高技术通讯 2013 年 第23 卷 第9 期:933 ~938

doi:10.3772/j.issn.1002-0470.2013.09.009

基于浏览记录挖掘的个性化偏好建模①

张晓宇②

(中国科学技术信息研究所 北京 100038)

摘 要 为提高个性化信息检索性能,提出了一种基于浏览记录挖掘的偏好建模算法。该算法从浏览记录出发,深入挖掘用户在域和域值这两个维度上的偏好,从而自动构建并累积更新偏好模型,对检索结果进行个性化优化;给定查询,相关结果能够自动根据现有浏览记录进行偏好建模以实现个性化排序,无需任何额外的用户操作。讨论了关键参数的优化,以进一步提升算法性能,使其更加符合实际应用的需求,从而在精确刻画用户偏好的同时有效提升了用户体验。实验结果表明,基于浏览记录挖掘的个性化偏好建模算法能够显著提高检索性能,对于海量信息的有效获取具有重要意义。

关键词 个性化检索,偏好建模,浏览记录挖掘,用户体验,相关反馈

0 引 言

随着信息处理、大规模存储以及交互技术的发展,各种各样的信息大量涌现。面对信息的汪洋大海,迫切需要高效的检索技术以确保对信息方便快捷的获取,最大限度地满足用户的信息需求。而用户的信息需求往往是个性化的,不同用户对同一查询的主观期望千差万别,甚至同一用户在同一查询的多次检索中的具体侧重也会有所不同。传统的检索机制完全基于内容的相关性而忽略了用户的不同偏好,从而导致检索结果单一,无法更好地契合不同用户的实际需求。由此可见,高效的信息检索应该是个性化的[1,2],检索系统应能够准确地获取用户偏好,并基于用户偏好对检索结果进行相应调整。同时,为了确保良好的用户体验,额外的用户操作应尽可能地避免。为此,本文对个性化检索方面现有相关工作进行了梳理,在此基础上,提出了一种用于个性化检索的基于浏览记录挖掘的偏好建模算法,并对其中的若干重点问题进行了讨论,与其它方法进行比较,最后给出了通过实验验证该算法的有效性的结果。

1 相关工作

实现个性化检索的关键是用户偏好建模。根据用户参与的程度,现有用户建模算法可以大致分为三类:用户指定、相关反馈和用户行为分析。

1.1 用户指定

用基于用户指定的方法,用户需要明确指定查询的一系列限定条件,这是一种最为直接的偏好建模手段。事实上,许多广泛使用的搜索引擎已经采用这种方法对用户的查询进行精细化处理,并以“高级检索”的形式供用户选用。

基于用户指定的偏好建模固然简单,但不足之处也是显而易见。首先,用户指定信息对于建模必不可少,对于用户而言是一项额外的负担,影响了检索过程的流畅性;其次,限定条件事先确定,用户只能从有限的几个方面对查询进行细化描述,难以充分揭示偏好的多样性;此外,有些个性化偏好非常抽象甚至无法描述,因此提供恰当的指定信息本身就具有相当的难度。

1.2 相关反馈

相关反馈[3,4]是一种通过人机交互获取用户信息的有效手段,其主要思想是:如果用户对于当前检

① 中央级公益性科研院所基本科研业务费专项资金(XK2012-2、ZD2012-7-2)和中国科学技术信息研究所预研基金(YY201208)资助项目。

② 男,1983 年生,博士;研究方向:模式识别与智能系统;联系人,E-mail:zhangxy@istic.ac.cn

(收稿日期:2013-01-10)

索结果不满意,可以对一些数据进行相关或不相关的标引;这些被标引数据随后作为训练数据,用以更新检索模型,从而使得检索结果更加符合用户需求。

由于只需给出相关性的二值标引,因此相比用户指定而言,对用户的要求大大降低。但用户标引仍然是一个费时费力的过程。信息检索中,相关反馈本质上是一个监督分类问题(即将数据分为相关和不相关两类),为了获得满意的分类结果,需要足够多的训练数据,因而对用户标引量提出了较高的要求。为了减轻标引负担,研究人员提出了一些解决方案,包括:主动学习[5,6]、动态批量采样[7-9]、伪相关反馈[10,11]等。

