中所需的信息,则用户会花费较多的时间对其进行浏览,反之亦然。而通过多不同历史用户相关的点击浏览行为,可以准确地判断相关数据在查询条件之下的信息重要程度。因此,可以通过对历史不同用户的点击浏览行为对查询结果数据的信息含量进行判断,进而实现对查询结果重要程度排序的反馈。

## 3 基于历史反馈的影响因子量化模型

### 3.1 个人用户反馈影响因子量化

在计算个人用户查询结果反馈时,可以首先将用户正在浏览的相关信息作为第一个匹配信息  $s_1$ 。同时,将用户浏览记录中的后一条记录  $s_2$  与信息  $s_1$  进行相似度对比,如果信息  $s_2$  与信息  $s_1$  之间的相似度满足所设定的阈值  $\eta$ ,则认为信息  $s_2$  也是与用户需求相关的历史浏览信息。重复上述操作直至所提取到的用户浏览信息与  $s_1$  的相似度小于所设定的阈值  $\eta$ 。最终得到与用户需求相关的数据文档信息集  $S(s_1, s_2, \dots, s_n)$ 。在获取到反映用户需求的相似数据集  $S$  之后,可以从需要进行优化排序的查询结果数据集  $C(c_1, c_2, \dots, c_n)$  中依次选取数据文件  $c_i$ , 将其与用户已浏览的相似数据集  $S$  中的数据进行相似度计算,并对最后的相似度结果进行平均化,得出数据文件  $c_i$  与数据集  $S$  的平均相似度。对于浏览记录中信息的相似度计算可以采用文档的余弦夹角相似法进行计算,具体公式如式(1)所示,其中  $d_1$ 、 $d_2$  表示文档的向量,  $|d_1|$ 、 $|d_2|$  表示文档的向量模<sup>[7]</sup>。

$$\text{sim}(d_1, d_2) = \frac{d_1 \times d_2}{|d_1| \times |d_2|} \quad (1)$$

然而,仅根据相似度对查询结果集  $C$  进行优化排序是不准确的。因为,如果查询结果数据集中的数据与用户已浏览的数据相似度过高甚至相同时,该数据明显不是用户所需要的数据。用户所需要的数据是与已查看的数据相似但不相同的数据;当然如果查询结果数据集与用户之前浏览的数据相似度特别低甚至为零,则说明所查询出的数据内部所包含的内容完全不属于用户的需求范围。因此,用户最需要的信息应是与之前浏览的信息相似但是具有一定的区分性的信息,而这类信息应是介于相似度为 0 和 1 之间的数据,也就是说优化权值会随着查询结果与用户已浏览信息的相似度逐渐增大,并在某一点处达到一个峰值,而当相似度大于这一峰值时,优化权值又会随着相似度的增加而不断地减小。

对于个人用户反馈优化权值与数据相似度的分布函数,可以考虑高斯分布函数,其分布函数如式(2)所示:

$$\omega_{\text{个入}} = \beta \frac{e^{-\frac{(ax-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} \quad 0 < x < 1. \quad (2)$$

式中,  $x$  为分布描述的变量, 在这里指查询结果与用户已浏览数据的相似性;  $\alpha, \beta$  为调节系数;  $\mu, \sigma$  为相应的参数. 其中  $\sigma$  控制曲线的形状,  $\sigma$  越小, 则函数曲线越陡峭. 很明显, 当  $\sigma$  越小, 用户越想得到相似度分布聚集在  $\mu$  值附近的区间; 反之, 当  $\sigma$  越大, 用户希望相似度权值的分布越平缓.  $\mu$  则表示曲线函数的峰值位置,  $\mu$  越靠近 1 的位置, 则表明用户越需要与已浏览数据相似的数据, 而  $\mu$  越靠近 0, 则表示用户更需要一些多样化的数据.  $\sigma$  和  $\mu$  分别取不同值时的分布函数图像如图 1 所示.

