基于用户偏好的垂直搜索算法

张 磊，陈俊亮，孟祥武，沈筱彦，郭 杰

(北京邮电大学网络与交换技术国家重点实验室 北京 海淀区 100876)

【摘要】提出并研究、实现了基于用户偏好的垂直搜索算法(PVSA)。以领域特征为基本出发点，PVSA借助领域主题偏好向量、领域元数据权重因子、检索名词差异化、行业词典库更新等4项策略，有效地挖掘、表征用户的领域个性化偏好，以此为基础构建基于用户偏好的垂直搜索算法。实验结果表明了PVSA算法的有效性和可行性。

关 键 词 词库； 差异化； 领域主题偏好向量； 元数据权重因子； 用户偏好

中图分类号 TP393 **文献标识码** A doi:10.3969/j.issn.1001-0548.2010.01.021

User Preference-Based Vertical Search Algorithm

ZHANG Lei, CHEN Jun-liang, MENG Xiang-wu, SHEN Xiao-yan, and GUO Jie

(State Key Laboratory of Networking and Switching, Beijing University of Posts and Telecommunications Haidian Beijing 100876)

Abstract Personalized search and vertical search are receiving more and more attention of users. User preference-based vertical search algorithm (PVSA) is proposed in this paper. By focusing on domain characteristics, PVSA uses domain topic preference vector, domain metadata weight factors, the strategy of distinguishing weights of input terms, and industry lexicon update to mine different domain preferences of different users. Experimental results show that the proposed algorithm is feasible and effective in mining users' personal preferences.

Key words dictionaries; differentiation; domain topic preference vector; metadata weight factors; user's preferences

Google通用搜索主题宽泛的特点使领域用户很难快速筛选出所需内容。垂直搜索(领域搜索、行业搜索)以其高度专业化的优势日益受到重视。不同用户具有不同领域偏好，在检索输入相同的前提下为所有用户提供统一的检索结果难以满足个性化检索需求，因此，提供个性化的搜索服务成为当前研究的重点。Google(http://www.google.com/coop/cse/)基于主题/领域特性为用户提供个性化搜索，但用户并不情愿提供偏好，仅靠用户提供的几个关键词也难以准确描述其喜好。饭统网是国内较著名的垂直搜索网站，但相关引擎主要基于数据库方式，即使具备全文检索能力，其结果也不甚理想。因此，以个性化服务为出发点，将个性化技术引入垂直搜索领域，基于域特征构建个性化垂直搜索算法(PVSA)，能较好地满足用户针对领域搜索的个性化需求。文献[1]以“社区”为搜索引擎引入个性化能力，挂载网页至不同社区，若用户访问过该社区中的某些网页，则认为用户对该社区及归于该社区的网页感兴趣。针对某次检索，计算网页得分除了采用pagerank，还考虑用户网页兴趣，该兴趣根据访问日志基于社区获得。文献[2]在元搜索中提出主题的概念，其含义与文献[1]中的“社区”类似，都是网页分类的类别标准。类似于文献[1]，若用户经常访问“狩猎”网页(狩猎即主题/社区)，当其检索“山鹿”时，在山鹿的多张网页中，猎鹿的网页应该是用户更感兴趣的。

不同的心理特点决定了用户对不同主题具有不同的兴趣。文献[1-2]以引入领域主题喜好至个性化搜索的基本出发点，基于域特征，PVSA从4个视角提升个性化域检索精度。电信级用户的行业搜索是PVSA的主要应用场景，在相对有效的基础上，算法会在性能和效率间取得折中。

1 PVSA构建个性化域搜索的4个基本策略

1.1 基于领域主题喜好(向量)的网页差异化策略

从电信级应用出发，采用直观的轻量级策略完

收稿日期：2008 – 08 – 14； 修回日期：2009 – 02 –25

基金项目：国家973重点基础研究发展规划(2007CB307100)；国家自然科学基金(60872051)；国家科技支撑计划重大项目(2006BAH02A11)；北京市教委产学研项目(zh100130525)

作者简介：张 磊(1980 –)，男，博士生，主要从事现代信息检索、个性化技术等方面的研究.

