ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, Supplement, December 2009, pp.59−65

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于时间访问轨迹的文件的智能推荐*

韩 爽+, 王 衡

(北京大学 计算机科学技术系,北京 100871)

Intelligent File Recommendation Based on Time Access Tracking

HAN Shuang+, WANG Heng

(Department of Computer Science and Technology, Peking University, Beijing 100871, China)

+ Corresponding author: E-mail: hanshuang@graphics.pku.edu.cn

Han S, Wang H. Intelligent file recommendation based on time access tracking. *Journal of Software*, 2009, 20(Suppl.):59−65. http://www.jos.org.cn/1000-9825/09008.htm

Abstract: Since the amount of information on today's computers grows, helping computer users to locate the required files in file systems has become an important topic in today's intelligent interaction model research. Past research has mostly concentrated in PIM (personal information management), to re-organize file hierarchies in a more understandable way for individual users. However, due to the numerous extra operations and the long period that needed for the re-organizing of users' knowledge systems, the preceding applications can hardly be adopted by users. Considering that there would be a certain topic or purpose when user accessing files (a user may always view several files that related to the same topic during the same time), this paper proposes file recommendation based on tracking user's file operations. An intelligent file recommendation desktop toolkit (IFRDT) is implemented, which will track user's file access history, recommend the most related files according to the currently being accessed file, to reduce time cost for finding desired information. Experimental results show that IFRDT can save more energy of searching files than history, and users can find over 50% of desired files in IFRDT and directly open them without searching in directories.

Key words: prediction; recommendation; time access tracking; machine learning; intelligent user interface; data mining

摘 要: 随着计算机用户个人信息量的日益扩大,如何帮助用户在系统中快速找到所需资源已成为当前智能交互行为模型的重要课题.过往的研究大多集中于个人信息管理,力求以更加便于用户理解的个性化方式重新组织计算机资源结构.然而,由于上述系统往往需要用户大量的额外操作,并且重构用户的知识系统需要较为漫长的时间而不被用户采用.考虑到用户访问文件的主题性和目的性(用户往往会出于同一目的在同一时间段内同时访问多个同主题相关的文件),提出基于用户时间访问轨迹的智能文件推荐,并设计实现基于时间访问轨迹的智能文件推荐桌面工具(intelligent file recommendation desktop toolkit,简称 IFRDT),将根据用户访问文件的轨迹,针对用户当前正在访问的文件向用户推荐最有可能被访问到的同主题的其他文件,以减少用户查找所需资

* Supported by the National Natural Science Foundation of China under Grant No.U0735004 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z159 (国家高技术研究发展计划(863))

Received 2008-09-20; Accepted 2009-04-09

源花费的时间开销.实验结果表明,使用 IFRDT 向用户推荐文件比仅仅向用户呈现访问历史更能为用户节省查找文件的时间;被试用户可以在 IFRDT 中找到一半以上的所需文件,这就是为用户节约了一半以上的查找开销.
关键词: 预测;推荐;时间访问轨迹;机器学习;智能用户界面;数据挖掘

当今用户的个人信息量越来越大,操作系统传统的目录存储结构已无法满足用户快速寻找到所需资源的需求:文件数量日趋扩大,目录层次日趋复杂[1],用户查找文件需要耗费更多精力.一些相关调查表明,用户在查找文件这一日常任务上耗费了大量时间和精力[2−4].

设计用于协助用户查找文件的智能工具可以缩短用户在个人信息系统中寻找信息的时间.过往的大量工作集中在个人信息管理(personal information management,简称PIM)上.许多系统采用搜索引擎的形式,支持用户根据文件属性进行查找[5,6];另一些系统则以文件的某几项重要属性(例如为文件添加标签)的层次结构来代替传统目录式存储结构[7].

在文献[8]中,Bao 等人提出按任务划分文件集:用户自定义若干任务,每项任务对应若干相关文件,以此为基础记录目录访问频度,以便向用户推荐最有可能需要访问的目录,将用户定位到目标文件夹的点击次数减到最少.然而,该系统需要正确判断出用户的当前任务,否则推荐很可能失效.智能文件推荐桌面工具(intelligent file recommendation desktop toolkit,简称 IFRDT)同样以主题(任务)为指导,但我们认为,任务与任务之间有时并没有明显的界线,并且未必一定要明确用户的当前任务才能明确需要推荐的文件,任务是隐式存在于上下文中的.IFRDT 利用用户访问文件系统的交互历史上下文,分析用户访问文件的同主题特性,在用户查找文件之前向用户推荐与最近访问文件最相关的文件,帮助用户快速获得所需信息.用户在使用 IFRDT 时无须进行打开/关闭文件之外的额外操作.

