DOI:10.13833/j.cnki.is.2014.11.032

基于用户主题偏好的智能检索算法及实现

周育忠[1],王　平[2]

(1.南方电网科学研究院,广东 广州 510080; 2.武汉大学 信息管理学院, 湖北 武汉 430072)

摘　要: 本文分析了用户对文献的查阅日志及用户间的关联关系,结合电力行业主题范畴表,获取用户的主题偏好。综合考虑检索相关度、用户主题偏好、文献来源权威性分析、引用关系分析等,建立新的排序模型,使结果排序更加准确,从而将与用户需求最相关的文献排到前面,提高检索功能的用户体验。基于lucene 4.3实现智能检索系统,并提供相关主题词提示、主题查询扩展、相关反馈等辅助功能。评测结果表明,该系统在检索满意度和检索效率等方面有显著提升。

关键词: 电力主题范畴;用户主题偏好;综合排序;智能检索;Lucene 4.3

中图分类号: G254.9　　**文献标识码:** A　　**文章编号:** 1007-7634(2014)11-07-06

Algorithm and Realization of Intelligent Retrieval Based on the User Topic Preference

ZHOU Yu-zhong[1], *WANG Ping*[2]

(1.*Research Institute of China Southern Power Grid, Guangzhou* 510080, *China;*
2.*School of Information Management, Wuhan University, Wuhan* 430072, *China)*

Abstract: The paper analyzes the correlation of documents logs and users, combines with the power industry subject category list and acquires the user's topic preference. Considering the retrieval relevance, user preferences, subject analysis, authoritative literature sources and reference relationship analysis, we establish a new scheduling model to make the results more accurate, which will make the most relevant articles of users' demands be on top to improve the user experience of retrieval function. It realizes the intelligent retrieval system based on lucene 4.3 and provides the auxiliary function of related subjects themes, query expansion and relevant feedback. The evaluation results show that the system has significant improvement in retrieval satisfaction and retrieval efficiency etc.

Key words: power subject category; user topic preference; comprehensive ranking; intelligent retrieval; Lucene 4.3

1 引　言

随着社会信息化程度的不断提高以及IT设备的高速发展,信息的存储呈指数上升趋势。而人们对信息的获取要求越来越高。如何利用检索技术快速找到所需的有用信息越来越困难。这在大型企业的信息化进程中得到充分体现[1]。当前,南方电网科研院收录大量论文、标准、专利等文献资源,仅文献的元数据就达到TB级别。一方面,海量的信息资源分布式存储在庞大的数据中心,并且不断更新增长中;另一方面,由于用户单次提交的查询请求描述信息有限,难以全面地体现用户的需求,这导致检索结果难以达到用户的满意程度。通过

收稿日期: 2014-05-04

基金项目: 国家自然科学基金项目(71303179、71073120)

作者简介: 周育忠(1978-),男,广东人,高级工程师,主要从事行业情报系统研发、运维服务与情报资源管理研究。

观察分析用户在较长一段时间内的检索行为，我们发现，在电力行业领域，通过专业主题词可以很好地表达用户的潜在需求[2]。为此我们以专业分类为基础，提炼出主题范畴表，又分析用户日志获取用户个性化偏好信息。

将两者结合，本文提出了基于主题偏好的个性化智能检索模型，并将显性相关反馈与隐性相关反馈技术融入其中，形成了完整的检索解决方案。通过本文智能检索模型的指导，在南方电网情报中心构建智能检索系统，有效的提高了文献资源检索的整体效率。

2 相关工作

用户个性化信息正越来越多地使用到信息检索领域中。相应地，有关用户个性化信息建模技术逐渐成为个性化服务中的基础技术研究。

Fragoudis 和 Likothanassis 对典型的个性化服务系统 LIRA、Letizia 等采用的建模方法进行了分析，指出用户建模在个性化服务系统中的重要地位[3]。

Chan 通过观察用户点击页面中超链接的操作行为，获取用户在页面偏好程度，从而提取训练样本，而后计算 term 间的期望互信息，选择期望互信息大的 250 个 term 来表征用户兴趣模型[4]。在具体应用中，用户个性化信息模型还有许多关键技术需要解决。