成基于主题喜好的个性化域偏好的挖掘和应用。以文献[1-2]为考虑起始点，基于领域特征，提出了在PVSA中的领域主题偏好策略：基于域网页日志挖掘行业主题偏好，根据领域(主题)兴趣提升域网页权重。PVSA中的基于域主题喜好的网页差异化策略是一种相对有效且高效的域兴趣分析、应用方法。

为方便理解，不采用xml和xml schema本身使用的标记及命名空间等术语。模式对实例进行约束，使实例具有一致规范。以餐饮域为例，任一酒店(xml文档)一般具有酒店名、地址等属性(可将这些属性简单理解为领域元数据，原始网页中相同域属性可能有不同属性名)。将餐饮域网页的schema简单标记为由“属性”构成的多元组(省略了与问题说明相关性不大的部分)。下面首先从总体上给出问题说明，然后依次给出具体的主题喜好度计算等。

设 $D=\{d_1,d_2,\cdots,d_n\}$ 为领域集， $\forall d_i\in D$ ，基于领域特征，定义域xml schema为 $X_i=\langle x_{i,1},x_{i,2},\cdots,x_{i,z}\rangle$ 。在少量人工干预下，使用正则式意义下的模式匹配思想[3-4]，将抓取的网页集转换为满足 X_i 的xml实例集 $P_i=\langle p_{i,1},p_{i,2},\cdots,p_{i,m}\rangle$ ，根据网页日志挖掘，得到任意用户 U 的网页喜好度 $FP_i^u=\langle\theta^u(p_{i,1}),\theta^u(p_{i,2}),\cdots,\theta^u(p_{i,m})\rangle$ ，且 $\sum_{p_{i,j}\in P_i}\theta^u(p_{i,j})=1$ 。

基于 d_i 的领域特征，提取领域主题向量 $\boldsymbol{T}_i=\{t_{i,1},t_{i,2},\cdots,t_{i,n}\}$ ， $\forall t_{i,k}\in\boldsymbol{T}_i$ ，相关联的网页集 $P_{i,k}=\{p_{i,1}^k,p_{i,2}^k,\cdots,p_{i,w}^k\}\subset P_i$ 。设 U 为目标用户，基于 U 的使用日志和关联等信息，得到 U 相对 $\boldsymbol{T}_i$ 的主题喜好向量 $\boldsymbol{FT}_i^u=\langle\vartheta^u(t_{i,1}),\vartheta^u(t_{i,2}),\cdots,\vartheta^u(t_{i,n})\rangle$ ，且 $\sum_{t_{i,k}\in T_i}\vartheta^u(t_{i,k})=1$ 。

以北京餐饮域为例，给出网页和主题喜好度的挖掘和应用策略。记 $x_{i,1},x_{i,2},\cdots,x_{i,z}$ 为餐饮域中酒店地址、消费价格及菜系等元数据，元数据的实例值(属性值)通俗地讲就是不同酒店的具体地址、消费价格等。首先，采用人工方式基于不同领域元数据的不同实例值提取领域主题 $t_{i,1},t_{i,2},\cdots,t_{i,n}$ ，包括东城区、海淀区、低档消费、中档消费、川菜及粤菜等。针对已有主题词，系统自动挂载域网页至相关主题词。不同于通用搜索根据整个文档进行关键词匹配的思想。基于领域特征，垂直搜索以领域元数据为基本出发点，针对任意 $t_{i,k}\in T_i$ 和 $p_{i,j}\in P_i$ ，使用 $p_{i,j}$ 的特定属性域的属性值与 $t_{i,k}$ 进行关联匹配，确保高效的同时，有效地去除“无关”属性域可能引入的噪音，显著提升网页与主题词的关联精度。

根据Web日志， $\forall p_{i,j}\in P_i$ ，假设已经得到 U 对 $p_{i,j}$ 的未归一化喜好度 $c_{i,j}^u\in C_i^u=\langle c_{i,1}^u,c_{i,2}^u,\cdots,c_{i,m}^u\rangle$ ，则归一化后的网页喜好度为：

$$\theta^u(p_{i,j})=\frac{c_{i,j}^u}{\sum_{c_{i,l}^u\in C_i^u}c_{i,l}^u}$$

类似于文献[1-2]，可以较直观地看到，从统计意义上讲，用户对主题下网页的喜好表征了用户对相关主题的喜好。根据基于域特征得到的与该主题关联的网页集 $P_{i,k}\subset P_i$ ，基于 U 归一化后的网页喜好 $\forall t_{i,k}\in\boldsymbol{T}_i$ ，不难得到归一化的用户主题喜好为：

$$\vartheta^u(t_{i,k})=\frac{\sum_{p_i^k,j\in P_{i,k}}\theta(p_{i,j}^k)}{\sum_{t_{i,e}\in T_i}\sum_{p_{i,h}^e}\theta(p_{i,h}^e)}$$