1 基于时间轨迹的智能推荐

为减少用户查找所需资源的时间开销,我们利用用户的文件访问时间轨迹,计算文件与文件访问时间重叠程度,测量文件之间的关联程度,以便在用户浏览文件的过程中即时地向用户推荐可能需要访问的文件.

1.1 同主题相关性

由于用户访问文件的密集性(用户会在一段时间内集中地访问多个文件)和有主题性(用户在特定时间段内访问的文件可能是与同一主题有关的),可以通过记录文件访问的日志,获取文件与文件之间的相关信息,即在某一时间段内曾被同时访问过的多个文件有可能在未来的某一时间段内再次被同时访问到,那么对于用户打开过的任意两个文件,都存在着某种程度的关联,我们称其为同主题相关性.由于每一个文件都与其他多个文件同主题相关,那么就可以得到一个与该文件同主题相关的其他文件列表.存在于这一列表中的与该文件同主题高度的文件很可能在将来再次与该文件同时被用户打开.

1.2 前提与假设

本文的论述是基于以下几个前提与假设进行的:

(1) 研究范畴只限于文档(office 系列文档,txt,pdf 等);

(2) 不考虑文件被转移或文件名更改等情况;

(3) 同一文件被多个应用程序同时打开,将其被访问的所有重叠时间序列合并成为一个时间序列看待;

(4) 用户一旦结束对某一文件的使用,便会及时将其关闭.

1.3 文件间的关联距离

用户在访问文件时往往带有明确的目的性,用户出于同一目的常常在同一时间访问多个同主题相关的文件,而这些文件经常并不集中存放于文件目录的相邻位置,假如用户不将它们手动整理归类,那么文件分布的混乱将给用户查找有用信息带来极大的不便.而用户访问文件的密集性和有主题性又决定了文件与文件之间是

存在关联性的,当用户集中处理同一主题的事务时,会打开一批与该主题相关的文件.我们将文件与文件的同主题相关性看作文件与文件的距离.这里的距离衡量的是文件之间的相关程度,一个文件与另一个文件的距离越近,它们的同主题相关性就越高.我们将这个距离称作关联距离.

如图 1 所示.图中的圆圈代表文件,各文件名分别为$f_0, f_1, f_2, f_3, f_4, f_5, f_6, f_7$,文件之间的连线代表文件间的关联距离,关联距离的数值标于连线上(其中关联距离为 0 的连线将被忽略).可见,f_0与f_5的关联距离是最近的,表明f_0与f_5相对于其他几个文件而言相关性最高.

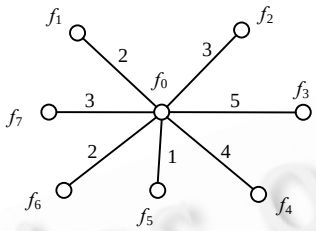


Fig.1 Distances between files

图 1 文件间的关联距离

1.3.1 关联距离表示

为了便于下文的讨论,我们将文件f_1与文件f_2的关联距离记为$D(f_1, f_2)$,其中f_1,f_2表示两个不同文件的绝对路径.

假设文件f_{dest}与n个文件$f_1, f_2,..., f_n$关联.在考查f_{dest}与其他文件的关联距离时,为避免混乱,以f_{dest}为目标文件与其他文件的关联距离公式以$D(f_{dest},f_i)(i=1,2,...,n)$为准.

1.3.2 文件访问时间重叠度

通常,文件之间的关联距离与文件的访问时间重叠程度有关.文件访问时间区间为打开文件的时间到关闭该文件的时间之间的时间区间.如果目标文件的历史访问时间与文件的历史访问时间的重叠度超过一定的相对数值,就可以认为这个文件与目标文件相关联.

对于任意文件f_x,它的访问时间区间序列可表示为$[ta_{x1},tc_{x1}],[ta_{x2},tc_{x2}],...,[ta_{xk},tc_{xk}]$,其中$k$表示文件$f_x$被访问过的总次数;$[ta_{xi},tc_{xi}](i=1,2,...,k)$表示第$i$次访问文件$f_x$时的访问时间,$ta_{xi}$为打开时间,$tc_{xi}$为关闭时间.易知:$\cap_{i=1}^{k}[ta_{xi},tc_{xi}]=\varnothing$.将$f_x$总的访问时间$[ta_{x1},tc_{x1}]\cup[ta_{x2},tc_{x2}]\cup...\cup[ta_{xk},tc_{xk}]$记为$U_x$.