1.3 用户行为分析

用户指定和相关反馈都或多或少地需要用户参与,而用户行为分析的方法则通过隐式的方法获取用户偏好,从而最大限度地减少用户操作。

用户行为分析旨在从用户行为出发深入挖掘其中蕴含的丰富个性化信息,由于用户行为记录易于获取,因此为偏好建模提供了大量可利用的数据。更重要的是,整个分析过程完全后台运行,无需用户干预,因此从用户体验角度出发无疑是更好的选择。

在以往的研究工作中,作者提出了基于相关反馈的偏好建模算法[12]。为了获取更优的检索结果,用户需要显式地对结果进行相关性标引作为偏好建模的训练数据。之后,作者又将算法改进为隐式相关反馈[13]。给定查询,初始检索结果按照相关性排序;当用户选择部分结果进行浏览之后,算法将被浏览结果作为相关结果自动更新偏好模型。上述方法都需要用户一定程度的干预。本文在已有工作的基础上,通过深入挖掘用户浏览记录获取个性化信息,在更加精确刻画用户偏好的同时进一步提升用户体验。

2 基于浏览记录挖掘的偏好建模

给定查询,所有相关的结果可以通过匹配算法获得。难点在于,如何根据用户的个性化偏好对结果集进行排序,从而将用户最感兴趣的内容安排在列表中靠前的位置。

对于特定的数据库,数据可以从不同角度进行描述,这些角度称为“域(field)”。例如,一篇文章可以用不同域进行描述,包括:标题、正文、作者、机构等。不同域对应不同的“域值(field value)”。对于同一查询,不同用户的个性化偏好体现在对于域和域值的不同关注度上。浏览记录可以作为获取用户偏好的重要依据,由于用户关注的域和域值在浏览记录中频繁出现,因此可以从浏览记录出发对域和域值的用户关注度进行量化,并基于此进行偏好建模、形成个性化排序准则,从而实现个性化检索。

2.1 算法

本文用 $\boldsymbol{f}=(f_1,f_2,\cdots,f_I)$ 来表示元数据中的所有域,其中 I 是域的数量。每个域 $\boldsymbol{f}_i$ 用一个 l_i 维向量表示,其中 l_i 是 $\boldsymbol{f}_i$ 中不同域值的数量。相应地,用户浏览记录可以用相同的数据结构表示为 $\boldsymbol{n}=(n_1,n_2,\cdots,n_I)$,其中 $\boldsymbol{n}_i$ 是一个 l_i 维向量,其元素 $\boldsymbol{n}_i(j)$ 表示 $\boldsymbol{f}_i(j)$ 被浏览的频率。图 1 描述了 $\boldsymbol{f}$ 和 $\boldsymbol{n}$ 所共同采用的“域—域值”结构。

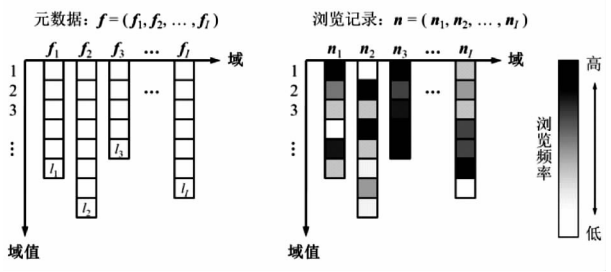

图 1 元数据“域—域值”结构

为了准确刻画用户的个性化偏好,需要重点关注两个关键因素:用户对不同域以及特定域不同域值的关注度(也即权重)。

首先,算法从浏览记录的一致性入手,量化用户对不同域的关注度。对于域 $\boldsymbol{f}_i$,其多样性定义为被浏览过的不同域值的数量:

$$d_i = \sum_{j=1}^{l_i} \varphi(\boldsymbol{n}_i(j),0) \tag{1}$$

其中

$$\varphi(x,y) = \begin{cases} 1, x > y \\ 0, x \leqslant y \end{cases} \tag{2}$$