一些研究认为，对网页的点击次数可基本表征用户的网页兴趣。引入类“时间戳”概念，以差异化不同阶段使用系统对主题偏好表征的影响。设 U 共进行过 τ 次检索， U 对 $p_{i,j}$ 的初始喜好(未归一化)为：

$$c_{i,j}^u=c_{i,j}^u(\tau)=\begin{cases}c_{i,j}^u(\tau-1)+\lg(\tau+\varsigma) & \text{用户第}\tau\text{次使用系统点击过网页}p_{i,j}\\ c_{i,j}^u(\tau-1) & \text{否则}\end{cases}\quad(1)$$

式中 $c_{i,j}^u(\tau-1)$ 表示截至第 $\tau-1$ 次检索完成后用户 U 对域网页 $p_{i,j}$ 的喜好度； ς 为偏移常数。

基于个性化的主题喜好度差异化Web网页，可得到个性化评分输出。针对一次检索，设 $p_{i,j}$ 的初始得分(score)为 $\delta_{i,j}$ ，则考虑用户领域主题偏好的 $p_{i,j}$ 得分(score)为：

$$\xi_{i,j}=\delta_{i,j}(1+\sum_{t_{i,k}\in\boldsymbol{T}_i}(\iota_{i,k}\cdot\vartheta^u(t_{i,k})))^{\frac{1}{3}}\quad(2)$$

式中 $\iota_{i,k}=\begin{cases}1 & p_{i,j}=p_{i,r}^k\in P_{i,k}\\ 0 & p_{i,j}=p_{i,r}^k\notin P_{i,k}\end{cases}$ 。

1.2 基于领域元数据提升域需求同检索输出关联度

海量Web网页不具备统一的元数据格式，故传统通用搜索不可能区分、加权特定元数据，最多只能在预处理时，针对若干张特定网页由人工指定不同网页的权重。针对领域搜索，不难发现，特定领域具有领域相关的领域本体，特定领域的领域特性使得能够定义统一的领域xml schema，从而较传统通用搜索，可以以元数据而非整个文档为基本单元，

更加细致、有效地进行网页——服务需求的关联匹配。需要指出，领域元数据方式是域网页信息组织的有效方式，对于构建领域信息平台将发挥重要作用。不难得到，不同属性域取值中的相同名词相对服务请求的重要性是不同的(同一个名词出现在名称域中和其出现在描述域中，前者能够更大程度地去表征一篇文档)，故为不同元数据(属性域)引入不同的权重因子以差异化领域元数据，可更有效地映像用户的领域服务需求，提升检索得到的目标网页与检索需求的匹配度。

具体而言，针对特定领域d_i，设基于领域特征预定义的领域xml schema为$X_i=\langle x_1,x_2,\cdots,x_z\rangle$(同样，本文的xml schema仅定义至属性元数据层面)，定义权重向量$\boldsymbol{W}_i=\langle w_1,w_2,\cdots,w_z\rangle$满足$w_k\geqslant 1$，通过差异化不同的领域元数据权重完成服务需求的细粒度镜像。

1.3　基于用户偏好的检索名词差异化策略

不同用户具有不同域偏好，相同检索下不同用户的侧重可以是不同的。以检索“小肥羊/水煮鱼”为例，很多用户主要关注“小肥羊”，而“水煮鱼”是附带信息，但也有用户对“小肥羊”和“水煮鱼”赋予基本相同的关注兴趣。如果用户经常检索“小肥羊”而很少查询“水煮鱼”信息，那么相对于该检索中的不同名词，该用户通常会对“小肥羊”赋予更高的兴趣度。相应地，针对TF-IDF计算得到Score相同的网页，假设网页1侧重于“小肥羊”，而网页2侧重于“水煮鱼”，则认为前者的PageRank要高于后者。有效地差异化检索中不同领域名词的权重，是个性化搜索的重要组成部分。