利用个性信息挖掘潜在需求，进行查询扩展是提高检索效率的一个重要途径。全局分析是较早出现的具有实际应用价值的查询扩展优化方法，其基本思想是对全部文档中的词或词组进行相关分析，计算每对词或词组间的关联程度。当一个新的查询到来时，则根据预先计算的词间相关关系，将与查询用词关联程度最高的词及词组加入原查询以生成新的查询。目前常见的全局分析方法包括隐式语义索引 (Latent semantic indexing，LSI)[5]、相似性词典[6]等。这些方法的巨大开销使得查询扩展研究向局部分析方向上转移，如局部反馈等[7-8]。

相关反馈作为一种自动扩展查询方法备受关注。其主要思想是检索系统在初始查询到一组样本文档的基础上，根据用户在样本文档中的相关性选择构造出改进的查询表达式，再次进行查询。通过调整检索策略来得到更准确的相关文献。相关反馈技术按照用户是否参与可以分为自动相关反馈和用户相关反馈[9]。前者通过假定检索结果列表的历史文献作为相关文献来进行反馈，不需用户做出相关性判断。后者融入用户参与因素，用户除了对检索出来的文献进行相关性判断外，还可以控制和修改查询。

3 个性化智能检索模型

为实现用户个性化检索，使检索结果更加贴近用户的需求，需要从用户主题偏好的发现与表达、检索资源与主题的关联关系、检索的扩展与个性化排序、检索与反馈的循环推进等多个方面，分析会影响到检索结果的相关因素。在此基础上，综合运用规范的主题词表，建立个性化智能检索模型，以此指导智能检索系统的构建。

3.1 用户的主题偏好模型

用户的查询词输入是用户检索过程中的显性需求。由于查询输入的规范性等因素的影响，仅靠用户输入得到的检索结果不尽如人意。而用户对文献资源的获取存在潜在的主题需求，特别是电力行业领域用户，这种隐性的主题需求更加明显。为此我们运用主题词范畴表对用户需求进行映射，发现用户在文献资源分类上的偏好，从而为智能检索提供良好的基础。主题偏好主要从以下两个方面进行考虑。

(1)用户主题偏好的预定义。

由于本系统主要面对的是企业员工用户，各员工对应着特定的岗位。因此，可以根据用户特征(主要是用户的岗位职能信息以及岗位文献)预先定义一些用户的主题偏好。如高压试验岗位的用户，对电力变压器、断路器、互感器等相关的文献资源有特殊需求[10]，我们从这些岗位文献中提取出主题词，结合岗位职能描述信息，将其映射到规范的主题范畴上，作为用户的需求偏好预定义。我们用向量空间模型的方法来表示用户的主题偏好。

首先，分析主题词分布情况，建立 N 维主题向量空间：

$$[(k_1, w_1),(k_2, w_2), \cdots(k_N, w_N)] \quad (1)$$

其中，k_i 为第 i 个主题词，w_i 为用户在 k_i 上的偏好程度，$i \in 1,2,\cdots,N$。

然后，从用户的岗位职能描述信息，以及岗位文献中，我们可以提取主题词，统计这些主题词的频率 p_{sub_i}，计算其概率分布。主题词频率可以通过

式(2)进行计算，

$$p_{sub_i} = freq_{sub_i} / freq_{sub\_total} \quad (2)$$

其中，$freq_{sub_i}$ 为主题词 sub_i 的词频，$freq_{sub\_total}$ 为主题词集合的总词频。

将 p_{sub_i} 经过一定的系统调整后用来表征用户在各个主题词 sub_i 上的偏好程度，从而得到预定义的用户主题偏好向量，表示为：

$$W_{pre} = (w_1', w_2', \cdots, w_n') \quad (2)$$

其中，$w_i' = \theta \cdot p_{sub_i}$，$i = 1,2,\cdots,n$，表示用户在主题 k_n 上预定义的偏好程度。

(2)从用户操作日志中挖掘用户主题偏好。

用户的检索行为，是用户获取信息的整体行为中的一部分。相关地，还有用户从系统中点击、下载、收藏文献的操作，这些操作都会被记录在系统日志中。我们可以从用户大量的操作日志信息中，挖掘出用户的主题偏好，为智能检索提供基础支撑。为此，我们建立完备的操作日志收集机制，将智能检索系统作为情报中心系统的子系统，从而满足这一要求。