域的多样性越小,其一致性越大,从而用户对其关注度越高。因此,域的一致性通过计算多样性的倒数获得:

$$c_i = \frac{1}{d_i} \tag{3}$$

通过标准化,域 $\boldsymbol{f}_i$ 的权重定义为

$$w_i = \frac{c_i}{C} \tag{4}$$

其中

$$C = \sum_{i}^{I} c_i \tag{5}$$

利用式(4)可以进而得到域权重向量 $\boldsymbol{w}=(w_1,w_2,\cdots,w_I)$。

其次,算法基于浏览频率,量化用户对特定域不同域值的关注度。对于域 $\boldsymbol{f}_i$,域值 $\boldsymbol{f}_i(j)$ 的权重定义为

$$v_i(j)=\frac{\boldsymbol{n}_i(j)}{N} \tag{6}$$

其中

$$N=\sum_{j}^{l_i}\boldsymbol{n}_i(j) \tag{7}$$

是用户浏览数据的总量,用于标准化。每个权重 $\boldsymbol{v}_i=(\boldsymbol{v}_i(1),\boldsymbol{v}_i(2),\cdots,\boldsymbol{v}_i(l_i))$ 是一个 l_i 维向量,其元素对应着域 $\boldsymbol{f}_i$ 的 l_i 个域值。综合各个 $\boldsymbol{v}_i$ 可以得到域值权重向量 $\boldsymbol{v}=(v_1,v_2,\cdots,v_I)$。

最后,对于每一个与用户查询相关的结果 $\boldsymbol{r}=(r_1,r_2,\cdots,r_I)$,其中 r_i 对应于域 $\boldsymbol{f}_i$ 的一个特定取值,$\boldsymbol{r}$ 的偏好值可以依据下式计算:

$$Score(r)=\sum_{i=1}^{I}w_i\sum_{j=1}^{l_i}v_i(j)\delta(r_i,f_i(j)) \tag{8}$$

其中

$$\delta(x,y)=\begin{cases}1, & x=y\\ 0, & x\neq y\end{cases} \tag{9}$$

如公式所示,用户偏好模型中的两个关键因素 $\boldsymbol{w}$ 和 $\boldsymbol{v}$ 均来源于浏览记录 $\boldsymbol{n}$,因此 $\boldsymbol{n}$ 的计算对于用户偏好模型构建至关重要。浏览记录是随着用户浏览行为的发生而不断改变的,每当一个新的数据被用户浏览,$\boldsymbol{n}$ 便会进行相应更新。因此,有必要给出 $\boldsymbol{n}$ 的增量表达式用于实时运算。假设用户浏览 N 个数据之后的浏览记录为 $\boldsymbol{n}^{(N)}$,则当第 $N+1$ 数据被浏览之后,新的浏览记录根据下式进行更新:

$$\boldsymbol{n}_i^{(N+1)}(j)=\boldsymbol{n}_i^{(N)}(j)+\delta(a_i^{(N+1)},\boldsymbol{f}_i(j)) \tag{10}$$

随着越来越多的样本被浏览,用户的个性化偏好将被刻画得越来越准确。

2.2 讨论

为使算法更加符合实际应用的需求,有必要对其中若干环节进行深入探讨。

2.2.1 域多样性自适应计算

在实际应用中,并非所有被浏览的内容都是用户真正感兴趣的,偶然的误点击普遍存在。这些无意中被浏览的数据无法准确反映用户的真实偏好,有时甚至与用户偏好截然相反。浏览记录中噪声的存在会极大地影响偏好建模的可靠性,因此需要有针对性地加以处理。