依统计意义，用户经常输入的行业名词是其比较感兴趣的，对这类词的喜好度要高于比较少输入或不输入的行业名词。提升用户经常检索的行业名词的权重，是PVSA个性化算法中提出的基本策略，并希望进一步地优化。在餐馆用餐时，用户一般会先看菜谱，菜谱除了说明能够提供的菜肴外，就用户而言，可以从中发现自己喜欢吃的东西，而喜欢的若干菜肴在没看菜谱之前用户未必能够想到；在菜市场、商场购物等都是类似的例子。用户检索输入是一个类似的过程，不能要求用户将行业名词全部浏览、过滤一遍，然后再进行检索输入；用户检索的菜肴是其比较喜欢的，但基于行业名词的海量性，若干比较喜欢的行业名词用户未必能够想到。换言之，针对具有偏好的所有行业名词，实际的检索输入往往针对用户想到的那部分其有偏好的名词展开。针对用户可能喜欢但没有给出过检索输入的名词，希望在初始值0的基础上适当提高这些名词的权重。PVSA的执行策略是引入个性化思想，根据实际检索输入提升用户兴趣名词权重，并在此基础上进行问题优化。

除了可以将地址名词库中的信息归入不同类别，领域名词的范畴可稍加扩展。还是以餐饮域为例进行阐述，基于多个行业词典库等先期工作，领域名词内部被划分为酒店名、菜肴名等不同“类别”。为简单起见，不强调“类别”的概念，但接下来的过程实际针对任一特定“类别”下的域名词展开。本文的基本策略是根据用户的兴趣偏好差异化不同检索名词的权重，用户比较感兴趣的域名词在检索输入中应具有相对更高的权重。为有效挖掘领域名词偏好，将同一“类别”下的域名词划归至两个不同的类型中，类型Ⅰ中域名词是用户历史检索中出现过的名词，基于检索次数等要素获取用户对这类名词的喜好；类型Ⅱ中域名词是检索日志中没有出现过的名词，基于域名词的海量性，很难简单地认为用户对这类名词不感兴趣。假设U经常检索水煮牛肉、白灼基围虾等行业名词，相对于其检索日志中没有出现过的盐水皮皮虾、水煮鲤鱼等，U完全可能具有相关偏好，有效挖掘用户对类型Ⅱ中名词的兴趣度(对Ⅱ中名词喜好度挖掘针对菜谱等“类别”而非所有“类别”展开，其余“类别”将自动跳过)，在提升个性化检索能力方面具有重要意义。

对于类型Ⅰ中的域名词，同样引入类“时间戳”概念，基于检索次数给出个性化偏好。设用户U共进行过τ次检索输入，类型Ⅰ中域名词集合记为$K_1^u=\langle k_{1,1}^u,k_{1,2}^u,\cdots,k_{1,s}^u\rangle$，$\forall k_{1,j}^u\in K_1^u$，$U$对$k_{1,j}^u$的初始兴趣度$g_{1,j}^u\in G_1^u=\langle g_{1,1}^u,g_{1,2}^u,\cdots,g_{1,s}^u\rangle$为：

$$g_{1,j}^u=g_{1,j}^u(\tau)=\begin{cases}g_{1,j}^u(\tau-1)+\lg(\tau+\zeta) & \text{用户第}\tau\text{次使用系统输入过名词}k_{1,j}^u\\ g_{1,j}^u(\tau-1) & \text{否则}\end{cases}$$

式中　$g_{1,j}^u(\tau-1)$为U第$\tau-1$次使用过系统后得到的对$k_{1,j}^u$的兴趣度；，ζ为偏移常数。若用户在前$\tau-1$次都没有检索过$k_{1,j}^u$，则$g_{1,j}^u(\tau-1)=0$。类型Ⅰ中域名词是用户检索兴趣的最为直接的表征，可以简单视为已知的用户检索兴趣。基于已获得的检索兴趣可挖掘用户对类型Ⅱ中域名词的兴趣度。本文引入协同算法[5-6]辅助问题求解，给出类型Ⅱ中域名词的兴趣度预测。

