结合用户交互行为和资源内容的资源推荐

杨智强[1,2)]，殷 钊[1,2)*]，王 衡[1,2)]
[1)](北京大学信息科学技术学院 北京 100871)
[2)](北京市虚拟仿真与可视化工程技术研究中心(北京大学) 北京 100871)
(yinzhao@graphics.pku.edu.cn)

摘 要：为了更好地管理任务以及与任务相关的资源，使用户集中注意力在任务本身上，减少用户的交互负担，提出一种基于隐式 Dirichlet 分配(LDA)模型的任务建模方法.通过将用户的交互行为按时间片进行切分，实现了时间片序列—任务—文件与 LDA 模型中的文章—主题—单词的对应，经过 LDA 方法的学习，得到了时间片—任务的概率分布和任务—文件的概率分布；为了对任务模型进行补充，进一步提出了基于资源内容的主题分析方法，并用 LDA 方法建立了主题模型；最后通过对资源的关联关系分析，实现了一个结合任务模型和主题模型的资源推荐系统.实验结果表明，任务模型能够有效地发现用户的主要任务和主要文件.

关键词：任务模型；内容分析；隐式 Dirichlet 分配；资源推荐
中图法分类号：TP391

Providing Resource Recommendation Based on Interactive Behavior and Resource Content

Yang Zhiqiang[1,2)], Yin Zhao[1,2)*], and Wang Heng[1,2)]
[1)](*School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871)
[2)](*Beijing Engineering Technology Research Center of Virtual Simulation and Visualization (Peking University), Beijing* 100871)

Abstract: To improve management of tasks and related resources, help users increase their concentration on tasks and reduce their interaction burden, a new task modeling method based on Latent Dirichlet Allocation (LDA) model is proposed. By segmenting the user's semantic behavior according to time slices, the mapping of time slice—task—file in user activity to document—topic—word in the LDA model is realized. After the learning process of LDA method, the probability distribution of time slice to task and the probability distribution of task to file are attained. In order to complement the task model, a topic analysis method based on the content of resources is proposed, and the topic model is built using LDA method. Finally, the relevance of resources is analyzed and a resource recommendation system based on task model and topic model is built. Experimental results show that the task model can find the user's main tasks and main documents effectively.

Key words: task model; content analysis; latent Dirichlet allocation; resource recommendation

科学技术的飞速发展使得人们获取信息变得更加容易，这是人类历史上任何其他时期都不能相比的[1].当今用户的个人信息量越来越大，过量的信息会使人们产生浪费时间、延迟决策、无法专注于主要任务以及压力等问题[2].知识工作者，如教授、律师和工程师等职业的人，对于信息过载的情况感受

收稿日期：2013-08-14；修回日期：2013-10-15.**基金项目**：国家“八六三”高技术研究发展计划(2011AA120301)；国家自然科学基金(60925007,61173080,61232014).**杨智强**(1985—)，男，硕士，主要研究方向为人机交互；**殷 钊**(1989--)，男，硕士研究生，论文通讯作者，主要研究方向为人机交互；**王 衡**(1960—)，女，博士，副教授，CCF 会员，主要研究方向为人机交互、图像处理与计算机图形学.

最为深刻,因为他们在日常工作中需要进行各种不同的任务,而在任务执行的过程中需要查找和处理大量的信息.这就不可避免地产生了一个问题:在任务被打断或者任务切换时,要为获得当前任务的相关信息或者资源付出大量的精力.

个人信息管理就是关于如何帮助人们解决这个问题的研究领域.为用户提供完美的个人信息管理会遇到很多心理学上的挑战,这些挑战可以归结为以下的两点:1)要把物品(如文件)进行分类在认知上是非常困难的;2)用户能够记住的关于物品的细节常常不能够用于检索.当前的研究从应对这两点挑战的角度提出了许多解决方案.

为用户提供更好的信息组织和呈现方式是个人信息管理的重要研究方向.Bergman 等[3]所实现的项目文件夹将用户的所有同主题信息(包括文档、邮件、收藏的页面等)存储于同一文件夹下,用户可以在同一目录下存储和找回同主题信息.