具体地，我们分析日志，获取用户操作文献的集合 $D_{op} = \{d_{op1}, d_{op2}, \cdots, d_{opN}\}$。对 $\forall d_i \in D_{op}$ $(i = 1,2,\cdots,N)$，统计用户对 d_i 的点击、下载、收藏等操作频次，并赋予不同操作权重，加权后计算得到用户对 d_i 的访问频率。根据文献的主题标引，可以得到 d_i 在主题词上的分布，再结合 d_i 的访问频率，即可得到用户在各个主题词上的访问频率，将其作为用户的主题偏好程度，对应到主题向量空间中，从而得到用户的主题偏好向量，表示为：

$$W_{op} = (w_1, w_2, \cdots, w_n) \quad (3)$$

通过将以上两种主题偏好进行加权，从而确定用户的主题偏好，即：

$$W = \alpha \cdot W_{pre} + \beta \cdot W_{op} \quad (4)$$

注意，根据日志分析得到用户偏好是随着时间变化的，需要建立相应的更新机制。

3.2 基于主题词的查询扩展模型

查询请求是用户查询需求的直接反应，其中蕴含着潜在的主题需求，这种主题需求在一定程度上反应了用户对所需文献的抽象和概括，更能反映用户的需求。同时主题词可以作为文献资源的标记，反应了文献的内容核心及分类信息，能更好地表达文献的本质。综合这两方面进行考虑，我们选择主题词进行查询扩展，必会从很大程度上提升检索的功效。

如果用户的检索时输入的直接就是规范的主题词，我们可以通过主题范畴表中的上位词、下位词等关联关系，找到相关的主题词进行查询扩展。但很多时候，用户输入的查询请求与潜在主题需求，两者没有显性的关联。这时，我们通过历史检索文献，以及主题标引文献为其建立关联关系。基本思想如下。

记用户检索请求Q对应的文档集合为：$D_{query} = \{d_{q1}, d_{q2}, \cdots, d_{qN}\}$。通过对 D_{query} 中各个文档进行分词，得到一组 *Term* 集合，记为 $T_{query} = \{t_{q1}, t_{q2}, \cdots, t_{qN}\}$。对 $\forall t_{qi} \in T_{query}$（$i = 1,2,$，统计概率 $p_{t_i} = freq_{t_i} / freq_{total}$，得到 D_{query} 对应的集合 T_{query} 的概率分布，记为 $F_{query} = (p_{qt_1}, p_{qt_2}, \cdots, p_{qt_N})$。其中，$freq_{t_i}$ 为 t_{qi} 的词频，$freq_{total}$ 为 T_{query} 中 *Term* 的词频总数。

对于主题向量空间的主题词，我们通过文献的主题标引，也可以得到一组文档集合，记为 $D_{subject} = \{d_{s1}, d_{s2}, \cdots, d_{sN}\}$。类似地，我们通过文档集合获取词条集合，再通过相应词频的计算，可以得到 $D_{subject}$ 对应的词条集合的概率分布，记为 $F_{subject} = (p_{st_1}, p_{st_2}, \cdots, p_{st_N})$。

在获取了这两方面的概率分布后，我们可以通过计算概率分布的相似性，找到与检索词最相关的主题词，进而用来做主题词的查询扩展。

在计算检索词和主题词对应的两组文档的概率分布相似性时，我们考虑使用Kullback-Leibler散度(Kullback-Leibler Divergence的简称，也叫做相对熵，Relative Entropy)进行计算。