假设在某一时刻t,文件f_x已被访问过m次,它的访问时间区间序列为$[ta_{x1},tc_{x1}],[ta_{x2},tc_{x2}],...,[ta_{xm},tc_{xm}]$;文件$f_y$已被访问过$n$次,它的访问时间区间序列为$[ta_{y1},tc_{y1}],[ta_{y2},tc_{y2}],...,[ta_{yn},tc_{yn}]$,$f_x$与$f_y$的总访问时间分别为

$$U_x=[ta_{x1},tc_{x1}]\cup[ta_{x2},tc_{x2}]\cup...\cup[ta_{xm},tc_{xm}],$$

$$U_y=[ta_{y1},tc_{y1}]\cup[ta_{y2},tc_{y2}]\cup...\cup[ta_{yn},tc_{yn}].$$

那么,$U_x\cap U_y$表示文件f_x与文件f_y访问重叠时间,记作U_{in_xy},亦即U_{in_yx};$U_x\cup U_y$表示文件f_x与文件f_y的访问覆盖时间(即f_x与f_y总访问时间的并集),记作U_{un_xy},亦即U_{un_yx}.

这样,文件f_x与文件fy之间的关联距离$D(f_x, f_y)$可通过下式来计算:

$$D(f_x, f_y)=\frac{|U_{in\_xy}|}{|U_{un\_xy}|} \tag{1}$$

1.3.3 关联文件的时效性

用户所关注的主题不是一成不变的,越接近当前时间的访问时间重叠度,对文件关联距离的影响越大.因此,在计算文件间关联距离时应将时效性也考虑在内.用户通常较倾向于访问最近访问过的文件,为此我们制定一个衰减因子α,满足 $0<\alpha<1$(α的经验值为 0.7 左右).每结束一个文件的一次访问,就意味着首先需要衰减所有已经存在的关联距离.假设文件系统中共有N个文件,衰减每一对f_x与$f_y(x,y=1,2,...,N,x\neq y)$的关联距离,就是令衰

减后的关联距离$D(f_x,f_y)^{attenuated}=\alpha\times D(f_x,f_y)=\frac{\alpha\times|U_{in_xy}|}{|U_{un_xy}|}$,这实际上是对$|U_{in_xy}|$的衰减.将$|U_{in_xy}|$记作$\theta_{xy}$,则有:

$$D(f_x,f_y)=\frac{\theta_{xy}}{|U_{un\_xy}|},\frac{\theta_{xy}^{attenuated}}{|U_{un\_xy}|}=D(f_x,f_y)^{attenuated}=\alpha\times D(f_x,f_y)=\frac{\alpha\times\theta_{xy}}{|U_{un\_xy}|},\text{故}$$

$$\theta_{xy}^{attenuated}=\alpha\times\theta_{xy} \tag{2}$$

当文件f_x结束它的一次访问(假设这次访问时间区间为$[ta_x,tc_x]$)时,已知文件f_x已经与n个文件$f_1, f_2,..., f_n$关联;另有m个正被用户访问但未曾与f_x关联过的文件$f_{n+1},f_{n+2},...,f_{n+m}$.对于每一个文件$f_i(i=1,2,...,n+m)$都可以得到$f_i$与$f_x$的访问重叠时间$U_{in_ix}=U_{in_xi}$,访问覆盖时间$U_{un_ix}=U_{un_xi}$.假如$f_i$正在被用户访问尚未关闭,它的最后一次打开时间为$ta_i$,不妨将访问覆盖时间$U_{un_ix}(U_{un_xi})$修改为$U_{un_ix}=U_{un_xi}=U_{un_ix}\cup[ta_i,tc_x]$.