除了更好地对信息进行组织和呈现之外,更强大的信息检索功能也是实现个人信息管理的重要手段.Dumais 等[4]实现了一个 Stuff I've Seen(SIS)系统.SIS 系统的设计有 2 个关键的方面:一是为不同组织结构的信息提供统一的标记,从而利用统一的标记实现统一的检索;另一个是利用如浏览的时间、文件的作者等用户比较容易记住的上下文信息为用户提供检索.

信息组织和呈现的方式需要用户预先把资源进行分类,但未能从根本上把用户从繁重的交互负担中解脱出来.信息检索的方式在一定程度中减小用户查找资源的开销,但是频繁检索会使得用户的任务产生中断,不能集中精力于当前的工作.

为了提供一种全自动的个人信息管理方法,让用户只需专注于任务本身,很多相关研究引入了"任务"的概念,并通过对任务的自动建模来组织和分析交互行为、交互资源之间的相关性.

Shen 等[5]提出了一个根据用户当前正在进行的事件(如编辑 word 文档等)预测用户当前任务的机器学习系统 TaskPredictor.该系统根据当前的活动窗口采集所需的事件信息(当前窗口的名字,文档路径等),并将此事件的特征表示成向量的形式;然后采用了朴素贝叶斯(naïve Bayes)和 SVM 结合的方法来推断当前事件所属的任务.在 TaskPredictor 系统之后,Shen 等[6]又与 IBM 沃森研究中心的研究者提出了一种以活动为中心(activity-centric)的资源推荐方法,该方法不直接推荐其他资源,而是先寻找与当前资源相匹配的活动,再根据该活动推荐相关资源,其中采用了朴素贝叶斯方法来预测资源与当前上下文的相关性.

另外,微软的 SWISH 系统[7]通过识别用户桌面上窗口之间的关系来推断用户当前正在进行的任务.SWISH 以 Windows 事件流的方式监视用户桌面上的活动,获取应用程序窗口的标题和窗口切换历史,对其处理之后再进行聚类,每一类对应相应的用户任务.

Palo Alto 研究中心的 Brdiczka[8]提出了一种根据用户对当前文档的交互行为来恢复用户任务的方法.该方法以完全非监督的方式记录用户对文档的交互行为(如打开、关闭、切换文档等,但不涉及文档的内容),并构建一个任务模型.根据文档获得焦点的时间等信息来求得 2 个文档之间的"相似度分数",并用一个相似度矩阵表示;然后,用谱聚类(spectral clustering)算法对这个相似度矩阵进行处理,最终得到很多组文档,每组内的文档具有较高的相似度,即确定这些文档属于同一个用户任务.

上述研究工作在自动识别用户任务、建立任务模型方面均取得了一定的成果,但在信息关联性分析和资源推荐等方面较少地考虑交互行为及资源文档之间的语义联系.本文采用一种时间片采样的方法对用户的交互行为进行处理,利用隐式 Dirichlet 分配(latent Dirichlet allocation, LDA)模型分析用户的交互行为,为任务建模提供了一种新的思路和方法;并利用 LDA 主题模型对资源内容进行分析,作为任务模型的辅助.基于任务模型和主题模型分析资源之间的关联性,本文提供了一种实时而准确的资源推荐服务.

1 信息关联分析

为了推荐与当前资源相关的其他资源,必须分析资源之间的关联关系.本文从用户的行为和资源的内容这 2 个方向出发,利用 LDA 模型分别建立基于用户交互行为的任务模型和基于资源内容的主题模型,以分析资源之间的关联关系,并最终根据这 2 个模型进行资源推荐.

1.1 LDA

LDA 是一个基于概率的生成模型,可以用来挖掘文章集合中潜在的主题信息;它有 3 个主要的概念,分别是文章、主题和单词.LDA 的基本思路是:文章可以用一组潜藏主题的概率分布来表示,每个

主题能够用词汇表中的词的概率分布来表示[9].

文章是由单词组成的，LDA 生成模型为文章中的所有单词定义了基于潜藏变量的概率采样过程. 假设语料库中有 M 篇文章，M 篇文章中存在 K 个主题，词汇表中有 V 个单词. 在 LDA 模型中文章的生成过程涉及到 2 个概率分布，一个是从每篇文章到 K 个主题的概率分布；另一个是每个主题到 V 个单词的概率分布.