这样，通过 $D_{KL}(F_{subject} \| F_{query})$ 即可计算出 $F_{subject}$ 相对于 F_{query} 的概率分布差异，取差异较小的主题词构建查询扩展。

$$D_{KL}(F_{subject} \| F_{query}) = \sum p_{st_i} \log \frac{p_{st_i}}{p_{qt_i}} \quad (5)$$

为获取更好的查询扩展效果，我们研究了查询请求和主题词在系统收录的文档向量上的分布情况，据此对上述计算进一步优化，选择Jensen-Shannon散度来平滑计算，通过计算 $D_{JS}(F_{subject} \| F_{query})$ 来衡量 $F_{subject}$ 和 F_{query} 的相互差异。

$$D_{JS}(F_{subject} \| F_{query}) = \frac{1}{2} D_{KL}(F_{subject} \| R) + \frac{1}{2} D_{KL}(F_{query} \| R)$$

(6)

其中，$R=\frac{1}{2}\left(F_{subject}+F_{query}\right)$。

当选择概率分布差异较小的主题词后，我们以一定权重将其加入检索向量中，构建扩展的查询向量，以提高检索效率。

3.3 个性化检索排序模型

在文档相关度排序的基础上，考虑用户的主题偏好进行加权排序，是个性化检索排序的核心。从用户主题偏好模型中，获取用户的主题偏好向量 W。对于检索得到的文档集合，我们可以根据文献主题标引情况，获取每篇文档的主题分布向量 $V=\left(v_1, v_2, \cdots, v_n\right)$。这样，我们可以通过计算 W 和 V 的向量相似度 $sim(V, W)$，来评判检索到的文档在用户偏好的主题上的得分。$sim(V, W)$ 计算值高的文档，其偏好得分也较高。

$$sim(V, W)=\frac{\sum_{k=1}^{n} v_k \times w_k}{\sqrt{\left(\sum_{k=1}^{n} v_k^2\right)\left(\sum_{k=1}^{n} w_k^2\right)}} \tag{7}$$

在考虑了用户主题偏好加权之余，文献的质量也是一项重要的加权指标。文献质量的评价因素有很多，本文主要从论文被引用频次、被下载的频次、发表期刊级别、是否为自建资源这4个方面的因素，对文献进行加分评价。其中自建资源主要是考虑本单位通过向资源商购买和自行采集两种方式收集文献资源。而根据专业自行采集的资源经过了人工审核，故具有较高的质量。各因子所占权重见表1。

表1 文献质量评判因子权重表

因子	$f_{引用}$	$f_{下载}$	$f_{期刊}$	$f_{自建}$
权重	0.5	0.1	0.2	0.2

通过文献元数据中相关字段的归一化计算，得出文献各因子的得分。加权后得到文献的质量评分，如式(8)。

$$G_{factor}=0.45\cdot f_{引用}+0.15\cdot f_{下载}+0.2\cdot f_{期刊}+0.2\cdot f_{自建} \tag{8}$$

通过对以上两方面得分以及文档与检索相似度得分的加权，计算检索结果文档的终排序得分，如式(9)。计算过程考虑不同权重的设置，具体根据系统使用情况和文献分布情况进行确定。

$$G_{sort}=\alpha\cdot G_{query}+\beta\cdot sim(V, W)+\gamma\cdot G_{factor} \tag{9}$$

3.4 结合主题的相关反馈模型

相关反馈作为检索请求的补充，可以有效提高检索的准确性。本文将相关反馈和伪相关反馈相结合，并通过主题范畴分类和用户的操作日志分析，有效界定相关文档和不相关文档的范围，从而使反馈达到更优的效果。

用户在一次检索之后，对检索结果进行相关性标注。我们根据用户的标注情况，建立相关文档向量集合 D_r 和不相关文档向量集合 D_{nr}。在获取相关文档和不相关文档之后，我们可以考虑在Rocchio算法思想的指导下，建立相关反馈检索向量，如公式(10)，

$$\vec{q}_m=\alpha\vec{q}_0+\beta\frac{1}{|D_r|}\sum_{\vec{d}_j\in D_r}\vec{d}_j-\gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j\in D_{nr}}\vec{d}_j \tag{10}$$