在浏览记录中,无意识的浏览会偶然地命中某些域值,这集中体现在 $\boldsymbol{n}$ 的稀疏性上。零星分布于 $\boldsymbol{n}$ 中的噪声对域值权重影响甚微,根据式(6),较低的浏览频率自然导致较低的权重。但这些噪声会严重影响域权重的精度,因为如式(1)所示,即使只有一次浏览也会对多样性产生影响,进而显著改变域权重。为使偏好模型对噪声具有更强的容错性,算法对式(1)进行改进:

$$d_i=\sum_{j=1}^{l_i}\varphi(n_i(j),\sigma) \tag{11}$$

其中 σ 是用于过滤噪声的门限值,从而确保低频浏览数据不参与域多样性的计算。

门限 σ 的引入固然可以弱化浏览记录中噪声的影响,但是作为一个经验性参数,对其恰当赋值却具有相当的难度;此外,固定门限值无法适用于不断增长的浏览记录,随着越来越多的数据被浏览,不可避免地会混入更多的噪声,从而使得既定门限失效。

鉴于固定门限的不足,本文进一步提出自适应的域多样性计算方法,其主要思想是:浏览记录中的噪声仅占一小部分,因此只需重点关注其主要部分。对特定的浏览记录 $\boldsymbol{n}_i$,算法首先对其降序排序得到 $\boldsymbol{n}_i^s$;然后不断累加 $\boldsymbol{n}_i^s$ 中排名前 t 的记录,直到总和超过 N 中预定的比例 τ;最终,t 即设定为域多样性。详细算法流程如图 2 所示。

输入	$\boldsymbol{n}_i(N=\sum_j^{l_i}\boldsymbol{n}_i(j))$, τ .
初始化	$\boldsymbol{n}_i^s=\boldsymbol{n}_i$ 降序排序, $N_i=\boldsymbol{n}_i^s(1)$, $t=1$.
循环	当 $N_i/N<\tau$ 时:
	$t=t+1$,
	$N_i=N_i+\boldsymbol{n}_i^s(t)$.
输出	$d_i=t$.

图 2 自适应域多样性赋值

采用自适应方法确保域多样性随着浏览记录的改变而动态调整,因而比固定门限方法更加鲁棒。

2.2.2 偏好时间窗设定

用户偏好大致可以分为两种:长期偏好和短期偏好。前者反映了用户在较长时间内的浏览偏好,它体现在浏览记录 $\boldsymbol{n}^{(N)}$ 中,随着浏览数据量 N 不断增大,长期偏好将越来越清晰;后者则反映了用户短期关注的行为,它往往在某一较短时间段的浏览记录中集中出现,与长期偏好相反,随着浏览数据量 N 不断增大,短期偏好会被其它浏览记录所淹没。

为了准确刻画用户的短期偏好,算法采用滑动

时间窗对浏览记录进行限定,以此定义短期浏览记录为

$$\boldsymbol{n}^{*(T)} = \boldsymbol{n}^{(N)} - \boldsymbol{n}^{(N-T)} \tag{12}$$

其中 T 是短期浏览记录中的浏览数据量。用 $\boldsymbol{n}^{*(T)}$ 代替 $\boldsymbol{n}$,便可以根据式(4)和(6)计算相应的域权重 $\boldsymbol{w}^{*(T)}$ 和域值权重 $\boldsymbol{v}^{*(T)}$,进而获得短期偏好模型下的偏好值:

$$Score^{*(T)}(\boldsymbol{r}) = \sum_{i=1}^{I} \boldsymbol{w}_i^{*(T)} \sum_{j=1}^{l_i} \boldsymbol{v}_i^{*(T)}(j)\delta(r_i, \boldsymbol{f}_i(j)) \tag{13}$$

值得注意的是,长期偏好与短期偏好并无显著差异,随着时间窗 T 的增大,短期偏好逐渐变为长期偏好。特别地,当 $T = N$ 时,由式(12)可得 $\boldsymbol{n}^{*(T)} = \boldsymbol{n}^{(T)}$。因此,式(8)可以看成是式(13)的一种特殊形式。

2.3 比较

在以前的研究中,作者也曾提出过偏好建模算法,将其与本文算法进行比较将有利于更好地理解算法。

在文献[12]中,偏好建模是通过显式相关反馈实现的。这种方法的不足之处在于:第一,相关反馈是一个"用户驱动(user-driven)"的过程,只有用户给出了明确的相关性标引之后才能建立偏好模型,离开用户标引个性化检索便无从实现;第二,相关反馈是"一次性优化(one-time optimization)",用户的标引只能用于当前查询结果的个性化,无法用于新的查询。