对类型Ⅰ中域名词的喜好度进行归一化处

理，$\forall k^u_{1,j} \in K^u_1$，$U$ 对 $k^u_{1,j}$ 的兴趣度 $\varphi^u(k^u_{1,j}) \in FK^u_1 = \langle \varphi^u(k^u_{1,1}), \varphi^u(k^u_{1,2}), \cdots, \varphi^u(k^u_{1,s}) \rangle$，且：

$$\varphi^u(k^u_{1,j}) = \frac{g^u_{1,j}}{\sum_{g^u_{1,h} \in G^u_1} g^u_{1,h}} \tag{3}$$

同样，针对 V，设类型Ⅰ中域名词集为 $K^v_1 = \langle k^v_{1,1}, k^v_{1,2}, \cdots, k^v_{1,t} \rangle$，$\forall k^v_{1,h} \in K^v_1$，$V$ 对 $k^v_{1,h}$ 的兴趣为 $\varphi^v(k^v_{1,h}) \in FK^v_1 = \langle \varphi^v(k^v_{1,1}), \varphi^v(k^v_{1,2}), \cdots, \varphi^v(f^v_{1,t}) \rangle$。$\forall (k^v_{1,h} \in K^v_1) \wedge (k^v_{1,h} \notin K^u_1)$，添加 $k^v_{1,h}$ 至 K^u_1，FK^u_1 中相应的名词喜好度置为0，得到新的 K'^u_1、FK'^u_1；$\forall (k^u_{1,j} \in K^u_1) \wedge (k^u_{1,j} \notin K^v_1)$，添加 $k^u_{1,j}$ 至 K^v_1，FK^v_1 中相应喜好度值置为0，得到新的 K'^v_1，FK'^v_1，根据信息检索中余弦向量方法[5-6]得 U、V 相似度为：

$$\mathrm{sim}(U,V) = \frac{\sum_{(k^u_{1,p} \in K'^u_1) \wedge (k^v_{1,q} \in K'^v_1) \wedge (k^u_{1,p} = k^v_{1,q})} \varphi^u(k^u_{1,p}) \cdot \varphi^v(k^v_{1,q})}{\sqrt{\sum_{(k^u_{1,p} \in K'^u_1)} (\varphi^u(k^u_{1,p}))^2 \sum_{(k^v_{1,q} \in K'^v_1)} (\varphi^v(k^v_{1,q}))^2}} \tag{4}$$

针对用户 U，$\forall k^u_{2,r} \in K^u_2$（设 K^u_2 为类型Ⅱ中域名词集），借助与 U 最邻近的 n 个邻居，给出类型Ⅱ中域名词的评价预测[5]：

$$\varphi^u(k^u_{2,r}) = \frac{1}{\sum_{m=1}^{n} |\mathrm{sim}(e_m, U)|} \sum_{m=1}^{n} \mathrm{sim}(e_m, U) \cdot \varphi^{e_m}(k^u_{2,r})$$

式中 $\varphi^u(k^u_{2,r})$ 为用户 U 对 $k^u_{2,r} \in K^u_2$ 的喜好预测；$\varphi^{e_m}(k^u_{2,r})$ 为邻居 e_m 对 $k^u_{2,r} = k^{e_m}_{1,l}$ 的喜好度。

至此，基于检索输入日志，已完成针对类型Ⅰ、Ⅱ中域名词的兴趣挖掘。针对菜肴等不同“类别”分别执行上述过程，得到检索中各相关名词的喜好度值。归并不同“类别”，基于同一类别下所属名词的输入总次数与不同类别下所有名词的输入总次数间的比例关系等，获得不同“类别”各自的权重。输入总次数同样基于“时间戳”意义。将域名词的兴趣度关联各自所属“类别”的权重得到最终喜好度值，本文仍用 $\varphi^u(\cdot)$ 表示U的名词喜好度。