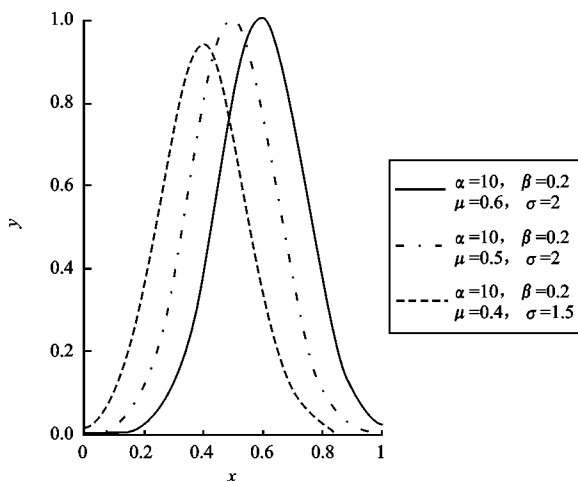


图 1 不同取值的高斯函数分布示例图

通过对高斯函数图像进行分析, 可以看出, 用户已浏览信息与用户查询结果的相似度的转化权值完全可以用该函数图像进行表示. 另外, 用户可以根据自己的搜索意图, 动态调整搜索结果的排序结果.

### 3.2 历史用户反馈影响因子量化

在对历史用户反馈影响进行量化时, 主要考虑装备保障数据的点击次数以及相应的点击时间. 另外, 用户在查询过程中, 采用单一查询词进行查询的相对较少, 大部分还是采用多个查询词进行查询. 因此, 在具体的计算中还应考虑词  $k$  在查询过程中的重要程度. 具体的计算公式如下:

$$\omega_{\text{历史}} = \frac{tf_{k,Q}}{|V_Q|} \times \sum_{i=1}^{|Q|} \frac{n_{d_i} \times \overline{T_{d_i}}}{\sum_{j=1}^n (n_{q_{ij}} \times \overline{T_{q_{ij}}})} \quad (3)$$

式中,  $T_{fk,Q}$  为历史记录集  $Q$  中查询词  $k$  的频率;  $V_Q$  为历史记录集  $Q$  中查询词总数;  $n_{d_i}, T_{d_i}$  为包含词  $k$  的第  $i$  次查询中数据  $d$  被点击的次数以及平均浏览时间;  $n_{q_{ij}}, T_{q_{ij}}$  为包含词  $k$  的第  $i$  次查询中数据  $j$  被点击的次数以及平均浏览时间;  $n$  为包含词  $k$  的第  $i$  次

查询中用户总的点击次数.

在实际应用中用户对页面的浏览时间还与页面内容的多少有关, 而对于不同的数据因为其所包含的内容大小而产生的用户时间也不同<sup>[8]</sup>. 因此, 在对数据的平均浏览时间的计算中, 需要对不同内容含量的数据所产生的浏览时间进行均量化. 设用户浏览时间用  $T = (t_1, t_2, \dots, t_n)$  表示,  $t_i$  为用户在查询结果集中浏览数据  $i$  的时间, 假设页面内容大小为  $m$ , 则用户浏览时间权重表示为

$$T = t_i / \sum_{i=0}^n t_i \cdot m_i / \sum_{i=0}^n m_i \quad (4)$$

### 3.3 整体优化

根据基于历史反馈的影响因素进行分析, 影响最终排序模型的因素有  $\omega_{\text{个人}}$  和  $\omega_{\text{历史}}$ , 而两者对查询排序结果的影响大小是不相同的. 对于有些用户来说其可能更想了解历史用户的选择, 因为历史用户点击反馈反映了大量的用户对查询结果进行浏览过程中的选择结果, 其展现了查询过程中用户的选择经验, 更能对相关的查询结构的排序进行优化. 而对于另一些用户来说, 其认为其查询意图相当明确, 自己所浏览的相关数据更能体现查询意图, 并且希望系统呈现出个性化的查询结果. 因此, 在查询结果排序中更依赖于个人历史反馈. 因此, 在具体的优化中可以根据用户自身的意愿调整相应优化权值的权重. 因而, 最终对查询结果的优化权值可以设计相应的权重进行计算.

$$\omega_{\text{总}} = \lambda_1 \times \omega_{\text{个人}} + \lambda_2 \times \omega_{\text{历史}} \quad (5)$$

式中,  $\lambda_1, \lambda_2$  分别代表不同影响因子的权重系数.

而对于最终的用户排序结果, 可以根据优化后的相似度进行排序. 优化后的相似度可以根据式(6)进行计算

$$SC_{\text{最终}} = \omega_{\text{总}} \times SC_{\text{原始}} \quad (6)$$

## 4 实验验证

本节采用某型坦克的 IETM 统对所提出的查询优化排序算法的查询效果进行验证. 采用的数据集包含 500 个数据信息, 其中故障类信息 100 个, 发动机相关数据信息 250 个. 假设发动机故障数据信息作为用户的查询数据, 计算参数为:  $\alpha = 10, \beta = 0.35, \eta = 0.6, \sigma = 2, \mu = 6$ , 优化排序算法与原始排序算法的查全率  $P$  与查准率  $R$  对比表 1 所示, 查询结果相关度对比如图 2 所示.