其中，$\vec{q}_0$ 是原始的查询向量，D_r 和 D_{nr} 是已知的相关和不相关文档集合。α、β、γ 是相应权重。

但是在本系统的使用场景下，直接使用上述公式，相关反馈效果无法达到最优。我们考虑从以下两个方面对模型进行改进：相关文档集合 D_r 及不相关文档集合 D_{nr} 的界定与过滤、反馈的文档向量与主题偏好向量相结合建立反馈后查询向量。

考虑到用户在一次检索之后，对文档的反馈标注操作有限，我们需要从用户检索历史和主题兴趣分布的角度出发，帮助界定哪些是相关文档，哪些是不相关文档。用户直接标注和判定的文档的相关性即为显式相关反馈，这部分是相关反馈的基础，在相关反馈计算中赋予较高的权重。而检索结果Top-N中，用户未标注的文档，可以通过计算文档主题向量与用户偏好主题向量的相似性，取相似性高的加入相关文档中，相似性低的加入不相关文档中，这两部分的文档在用户相关反馈计算时，可以考虑用偏好主题相似性评分作为其权重 l_j。这样，在缓解用户操作负担的同时，有效获取反馈检索所需的文档集。

在确定了 D_r 和 D_{nr} 的文档范围后，我们记 $D_r=\{\vec{d}_{r1}, \vec{d}_{r2}, \cdots, \vec{d}_{rN}\}$ 为相关文档向量集合，记 $D_{nr}=\{\vec{d}_{nr1}, \vec{d}_{nr2}, \cdots, \vec{d}_{nrM}\}$ 为不相关文档的向量集合。对 $\forall\vec{d}_i\in\{D_r, D_{nr}\}$，我们取高频词条及其词频，建立文档向量，记为 $\vec{d}_i=\left(freq_{t1}, freq_{t2}, \cdots, freq_{tN}\right)$。其中，$freq_{ti}$ 为文档中的词频。

在确定了反馈文档向量后，我们进一步对其权

重进行调整。用户直接标记的文档权重赋1，其它文档根据文档主题向量与用户主题偏好向量相似性评分来计算。从而将反馈文档以相应的权重加入到反馈检索向量中。同时也将用户的主题偏好向量以权重 δ 加入到反馈向量中。根据使用统计分析，δ 取0.2–0.3之间效果较优。另外，由于不相关文档主要是系统自动从用户未标注的文档中挑选的，不确定性高。为加强反馈检索的稳定性，我们通过相似性计算取最不相关的文档代表 D_{nr}，加入到计算中。即不相关文档集合中只取 $\arg\max_{\vec{d}\in D_{nr}} sim\left(\vec{d},\vec{q}_0\right)$ 进行计算。

综合以上考虑，得到改进的反馈查询公式(2)进行反馈检索的查询扩展。

$$\vec{q}_m=\vec{q}_0+\sum_{\vec{d}_j\in D_r} l_j\vec{d}_j+\delta\cdot W-\arg\max_{\vec{d}\in D_{nr}} sim\left(\vec{d},\vec{q}_0\right) \quad (4)$$

其中，$\vec{q}_0$ 是原始的查询向量，D_r 和 D_{nr} 是已知的相关和不相关文档集合。l_j 是各个相关文档的权重。W 为用户的主题偏好向量，δ 为 W 的权重。通过该公式计算得到查询扩展进行二次反馈检索，提高检索准确率和召回率。

4　系统设计及验证分析

4.1 个性化智能检索系统设计

作为南方电网情报中心系统的子系统，智能检索系统充分利用全方位收集的用户日志信息及主题词范畴表，深度挖掘用户的需求偏好，并以此为支撑，实现用户个性化检索的需求，提高检索的准确性和满意度。系统采用Lucene 4.3作为底层检索技术，提供统一检索入口、智能提示、检索纠错、相关反馈、相关主题提示等功能，构建个性化智能检索系统。具体从以下几点进行系统设计：