针对 U 的前 τ 次检索输入日志获得针对不同“类别”下的不同域名词的兴趣值，第 $(\tau+1)$ 次检索输入 $q_{\tau+1}$，对于分词得到的 $q_{\tau+1}$ 中任一名词 k_o。用 $w_{k_o, q_{\tau+1}}$ 表示传统信息检索中 $\langle k_o, q_{\tau+1} \rangle$ 的权重，则不同于传统的信息检索方法，相对 U，新的引入个性化偏好的 k_o 的权重 $w'^u_{k_o, q_{\tau+1}}$ 为：

$$w'^u_{k_o, q_{\tau+1}} = \begin{cases} (1+\varphi^u(k^u_{1,c}))^{0.5} \cdot w_{k_o, q_{\tau+1}} & \text{若} k_o = k^u_{1,c} \in K^u_1 \\ (1+\gamma \cdot \varphi^u(k^u_{2,d}))^{0.5} \cdot w_{k_o, q_{\tau+1}} & \text{若} k_o = k^u_{2,d} \in K^u_2 \end{cases} \tag{5}$$

式中 $(1+\varphi^u(k^u_{1,c}))^{0.5}$ 和 $(1+\gamma \cdot \varphi^u(k^u_{2,d}))^{0.5}$ 用于表征当前检索的个性化偏好加权；$\varphi^u(k^u_{1,c})$ 和 $\varphi^u(k^u_{2,d})$ 分别是类型Ⅰ、Ⅱ中域名词的用户偏好。类型Ⅰ中域名词的喜好度根据 U 的输入日志直接得到，而类型Ⅱ中域名词的喜好度借助“邻居”预测得到，引入因子 γ 降低类型Ⅱ中域名词喜好度的权重（γ 的经验取值为0.1～0.4）。相对于简单地将 U 对类型Ⅱ中域名词的喜好度直接置0的方法，通过上述过程可更为合理有效地获得针对当前检索的偏好加权 $(1+\gamma \cdot \varphi^u(k^u_{2,d}))^{0.5}$，直接置0的偏好加权为 $(1+0)^{0.5}$。

1.4 词典库更新策略

在PVSA系统中，除了基础性的通用词库，系统需要根据不同的领域搜索应用引入领域对应的行业词库。词库在搜索引擎中扮演重要角色，包括索引构建、有效检索需求获取等，皆离不开词库的支持，尤其是行业词库，在索引构建以及域名词差异化等环节起着非常重要的作用。所以，在原有词库基础上，针对瞬息万变的Web信息如何有效更新词库，也是个性化域搜索需要解决的问题。不同于产生式模型，条件随机场(CRF)[7]是通过模型训练最大化条件概率的无向图模型。该模型具有良好的特征知识融入、非马氏无后效性约束等特性，在句法分析等人工智能领域获得了广泛的应用[8-9]。在传统CRF中文分词、未登录词识别的基础上，基于领域元数据和行业词库等领域特征下的若干专有“指导信息”，使用条件随机场模型可以相对有效地识别包括行业名词在内的新词，在略微的人工辅助下，高效可靠地进行词典库的更新。

基于领域元数据，使用分步式策略完成未登录专有名词的识别。领域特征元数据对于词典库的更新同样具有重要的辅助作用；分步识别方式是指算法先就比较容易识别的元数据的属性值(实例值)进行新词识别，将这些新识别的名词应用到其他元数据实例值的新词识别中，已识别出的新词对于之后的新词识别具有良好的“指导”作用。为了清晰地阐释，以餐饮域为例进行说明。基于正则式匹配思想针对酒店名称域，地址域等进行酒店名称，地址等未登录专有名词识别，其中address实例值的基本正则式可以描述为“*市*区+路|大街|里|街+号*”。依上述表达式自左至右依次判别。为确保最高精度，新词识别的最终结果需引入人工干预。追加新识别得到的专有名词至相关词典，借助新行业词典，针对酒店描述等属性域，基于CRF模型进行菜谱名称等行业专有名词的识别。不同于通用搜索，以领域元数据、行业词典等域特征为识别出发点。从行业

特征出发，识别过程中，得到并引入特定领域的专有特征。例如，当大小为k的窗口中同时出现多个菜名时，窗口中其他词作为菜名的概率较高。菜名可以采用“招牌菜”、“推荐菜”等词语作为窗口特征词，对菜名的识别和抽取具有良好的指导作用。不难得到，购物域等其他领域的新词识别具有类似的识别过程。

以上是词典库的基本更新策略。面对瞬息万变的Web信息，行业词典的有效更新使系统能够更可靠地针对行业名词进行识别。能否准确识别行业名词对检索名词差异化等策略具有直接的影响，故词典库的有效更新也是PVSA有效工作的必要基础。