因此，LDA 模型主要需要推断 2 个概率分布，一个是每篇文章 m 的主题分布，记为 $p(z|d=m)=\theta_m$；另一个是每个主题 k 的词的分布，记为 $p(w|z=k)=\varphi_k$. 本文采用 Gibbs 采样[10]方法来推断 LDA 模型中的参数.

1.2　任务模型：交互行为分析

建立任务模型的目的是从用户交互行为的历史中自动挖掘用户的任务. 之所以把挖掘的基础建立在用户的交互行为数据上，是因为用户交互任务必须通过一系列的交互行为来实现，用户的一系列交互行为的背后必然潜藏着用户的交互动机和交互目的. 通常意义上说，在某种情况下经常共同出现的程序和资源应该是服务于同一任务，本文的任务挖掘方法基于此假设.

我们通过采集用户在 PC 上的活动窗口切换记录获取用户的交互行为，每一条记录由一个五元组构成，即〈开始时间，结束时间，活动窗口标题，应用程序名称，关联文件〉；每个交互行为以其使用的应用程序和所关联的文件来表征.

为了建立基于用户交互行为的任务模型，我们需要对用户一段时间内的交互行为进行处理，建立与 LDA 模型中文章—主题—单词 3 层概念的对应关系；然后通过 LDA 方法来挖掘交互行为的历史数据中的交互任务.

任务模型的建立方法如下：

Step1. 时间片采样. 将用户一段时间内的语义行为用时间片切分，每个时间片中的语义行为转化为词频的方式去表示.

首先需要定义 2 个参数：时间片的长度 t 和采样周期 c，表示每过周期 c 的时间就采样一次，这里 t 和 c 的单位都定义为 s；还需要定义采样的向量 $\boldsymbol{A}_j=\{a_1,a_2,\cdots,a_n,\cdots,a_N\}$，其中 $a=\{0,1\}$，n 为语义行为实例的编号，N 为语义行为实例总个数，j 表示第 j 次采样，$\boldsymbol{A}_j$ 反映了第 j 次采样时所有语义行为的状态.

以周期 c 来采样时间片 t，我们可以得到 m 个向量 $\boldsymbol{A}$，其中 $m=\frac{t}{c}$. 最终定义 $\boldsymbol{K}=\sum_{i=1}^{m}\boldsymbol{A}_i$，其中 $\boldsymbol{K}$ 是时间片序列，表示了时间片内每个语义行为的运行状态. 如图 1 所示，假设我们定义 $t=30$，$c=10$，于是 $m=3$. 值得注意的是，代表语义行为的横线在一个时间片内最多与代表采样时间点的竖线有 4 个交点，但是我们只采样 3 次，选择的是时间片中的前 3 个点. 那么，在 0～60 s 的时间内可以得到 $\boldsymbol{K}_1=\{3,2,1\}$，$\boldsymbol{K}_2=\{3,1,3\}$. 通过这种方式，我们就实现了语义行为到词频的转换.

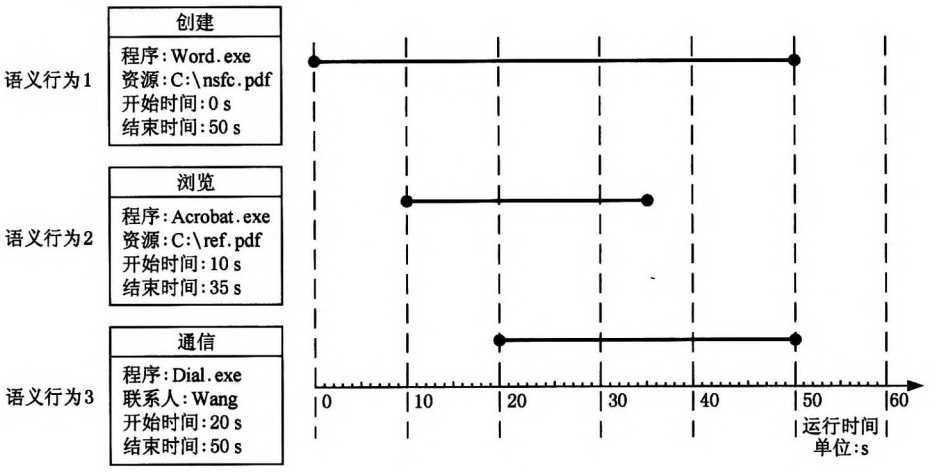

图 1　时间片采样示例

Step2. 利用 LDA 模型学习任务分布及任务中语义行为的分布.