由于此前已经首先对文件系统中包括f_i与f_x的关联距离在内的所有已经存在的关联距离进行了衰减,那么f_i与f_x的关联距离已经变为$D(f_i,f_x)^{attenuated}=D(f_x,f_i)^{attenuated}=\frac{\theta_{ix}^{attenuated}}{|U_{un_ix}|}$;而$f_i$与$f_x$的新的访问重叠时间和新的访问覆盖时间应分别变为

$$U'_{in\_ix}=U'_{in\_xi}=(U_{un\_ix}\cap[ta_x,tc_x])\cup U_{in\_ix} \tag{3}$$

$$U'_{un\_ix}=U_{un\_xi}=U_{un\_ix}\cup[ta_x,tc_x] \tag{4}$$

将$|U_{un_ix}\cap[ta_x,tc_x]|$记作$\Delta t_{ix},\Delta t_{ix}$即为文件$f_x$的最近一次访问为$f_i$与$f_x$带来的额外访问重叠时间,那么$f_i$与$f_x$新的关联距离为

$$D(f_i,f_x)'=D(f_x,f_i)'=\frac{\theta_{ix}^{attenuated}+\Delta t_{ix}}{|U'_{un\_ix}|} \tag{5}$$

1.3.4 用户评价

用户可以评价IFRDT推荐的文件,执行从推荐文件列表中删除不相关文件的操作.用户的删除操作也将成为影响文件间关联距离的因素之一.对文件f_x和f_y,假如用户将f_y从f_x的相关文件列表中删除,θ_{xy},θ_{yx}将被自动置零,相当于将f_x和f_y的相关性清零.

1.4 度量文件间的相关性

既然已经有了文件间关联距离的概念和计算方法,就可以用它度量文件间的相关性.为了向正在访问某一文件的用户推荐该文件的相关文件,就必须考查该文件与其他文件的相关程度.我们将目标文件f_{dest}与其他文件f_x的同主题相关程度记为$R_{dest}(f_x)$.假如文件f_x比文件f_y对目标文件f_{dest}而言更加相关,则记为$R_{dest}(f_x)>R_{dest}(f_y)$;如文件$f_y$比文件$f_x$对目标文件$f_{dest}$而言更加相关,则记为$R_{dest}(f_x)<R_{dest}(f_y)$;如文件$f_x$与文件$f_y$对目标文件$f_{dest}$而言同等相关,则记为$R_{dest}(f_x)=R_{dest}(f_y)$.目标文件与其他文件的同主题相关程度和它与其他文件的关联距离之间的关系满足下述属性:

(1) 对于目标文件f_{dest}:

$$D(f_{dest},f_x)<D(f_{dest},f_y)=>R_{dest}(f_x)>R_{dest}(f_y);$$

$$D(f_{dest},f_x)>D(f_{dest},f_y)=>R_{dest}(f_x)<R_{dest}(f_y);$$

$$D(f_{dest},f_x)=D(f_{dest},f_y)=>R_{dest}(f_x)=R_{dest}(f_y).$$

(2) 两个不同的目标文件f_{dest1}与f_{dest2}与其他文件的同主题相关程度不具备可比性.

如图2所示,对于目标文件f_0,$D(f_0,f_5)<D(f_0,f_1)=D(f_0,f_6)<D(f_0,f_2)=D(f_0,f_7)<D(f_0,f_4)<D(f_0,f_3)$,故$R_0(f_5)>R_0(f_1)=R_0(f_6)>R_0(f_2)=R_0(f_7)>R_0(f_4)>R_0(f_3)$;

对于目标文件f_{10},$D(f_{10},f_3)=D(f_{10},f_8)<D(f_{10},f_9)<D(f_{10},f_2)$,故$R_{10}(f_3)=R_{10}(f_8)>R_{10}(f_9)>R_{10}(f_2)$.

但由$D(f_0,f_3)<D(f_{10},f_3)$并不能得到$R_0(f_3)>R_{10}(f_3)$,因为f_0和f_{10}是两个不同的目标文件,f_0和f_{10}与f_3的同主题相关程度不具备可比性.

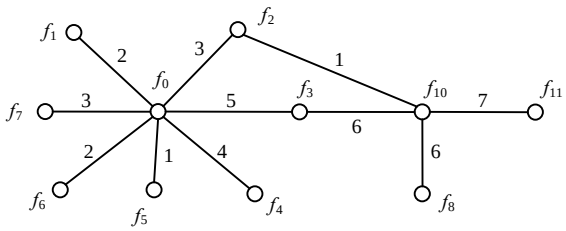


Fig.2 Distances between different target files and other files

图 2 不同目标文件与其他文件的关联距离

2 用户界面

IFRDT 的用户界面如图 3 所示.通过探测用户正在访问的目标文件,IFRDT 将与目标文件最相关的文件以精简的悬浮界面形式呈现给用户.如果用户当前并没有访问任何文件,IFRDT 就将显示用户最后关闭的几个文件.用户可以在智能文件推荐界面中直接双击打开所需文件.如果用户认为 IFRDT 推荐的某些文件与当前正在访问的文件并无关联,就可以通过取消智能文件推荐界面中不相关文件前 Check 框的勾选并点击 refresh 按钮刷新列表来达到评价推荐文件与当前访问文件相关性的目的.如用户无法在智能文件推荐界面中找到所需文件,则将前往真实文件系统中找到并打开,这一动作会被 IFRDT 记录并作为进一步推荐的依据.