2　PVSA算法

设U第$\tau+1$次使用系统，针对领域d_i，U的任一检索输入为$q_{\tau+1}$。算法包括以下步骤：(1) 对$q_{\tau+1}$进行分词处理，基于式(3)～式(5)等进行检索名词差异化，计算每个检索名词的权重；(2) 借助http://jakarta.apache.org/lucene和前期构建的行业词典库等，基于领域元数据权重因子和差异化的检索词权重，计算$p_{i,k}$的初始得分$\delta_{i,k}$；(3) 由式(1)和式(2)引入领域主题喜好向量，计算基于用户个性化偏好的$p_{i,k}$的得分值$\xi_{i,k}$；(4) 根据网页得分值$\xi_{i,k}$对领域网页集P_i中的网页降序排序，得到满足用户个性化偏好的领域网页输出集。本文中，词库更新等部分作为算法的“后台支持”，没有直接显示在PVSA中。

3　试验评测

PVSA构建于Lucene平台之上，应用服务器采用Tomcat，使用两台PC主机服务器，分别部署Lucene和PVSA信息检索系统。Lucene自带分词器精度差，不同算法皆采用先期开发的基于行业词典的中文分词器。使用爬虫在“北京餐饮域”抓取6 196张网页并简单预处理后使用Lucene建立索引，实验部分给出PVSA与Lucene的对比评测。

选择100名研究生志愿者(其中的6位博士生是最终完成评测的人员)。在系统偏好采集阶段，每人根据自己的兴趣偏好，针对餐饮域给出不少于40个查询。实际应用中，用户使用系统会自动保留日志信息。

相同的两台PC，分别部署根据Apache开源包搭建的Lucene搜索平台和PVSA搜索平台。6位博士生基于各自的兴趣进行检索输入，针对系统的网页输出，根据自己的个性化偏好进行性能评测，给出系统的精度指标。每位博士生在两台部署有不同检索系统的PC上分别执行上述过程30次以上。实验目的是评价系统的领域个性化性能水平，所以实验过程中用户没有必要过多地给出太过复杂的查询输入。另外，在先前进行个性化评测时采用的是查准/查全率评价准则，但该准则实际上并不能很好地表征个性化偏好水平，因为查准/查全率并不能有效地描述较为关注的网页的相对“优劣”。若网页1和2都是满足检索要求的网页，但U_1比较喜欢低价位的川味饭馆，故他认为网页1比2要好，而U_2认为网页2要优于1，评测指标要求能够针对“相对优劣”等个性化因素进行明确的表征。采用www06的pairwise accuracy(对精度)[10]进行评价，对精度指标可以有效针对个性化精度进行表征。

设目标用户为U，针对任意一次查询，设$P_{i,n}\subset P_i$是检索结果输出中的前n张网页的集合(P_i是领域网页集)，定义集合X和Y，X基于用户判断进行构建，而Y基于机器判断进行构建，则：

$$X=\{\langle p_{i,j},p_{i,k}\rangle \mid p_{i,j}\in P_{i,n},p_{i,k}\in P_{i,n}\}$$

$$Y=\{\langle p'_{i,j},p'_{i,k}\rangle \mid p'_{i,j}\in P_{i,n},p'_{i,k}\in P_{i,n}\}$$

式中　$\langle p_{i,j},p_{i,k}\rangle$和$\langle p'_{i,j},p'_{i,k}\rangle$分别是由用户判断和由机器判断得到的网页“相对优劣”。根据文献[10]，对精度计算如下：

$$\text{pairwise accuracy}=\frac{|X\bigcap Y|}{|X|}$$

实际应用中，通常用户仅会针对检索结果中最前面的若干张网页进行浏览，本文选择n=20，即对于U的任意一次查询，针对检索输出中的前20张网页(不够20则选取实际网页数目)，U基于自己的个性化偏好给出这些网页与实际需求的“关联度”排序，据此得到对精度的X；根据机器的检索结果输出，得到对精度的Y；根据X和Y完成相应的对精度计算。最后，将参与评测的6位博士生最终得到的对精度进行平均值运算，得到系统的对精度指标。通过实验计算得到的PVSA算法、Lucene算法和完全随机情况下的对精度如表1所示。