从表1可以看出,基于用户反馈的 IETM 数据查询结果优化排序算法的查准率要明显高于传统的排序算法。由于查全率与查准率相对互补,一般情况下,提升查准率会导致相应的查全率降低。但是,本文所提出的优化排序算法在原始的排序算法中考虑了用户使用查询功能时的行为反馈,改变了原始的相关度值,因此,并没有降低相应的查全率,而是与原系统的排序算法基本持平。

图2反映了十次查询实验中,改进排序算法与原始排序算法查询结果相关度值的比较。从图中可以看出,由于考虑了历史用户点击的记录、页面浏览时间以及个人用户的相关浏览记录,改进后的排擅算法查询结果相关度值要优于原算法。

表1 两种方法计算结果比较

| 返回数据量 | 优化算法 |      | 原始算法 |      |
|-------|------|------|------|------|
|       | P    | R    | P    | R    |
| 40    | 0.48 | 0.98 | 0.47 | 0.83 |
| 60    | 0.52 | 0.89 | 0.52 | 0.77 |
| 80    | 0.59 | 0.85 | 0.60 | 0.70 |
| 100   | 0.67 | 0.79 | 0.68 | 0.66 |
| 120   | 0.82 | 0.73 | 0.84 | 0.59 |

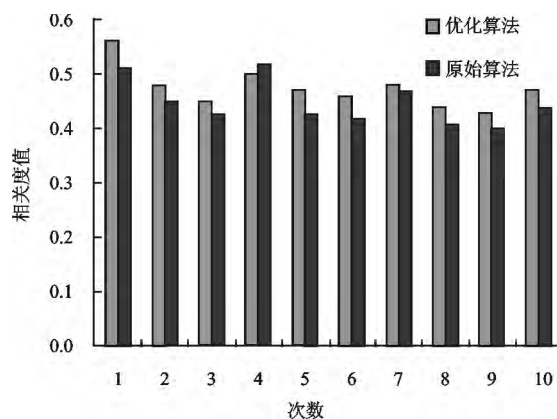


图2 两种排序模型相关度值比较

## 5 结束语

本文针对集成数据环境下 IETM 数据查询结果排序问题,采用历史反馈方法,提出了一种基于历

史反馈的 IETM 查询结果优化排序方法。该方法通过对用户在使用 IETM 查询功能过程中的用户行为进行分析,从历史用户行为反馈和个人用户行为反馈两方面对相关影响因子进行量化,并最终实现查询结果的优化排序。本文结合某型坦克 IETM 系统对所提出的算法进行验证,实验取得了良好的效果。

## 参考文献:

- [1] 徐宗昌. 装备 IETM 研制工程总论[M]. 北京:国防工业出版社,2012.
- [2] 徐宗昌. 装备 IETM 的互操作性与交互性[M]. 北京:国防工业出版社,2014.
- [3] 杨岳湘,苏国荣,邓劲生. 基于用户行为分析的校园网搜索引擎排序方法[J]. 计算机工程,2010,36(24): 275-277.
- [4] 罗成,刘奕群,张敏,等. 基于用户意图识别的查询推荐研究[J]. 中文信息学报,2014,28(1): 64-72.
- [5] Yi J, Maghoul F. Mobile search pattern evolution: The trend and the impact of voice queries[C]// Proceedings of the 20th international Conference Companion on World Wide Web. New York: ACM, 2011:165-166.
- [6] 王伟. 基于数据挖掘的图书馆用户行为分析与偏好研究[J]. 情报科学,2012,30(3): 391-394.
- [7] De Smet W, Moens MF. Representations for multidocument event clustering [J]. Data Mining and Knowledge Discovery, 2013,26(3):533-558
- [8] 杨岳湘,苏国荣,邓劲生. 基于用户浏览时间的点击模型[J]. 华南理工大学学报:自然科学版,2011,39(4): 26-31.

## 作者简介:

呼凯凯 男,(1987-),博士研究生.研究方向为装备保障特性与综合保障. E-mail:hukaikai1987@126.com

徐宗昌 男,(1941-),教授,博士研究生导师.研究方向为装备保障特性与综合保障.