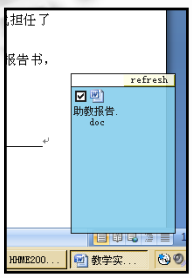


Fig.3 User interface of IFRDT

图 3 IFRDT 的用户界面

3 用户实验

本实验是为评价 IFRDT 的可用性而设计的.我们准备了针对 3 个不同任务的 3 张表格,每张表格都有若干问题需要被试者填写,这些问题的填写需要被试者参照计算机的若干个不同位置的文件才能完成,并且需要填写的内容足够长而难以记忆.完成这 3 个表格的填写分别需要被试者打开 6 个/10 个/15 个文件.

在实验前,被试者将被告知每个文件都有可能在表格中对应多个问题,以降低被试者完成一个问题就关闭一个文件的意愿.被试者每次填写完表格都必须关闭所有访问过的文件.在预实验中,我们要求 5 名被试者在 3 个不同的时间段内分别填写完这 3 个表格.预实验完成后,正式实验分两次进行.在第 1 次正式实验中,我们要求这 5 名被试者完成与预实验同样的工作,与此同时,在被试者使用的计算机屏幕右下方将呈现与 IFRDT 相同的界面,该界面中展示的是被试者最后访问过的历史文件,被试者可以选择从访问历史中直接打开历史文件,或前往计算机目录找到再打开所需文件,第 1 次实验将成为用户实验的 baseline;在第 2 次正式实验中,我们也要求被试者完成与预实验相同的工作,与此同时,在被试者计算机屏幕右下方呈现的是 IFRDT 的推荐界面,被试者可以选择从 IFRDT 的推荐结果中直接打开推荐文件,或前往计算机目录找到再打开所需文件.我们要求被试者保证相同任务下预实验和正式实验的顺序,即保证在相同任务下,预实验发生在两次正式实验之前,而对任务进行的顺序则不加限定.每个实验进行的时间并不连续,用户可以在实验的间歇用计算机做自己的其他事情.

最后我们得到如图 4~图 6 所示的结果.图 4 展示了使用两种不同的推荐方法,任务 1 的实验结果对比;图 5、图 6 分别为任务 2、任务 3 的实验结果对比.两种柱状体分别表示被试者直接从访问历史或 IFRDT 的推荐结果中打开的文件占全部打开文件(包括表格文件本身)的比率,其中,实心柱状体表示向被试者推荐访问历史的实验结果,斜纹柱状体表示向被试者呈现 IFRDT 推荐结果的实验结果.可以看到,访问历史的推荐结果百分比有时为 0,这是因为历史记录太长反而更加难以查找到所需文件,从而迫使被试者前往计算机目录查找文件;有时却接近 100%,这是因为历史记录中的文件数目并不受限,如果被试者执着地在历史中寻找文件,那么只要找到与

任务相关的其中一个文件也就找到了全部相关文件.因此,依靠历史记录查找文件,其实验结果并不稳定,历史记录并不是方便用户查找文件的最佳途径.而 IFRDT 的推荐结果在整体上优于访问历史的推荐结果,且效果较为稳定.随着任务 1~任务 3 所需要的文件数增多,IFRDT 实验结果整体呈下降趋势.我们认为用户在完成任务时需要用到 15 个以上的文件的概率是较低的,大多数任务并不需要太多的相关文件.根据 IFRDT 的推荐,至少有一半的文件可以由用户直接在 IFRDT 中打开,为用户节省了至少一半的查找开销.