在文献[13]中,算法改进为隐式相关反馈,用户浏览的数据自动作为相关结果进行模型训练,因此用户无需显示地进行标引。但是这种方法仍存在诸多问题:第一,隐式相关反馈仍然是用户驱动的,用户提交查询后需要先浏览部分结果算法才能进行偏好学习;第二,算法仍然是一次性优化,用户浏览信息仅对当前查询有效;第三,隐式相关反馈算法复杂度过高。文献[13]中的偏好值可以表示为

$$Score^{RF}(r) = \max_{b_k \in B^{(K)}(1 \leqslant k \leqslant K)} \sum_{i=1}^{I} w_i y_{ki} \delta(r_i, b_{ki}) \tag{14}$$

其中 $\boldsymbol{B}^{(K)}$ 表示用户浏览过的数据集合,K 是其中的数据量(显然 $K = N$),y_{ki} 是域值权重矩阵 $\boldsymbol{Y}_{K \times I}$ 的元素。由式(14)可知,偏好值的计算需要比较所有结果的所有域,因此其计算复杂度为 $O(I \times K)$。域的数量 I 是有限的且相对固定的,而浏览数据量 K 则是不断增长的,随着越来越多的数据被浏览,偏好值的计算复杂度将越来越高,直至难以接受。

相比之下,本文算法则在各方面均优于以上方法。第一,算法是"自驱动(self-driven)"的,给定查询,算法将自动依据现有浏览记录进行个性化结果组织,而无需用户对当前结果进行操作;第二,偏好模型能够"累积优化(accumulative optimization)",在以往查询过程中学习到的个性化信息可以用来处理新的查询,从而保证模型不断优化;第三,算法计算复杂度有限。由式(8)或(13)可知,偏好值的计算基于对所有域及所有域值的比较,因此其计算复杂度为 $O(I \times J)$,其中

$$J = \max_{1 \leqslant i \leqslant I} l_i \tag{15}$$

是域值数量的最大值。由于域数量 I 和域值数量 J 都是有限的而且与浏览记录数无关,因此随着浏览记录的增加,计算复杂度不会改变。

综上,基于浏览记录挖掘的偏好建模算法在性能和效率方面均优于以往的方法。

3 实 验

本文将基于浏览记录挖掘的偏好建模算法用于体育视频个性化检索中,以验证其有效性。

实验数据包括 30h 的足球视频和 30h 的篮球视频。采用[12,13]中的方法,借助网络直播文字,原始视频被自动切割为视频片段并自动标引为 4 个域:球员、球队、对手球队、事件。20 名用户被邀请参加算法评估,每名用户根据各自偏好提交 10 个查询(允许有重复),对每个查询,用户选取 10 个最感兴趣的内容进行浏览,从而形成一系列浏览记录:$\boldsymbol{n}^{(N)}$ ($N = 10, 20, 30, \cdots, 100$)。基于浏览记录,偏好模型自动构建并不断完善。具体而言,对第 q 个查询($q = 1, 2, 3, \cdots, 10$),结果排序所依据的偏好值来源于浏览记录 $\boldsymbol{n}^{((q-1)\times 10)}$。特别地,对第 1 个查询,浏览记录为 $\boldsymbol{n}^{(0)}$,这意味着初始状态的首次检索不引入偏好信息。实验采用排名前 m 结果的正确率作为平均指标,也即 P@m[14,15]:

$$P@m = \frac{R_m}{m} \tag{16}$$

其中 R_m 是排名前 m 的结果中相关的数量。实验分别计算在 $m = 10$ 和 20 时所有用户的 P@m 平均值。

如表 1 所示,实验比较了基于浏览记录挖掘的偏好建模算法(包含 4 种不同设定)和相关反馈算法(反馈前后 2 种结果)。

表 1　算法设定

P0	基于浏览记录挖掘的偏好建模:利用式(1)计算域多样性,时间窗 $T=N$
P1	基于浏览记录挖掘的偏好建模:利用式(11)计算域多样性($\sigma=2$),时间窗 $T=N$
P2	基于浏览记录挖掘的偏好建模:利用图 2 自适应计算域多样性($\tau=90\%$),时间窗 $T=N$
P3	基于浏览记录挖掘的偏好建模:利用图 2 自适应计算域多样性($\tau=90\%$),时间窗 T 由用户指定
B0	相关反馈:反馈前,初始返回结果
B1	相关反馈:反馈后,利用浏览数据作为隐式反馈