(1)系统分析用户操作日志，按主题分类统计对应的点击、下载、收藏等操作次数，并按操作类型的权值计算各个主题的访问热度得分，作为用户对主题的偏好程度。该计算涉及大数据量的日志分析，单机运行难以支撑。系统使用Hadoop平台，通过MapReduce分布式计算，实现日志的分析。

(2)当用户提交检索请求，系统使用ICTCLAS分词器对检索语句进行分词。通过Jensen–Shannon散度衡量方法计算检索分词与主题词之间的相关度，取相关度高的主题词提示给用户，帮助用户更加明确地表示自己的检索需求。

(3)系统提供多种排序接口。在综合排序中，以文档与检索词的相关度作为排序的基础。考虑用户主题的偏好，将其表示为主题偏好向量。计算文档在主题上的空间向量与用户主题偏好向量的距离，作为文档在用户偏好上的加权得分，累加到总体排序得分中。

(4)计算文档的质量评分。对文档的引用频次、下载频次、期刊影响因子等分别进行归一化处理，乘以对应的权重后，得到文档的质量评分，再以一定权重累加到总体排序得分中。

(5)对于第一次排序结果，由用户标记指出哪些是相关的文档。从用户翻看过的结果页面中，收集未标记的文档，作为初始不相关文档。再根据用户日志分析的概率结果，从中过滤掉意向不明的文档。对挑选出来的相关文档和不相关文档，通过公式(4)进行查询扩展，进行二次反馈检索，进一步聚焦到用户最想要的检索结果。

个性化智能检索系统设计的整体流程如图1所示：

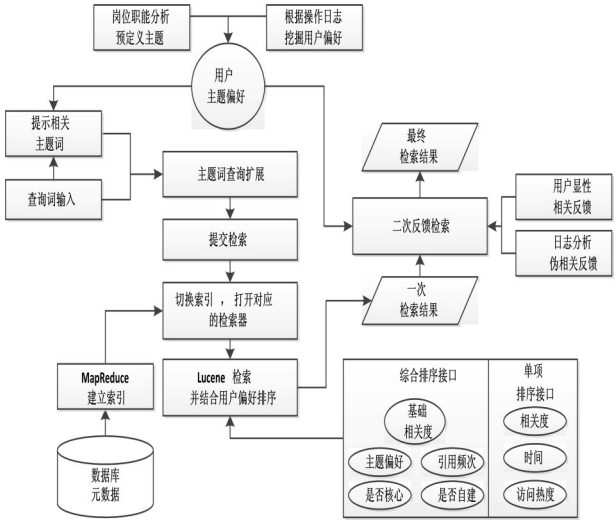

图1　智能检索系统整体流程

其中对于相关反馈模块，系统考虑用户的显式相关反馈和隐式相关反馈两个方面，综合运用，保障反馈有效性的同时尽量减轻用户交互的负担。首先系统获取用户指定的相关度高的文档，从Lucene索引中提取相应的文档词频向量，取权重高的词项，形成备选的扩展查询向量，并根据用户对这些文档的主题偏好，赋予相应的权重。然后根据用户在检索结果中的翻页情况，将用户未标记的文档作为初始不相关文档。

通过偏好相关度计算，只取偏好较低的前20篇

文档,从而缩小不相关文档的范围。最后根据公式(4),进行扩展后检索相向的构建,将其提交后,利用Lucene检索接口实现二次反馈检索。反馈检索可以根据用户需要循环推进,逐步逼近用户满意的检索结果。相关反馈流程如图2所示:

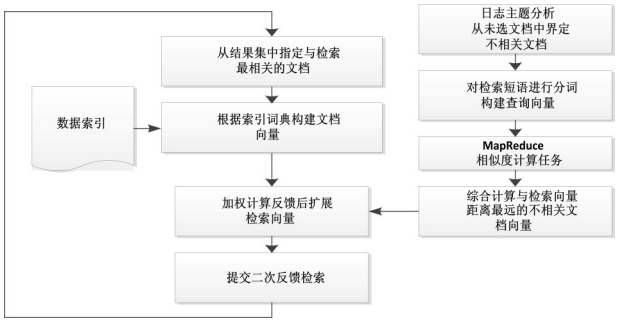

图2 相关反馈具体流程

4.2 检索效果及性能分析

将本文提出的引入用户主题偏好的个性化智能检索,与普通相似度检索进行实验对比,从检索的用户满意度,以及检索的召回率和准确率上,分析本系统的检索效果。

在用户满意度方面,我们主要统计南方电网情报中心旧版系统的检索日志中,用户提交查询后,在检索结果列表中平均点击文献的次数 Num'_{click} 及被点击文献的平均驻留时间 $Time'$。相应的,在本系统的试用期内,统计用户对个性化智能检索结果的平均访问次数 Num_{click} 及结果平均驻留时间 $Time$。我们以 $S=Num_{click}*Time$,即用户对检索结果的关注时间,来反映用户的满意度。用户停留在检索结果中获取信息的有效时间越长,表示对检索结果越满意。经对比发现,本系统个性化智能检索的 S 值比旧版系统中普通检索有60%的提升。

表2 普通检索与基于主题偏好检索的效果对比

	普通检索	基于主题偏好个性化检索
10%	0.6619	0.7001
20%	0.5431	0.5994
40%	0.431	0.4931
60%	0.317	0.3792
80%	0.2173	0.2501
100%	0.1007	0.1450

在召回率和准确率实验分析方面,需要先建立测试数据集。我们从南方电网情报中心系统中,收集用户的检索日志记录,通过MapReduce任务提取特征明显的检索请求,并从其检索结果中获取用户点击驻留时间长,有下载、收藏操作的部分作为检索结果,构成测试数据集。利用Lucene 4.3对南方电网现有的外购资源数据及自建文献数据进行统一的倒排索引构建。将测试检索请求提交到系统,在该索引中进行检索测试。实验结果对比如表2所示。

通过对比不同召回率下的检索准确率,我们发现使用了用户主题偏好的检索在检索精度上有明显提升。这是因为南方电网内部用户在文献检索时,结合自身岗位需求,有明显的主题偏好。将这种偏好融入到查询请求中,建立查询扩展和相关反馈机制,检索结果将更能满足用户需求。

在引入用户主题偏好后,检索排序将产生额外的计算量,检索时间性能会受到影响。但由于文档的主题标引计算,用户的主题偏好计算等,是通过MapReduce分布式计算在后台预处理中进行的,并在用户检索前已经通过特殊的格式存储并加载到Cache中。检索的额外计算主要集中在文档主题分布向量和用户主题偏好向量的相关度计算中。该部分计算时间与Lucene的检索相似度计算时间相当,故检索的总时间影响已被降到最低。经实验统计,普通检索平均耗时436ms,引入主题偏好后,检索平均耗时为953ms。检索时间在用户可接受的范围内。

5 结 语

用户主题偏好的引入从很大程度上改进了检索系统的性能,这一点通过本系统的实施得到验证。本文提出的基于用户主题偏好与文献质量的综合检索排序,以及将相关反馈技术与个性化信息相结合的检索模型,充分考虑了面对专业领域中检索运用的关键问题,并设计了高可用的框架和整体解决方案,使个性化信息与相关反馈技术能够更好地发挥其作用。目前,在用户个性化信息的识别与筛选方面,主要依靠对用户的操作日志的分析。从语义角度分析用户兴趣,使得系统对用户需求有更加精准的把握,是我们未来的主要研究方向。另外,日志的分析是后台异步执行的,如何在实时检索中有效利用个性化信息也有待进一步深入研究。

参考文献

1 龚 婷, 周育忠, 韦嵘晖, 王庆红. 南方电网技术情报中心系统的研发与应用[J]. 南方电网技术, 2012,6 (1):74-77.
2 陈锦攀,等. 广东电网公司竞争情报体系建设初探[C]// 2011年度广东省科学技术情报学会综合研讨会, 2011.
3 Fragoudis D. User Modeling in

(下转第18页)