表1　PVSA同Lucene等方法的对精度对比

检索算法	Random	Lucene	PVSA
对精度/(%)	50	62.795	70.118

随机情况下的对精度值是指在未施加任何操作的基础上，随机地确定两个网页的“好坏”关系，

此时的对精度指标为50%。Lucene具有相对有效的索引构建等机制，实验评测得到的Lucene算法的对精度值为62.795%。进一步引入个性化偏好，基于领域主题喜好向量等策略的PVSA的对精度指标达到70.118%，较完全随机情况下的对精度(50%)和Lucene的对精度(62.795%)的提升比例分别达到40.036%和11.661%，收效非常明显。Lucene开源包本身就pageRank计算的若干方面进行优化处理，Lucene算法的对精度由随机情况下的50%增加到62.795%，PVSA在Lucene之上搭建个性化的信息检索能力，引入领域主题喜好向量挖掘、基于用户偏好的检索名词差异化等策略、基于用户偏好给出个性化的领域检索服务输出，可进一步有效地提升系统的对精度指标。

PC基本配置为2.66 GHz、1.46 GB、200 rpm。将若干次检索耗时平均，算法平均耗时如表2所示。

表2 PVSA算法和Lucene一次检索的平均耗时

检索算法	Lucene	PVSA
一次检索平均耗时/ms	11.503 1	11.782 4

4 总 结

本文提出并研究、实现了基于用户偏好的垂直搜索算法PVSA。基于领域特征，PVSA给出行业词典的更新策略，借助基于领域主题偏好向量的网页差异化策略、领域元数据权重因子和检索名词的差异化策略等以有效挖掘、表征用户偏好，并以此为基础构建个性化的领域搜索算法，有效提升了检索系统的性能指标。今后希望在人机交互层面进行研究，更好地满足个性化需求。

参 考 文 献

[1] KRITIKOPOULOS A, SIDERI M. The Compass filter: search engine result personalization using web communities [J]. Lecture Notes in Computer Science, 2005, 3169: 225-240.

[2] 杨炳儒, 王 敏. 基于主题的个性化元搜索引擎的设计与实现[J]. 情报杂志, 2005, 24(7): 57-58.
YANG Bing-ru, WANG Min. The design and accomplishment of title-based personalized meta-search engine[J]. Journal of Information, 2005, 24(7): 57-58.

[3] 李效东, 顾敏清. 基于DOM的Web信息提取[J]. 计算机学报, 2002, 25(5): 526-533.
LI Xiao-dong GU Min-qing. DOM-based information extraction for the web sources[J]. Chinese Journal of Computers, 2002, 25(5): 526-533.

[4] 杜轩华, 袁 方. Perl在Web上的应用[J]. 微型机与应用, 2000, 19(3): 29-32.
DU Xuan-hua, YUAN fang. The application of perl on the web[J]. Microcomputer & Its Applications, 2000, 19(3): 29-32.

[5] DELGADO J, ISHII N. Memory-based weighted-majority prediction for recommender systems[C]//Proc of ACM SIGIR'99 Workshop Recommender Systems: Algorithms and Evaluation. Berkeley, USA: The Association for Computing Machinery, 1999.

[6] BREESE J S, HECKERMAN D, KADIE.C. Empirical analysis of predictive algorithms for collaborative filtering [C]//Proc of the Fourteenth Conference on Uncertainty in Artificial Intelligence. Stockholm: [s.n.], 1998: 43-52.

[7] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proc of the 18th International Conf on Machine Learning. Williamstown, MA: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[8] SHA F, PERDIRA F. Shallow parsing with conditional random field[C]//Proc of HLT-NAACL. Columbia University, Edmonton, Canada: [s.n.], 2003.

[9] PENG F C, FENG F F, MCCALLUM A. Chinese segmentation and new word detection using conditional random fields[C]//Proc of the 20th International Conference on Computational Linguistics (COLING 2004). Geneva, Switzerland: Morgan Kaufmann Publishers, Inc, 2004: 282-289.

[10] RICHARDSON M, PRAKASH A, BILL M. Beyond pagerank: machine learning for static ranking[C]//Proc of the 15th Int'l Conf World Wide Web. New York: Association for Computing Machinery, 2006: 705-715.

编 辑 蒋 晓