实验结果如图 3 所示。

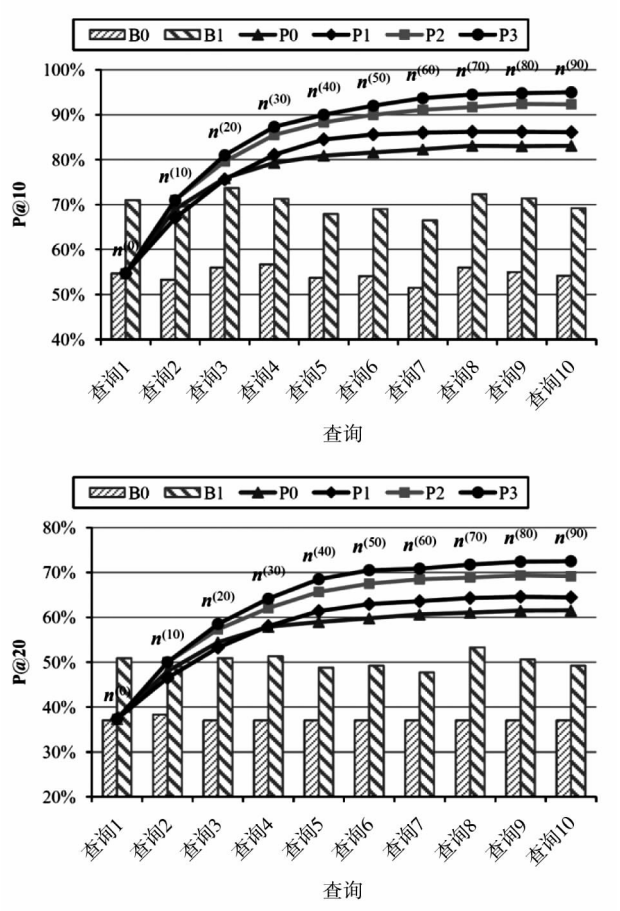

图 3　个性化检索平均 P@m 值($m=10$、20)

实验结果分析如下:

● 基于浏览记录挖掘的偏好建模算法是一种有效的个性化检索方法(P0、P1、P2、P3)。给定查询,算法依据偏好模型自动实现个性化结果组织,无需用户干预。随着更多的数据被浏览,偏好模型累积更新,从而使得结果更加契合用户的个性化需求。

● 相关反馈算法的初始结果由于没有反馈信息,因而无法实现个性化检索(B0)。只有用户浏览了部分检索结果之后,才能通过隐式反馈实现个性化组织(B1)。但是作为一次性优化算法,相关反馈中获得的个性化信息只能用于本次查询结果优化,而无法累积用于后续查询,因此限制了算法性能提升。

● 基于浏览记录挖掘的偏好建模算法中不同设定也效果各异。不设门限计算域多样性(P0)受噪声影响较大,效果最差;采用固定门限(P1)能够一定程度上降低噪声影响,提高检索精度;采用自适应方法计算域多样性(P2)能进一步提升检索性能,这表明自适应的方法更加适用于不断增长的浏览记录;滑动时间窗是处理长期偏好和短期偏好的有效手段,如果用户可以自主设定时间窗(P3),则检索结果最优。