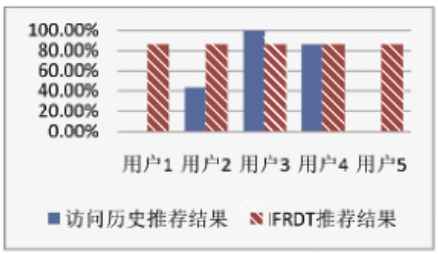


Fig.4 Contrast of results for Task 1

图 4 任务 1 的实验结果对比

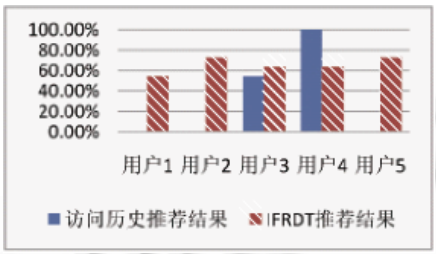


Fig.5 Contrast of results for Task 2

图 5 任务 2 的实验结果对比

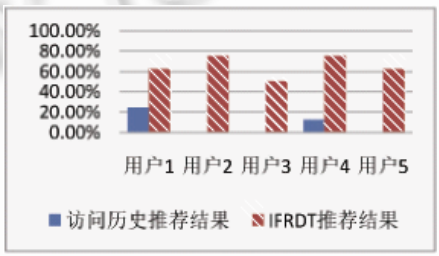


Fig.6 Contrast of results for Task 3

图 6 任务 3 的实验结果对比

4 总结与展望

本文提出了智能文件推荐算法,设计实现了智能文件推荐桌面工具 IFRDT,利用用户访问文件时的同主题特性,通过记录用户访问文件的历史并加以学习来向用户推荐可能将被访问到文件,以达到减少用户查找文件时间开销的目的.5 位被试者的实验结果表明,使用 IFRDT 向用户推荐文件比仅仅向用户呈现访问历史更能为用户节省查找文件的时间,IFRDT 可以为用户节约一半以上的查找开销.用户除了正常的打开/关闭文件操作外,无须进行其他额外操作.

本文提出的智能文件推荐算法主要基于文件访问时间,并未考虑到文件的实质内容等因素.但很明显,文件内容也是影响文件关联程度的重要因素.我们的未来工作将集中在向算法中加入文件内容或标签因子,改善智能文件推荐算法.文件标签除了来自用户的手动文件,还可以来自文件标题及内容的关键词提取.一个文件的手动添加标签数目的多少可以决定这个文件对用户的重要程度.因为越重要的文件,用户为它添加标签的意愿可能越高.最后,除了本文已经实现的智能文件推荐算法,未来还希望实现标签推荐,以最大程度地降低用户的额外工作量.用户除了能够评价 IFRDT 推荐的文件之外,也可以对推荐的标签进行评价.

References:

[1] Boardman R, Sasse MA. “Stuff goes into the computer and doesn’t come out”: A cross-tool study of personal information management. In: Proc. of the CHI 2004, ACM Conf. on Human Factors in Computing Systems. 2004.

[2] Barreau D, Nardi BA. Finding and reminding: File organization from the desktop. ACM SIGCHI Bulletin, 1995,27(3):39–43.

[3] Jul S, Furnas GW. Navigation in electronic worlds: Workshop report. ACM SIGCHI Bulletin, 1997,29(2):44–49.

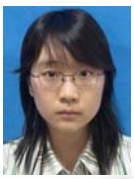
[4] Ko AJ, Aung HH, Myers BA. Eliciting design requirements for maintenance-oriented IDEs: A detailed study of corrective and perfective maintenance tasks. In: Proc. of the Int'l Conf. on Software Engineering. St. Louis, 2005. 126−135.

[5] Dourish P, Edwards WK, LaMarca A, Lamping J, Petersen K, Salisbury M, Terry DB, Thornton J. Extending document management systems with user-specific active properties. ACM Trans. on Information Systems, 2000,18(2):140−170.

[6] Dumais S, Cutrell E, Cadiz J, Jancke G, Sarin R, Robbins DC. Stuff I've seen: A system for personal information retrieval and re-use. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2003). 2003. 72−79.

[7] Cutrell E, Robbins DC, Dumais ST, Sarin R. Fast, flexible filtering with Phlat—Personal search and organization made easy. In: Proc. of the CHI 2006, ACM Conf. on Human Factors in Computing Systems. 2004. 261−270.

[8] Bao XL, Herlocker JL, Dietterich TG. Fewer clicks and less frustration: Reducing the cost of reaching the right folder. In: Proc. of the 11th Int'l Conf. on Intelligent User Interfaces. 2006. 178−185.



韩爽(1983－),女,浙江湖州人,硕士,主要研究领域为人机交互.

王衡(1960－),女,博士,副教授,主要研究领域为人机交互.