上看,在发展效果指数上,湖北省处于中下游水平,4项指标全部落后于全国平均水平。值得注意的是拥有高校数量全国排名第3的湖北省,而在百万人拥有专利数上却掉到了第13名,落后于全国平均水平,而江苏在这一指标上全国排名第1。在社会资金周转率方面,湖北省更是排在在第23名,不仅落后于全国平均指数,甚至也落后于中部平均水平,表明湖北省经济运行质量需要进一步提高。

表6 发展效果指数三级指标比较

指标	发展效果指数					
	湖北		江苏		中部平均指数	全国平均指数
	指数	排名	指数	排名		
人均国内生产总值	1.04	13	2.64	4	0.72	1.28
百万人均拥有专利数	0.44	13	4.00	1	0.36	0.80
综合能耗产出率	1.52	18	2.84	4	1.60	1.60
社会资金周转率	1.20	23	0.92	25	1.48	1.76

4 结 语

信息化评价是信息化建设的重要任务之一,它为信息化建设的量化分析和科学管理提供了依据和手段。但由于信息化发展的疾进及新兴信息技术的深入运用,势必会对信息化评价带来较大的冲击,这就需要我们对信息化评价指标体系进行不断的优化与完善,既要纲举目张,抓住信息化的本质与核心;又要见微知著,体现信息化发展的新内容、新趋势,以构建全面、客观、真实的信息化测度体系,引导我国区域信息化建设的持续、健康发展[10]。

参考文献

1 国家统计局统计科研所信息化统计评价研究组.信息化发展指数优化研究报告[J].管理世界,2011,(12):1-4.
2 邓小昭,邬晓鸥,韩 毅,樊志伟. 论信息化指标体系研究中的几个理论问题[J].情报学报,2003,(1):99-100.
3 任剑婷,李瑜婷. 对我国信息化测度的建议[J].图书情报工作,2011,(8):26-27.
4 张文娟.中外社会信息化测度方法研究[J].情报科学,2009,(6):954-956.
5 汪卫霞,汪 雷.社区信息化评价:模型构建与分析--以安徽省合肥市为例[J].情报理论与实践,2011,(12):96-98.
6 郭春丽,王 健. 基于国家信息化水平指数分析河北省信息化水平[J].情报科学,2006,(4):515-517.
7 孙艳丽.高校档案信息化建设的影响因素分析及对策研究[J].情报科学,2012,(2):255-257.
8 王 明.对我国信息化水平测度方法的思考[J].情报杂志,2005,(1):96-98.
9 刘文云,葛敬民. 国内外信息化水平测度理论研究比较[J].情报理论与实践,2004,(8):145-146.
10 郑建明.信息化指标构建理论及测度分析研究[M].北京:中国社会科学出版社,2011:150-153.

(责任编辑:赵立军)

(上接第12页)

Information Discovery[C]//Proceedings of Advanced Course on Artificial Intelligence (ACA199), Greece, 1999.
4 Chan Philip K. A Non-invasive Learning Approach to Building Web User Profiles[C]//Proceedings of KDD-99 Workshop on Web Usage Analysis and User Profiling. New York:ACM Press, 1999: 7-12.
5 Deerwester S, Dumai S T, Furnas G W, et al. Indexing by Latent Semantic Analysis[J]. Journal of ACM Transactions on Information Systems, 2000,18(1):79-112..
6 Qiu Y, Frei H. Concept Based Query Expansion[C]// Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1993. 160-169.
7 Buckley C, Salton G, Allan J, Singhal A. Automatic Query Expansion Using SMART[R]. Technical Report, TREC-3, 1995: 69-80.
8 Ricardo B-Y, Berthier R-N. Modern Information Retrieval[M]. England: Pearson Education Limited, 1999:89.
9 He D Q. A Study of Self-Organizing Map in Interactive Relevance Feedback[C]//Proceedings of the 3rd International Conference on Information Technology. New Generations, Las Vegas, IEEE. 2006:394-401.
10 郑阿花. 电力企业竞争情报系统建设探讨[J]. 湖南电力, 2010, 30(5):17-20.

(责任编辑:赵立军)