值得一提的是,实验中仅仅使用体育视频作为一个应用来评估算法的有效性,基于浏览记录挖掘的偏好建模算法本身是一种通用的方法,适用于各种信息的个性化检索。

4　结　论

本文提出了一种基于浏览记录挖掘的偏好建模算法用于个性化信息检索。通过深入分析浏览记录,算法从域和域值两个维度上挖掘用户个性化偏好,从而实现个性化检索。与相关研究相比,本文算法的贡献在于:第一,作为一种自驱动算法,个性化检索能够自动地、隐式地完成,无需额外的用户干预;第二,偏好模型充分挖掘浏览记录,能够随着浏览记录的增加累积更新,确保检索性能不断提升;第三,算法计算复杂度较低且相对稳定,并不随浏览记录的增加而改变。实验表明,基于浏览记录挖掘的偏好建模算法能够准确刻画用户个性化需求,对于提高个性化检索性能具有重要意义。

参考文献

[1] Foltz P W, Dumais S T. Personalized information delivery: an analysis of information filtering methods. *Communications of the ACM*, 1992, 35(12): 51-60

[2] Liu F, Yu, C, Meng W. Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 2004, 16(1): 28-40

[3] Rui Y, Huang T S, Ortega M, et al. Relevance feedback: a power tool for interactive content-based image retrieval.

IEEE Transactions on Circuits and Systems for Video Technology, 1998, 8(5): 644-655

[4] Salton G, Buckley C. Improving retrieval performance by relevance feedback. *Readings in Information Retrieval*, 1997: 355-364

[5] McCallum A, Nigam K. Employing EM in pool-based active learning for text classification. In: Proceedings of the 15th International Conference on Machine Learning, Madison, USA, 1998. 350-358

[6] SchohnG, CohnD. Less is more: active learning with support vector machines. In: Proceedings of the 7th International Conference on Machine Learning, Stanford, USA, 2000. 839-846

[7] Zhang X Y, Cheng J, Lu H Q, et al. Weighted co-SVM for image retrieval with MVB strategy. In: Proceedings of IEEE International Conference on Image Processing, San Antonio, USA, 2007, 517-520

[8] Zhang X Y, Cheng J, Lu H Q, et al. Selective sampling based on dynamic certainty propagation for image retrieval. In: Proceedings of the 14th International Conference on Advances in Multimedia Modeling, Kyoto, Japan, 2008. 425-435

[9] Zhang X Y. Dynamic batch selective sampling based on version space analysis. *High Technology Letters*, 2012, 18(2): 208-213

[10] Sakai T, Manabe T, Koyama M. Flexible pseudo-relevance feedback via selective sampling. *ACM Transactions on Asian Language Information Processing*, 2005, 4(2): 111-135

[11] Yan R, Hauptmann A, Jin R. Multimedia search with pseudo-relevance feedback. In: Proceedings of International Conference on Image and Video Retrieval, Urbana-Champaign, USA, 2003. 238-247

[12] Zhang Y F, Zhang X Y, Xu C S, et al. Personalized retrieval of sports video. In: Proceedings of the International Workshop on Multimedia Information Retrieval, Augsburg, Germany, 2007. 313-322

[13] Zhang Y F, Xu C S, Zhang X Y, et al. Personalized retrieval of sports video based on multi-modal analysis and user preference acquisition. *Multimedia Tools and Applications*, 2009, 44(2): 305-330

[14] Wang X J, Zhang L, Jing F, et al. AnnoSearch: image auto-annotation by search. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, NY, USA, 2006. 1483-1490

[15] Liu J, Wang B, Li M, et al. Dual cross-media relevance model for image annotation. In: Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, 2007. 605-614

Personalized preference modeling based on browsing history mining

Zhang Xiaoyu

(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract

A novel preference modeling algorithm based on browsing history mining is proposed to improve the personalized retrieval performance of informafion retrieval systems. Based on the browsing log, the algorithm deeply explores users' interest in both the dimensions of field and field value, and automatically constructs and accumulatively updates the preference model to optimize the personalized retrieval. Given a query, the relevant retrieval results can spontaneously be ranked according to their corresponding preference score without any extra user interference. Advanced settings are subsequently discussed to further improve the algorithm for practical use. The experimental results demonstrate the advantages of the proposed algorithm over the previous work.

Keywords: personalized retrieval, preference modeling, browsing history mining, user experience, relevance feedback