本文定义待挖掘的任务对应着 LDA 模型中的主题，每个时间片的描述 $\boldsymbol{K}$ 对应着 LDA 模型中的一篇文章，每个语义行为对应着 LDA 模型中的单词. 通过 Gibbs 采样方法，可以求得某个时间片序列中任务的概率分布；同时，也能够获得某个任务中语义行为发生的概率分布. 图 2 说明了这个对应关系.

从交互行为的转化方式可以看到，时间片对应的向量描述了用户在某个特定时段的交互行为，代表了

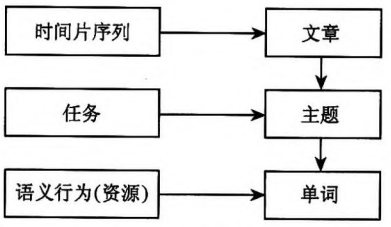

图 2 任务模型与 LDA 模型的对应

用户的一种交互状态. 通过 LDA 模型的学习方法，可以得到用户在这样一种状态下的任务分布概率. 通过任务分布概率，可以衡量不同时间片所代表的用户状态的相似性. 另外，我们通过 LDA 模型得到了任务关于交互行为的分布概率，在同一个任务中，概率高于某一个阈值的交互行为之间具有比较高的相关性，这些交互行为的相关性是由于同属于某个任务而发生的.

1.3 主题模型：资源内容分析

任务模型主要是基于用户的交互行为建立的，并未考虑行为所涉及的资源的内容. 然而，资源内容的之间的关联性可以作为信息关联的一个衡量标准. 为了对任务模型进行补充，本文建立了基于资源内容的主题模型来分析文件之间的关联关系.

最简单的内容分析方法是使用 TF-IDF 来衡量 2 个文件之间的相似性. TF(term frequency)是词频，指的是单词在文件中出现的次数；IDF(inverse document frequency)逆文本频率，指的是所有文件数除以出现过该单词的文件数. TF-IDF 方法的优点是简单、易实现，但是它不能有效地分析文章的语义信息.

所以，为了挖掘文件的语义信息，特别是挖掘不同文件之间的主题信息，本文决定采用 LDA 主题模型分析内容.

基于文件内容的主题发现就是从文件内容的角度进行分析，发现文件所潜藏的主题信息. 得到主题之后，可以进一步利用文件的主题信息来衡量文件之间的相关性. 文件内容的预处理过程如图 3 所示.

Step1. 去除标点符号. 标点符号没有实际意义，对内容分析会产生负面影响，需要预先把它们去除.

Step2. 中文分词. 中文并不像英文一样具有天然的分隔符，因此需要对文章内容进行分词. 本文采用中国科学院计算技术研究所的汉语词法分析系统 NLPIR① 进行分词，可以得到比较准确的结果.

Step3. 根据停用词表去除停用词，得到词汇表. 不是所有的词都能用来分析，必须删除一些停用词(stop words)[11]；删除的方法主要是依靠停用词表，最终，可以得到整个文章集合的词汇表.

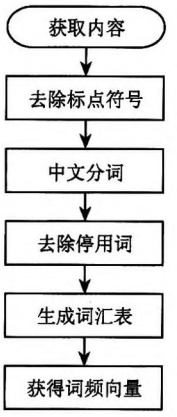

图 3 资源内容的预处理过程

Step4. 统计词汇表，得到文档向量.

经过以上几步，我们得到了每一篇文章的单词集合，通过词汇表即可得到文章的词频向量化表示. 然后，利用 LDA 的方法进行分析，可以得到基于资源内容的主题模型. 主题模型中包括文章—主题分布和主题—单词分布.

1.4 关联分析

在经过任务建模之后，我们可以得到任务—交互行为的分布概率. 在某个任务下出现概率比较高的一些交互行为(资源)，基于这个任务本身而具有一定的关联关系.

当需要衡量任务模型中没有的资源与任务模型中资源的关联关系时，就需要用到时间片的分析方法.

时间片分析是任务模型中的概念. 用户的交互行为涉及了一个新的资源 w 时，由于之前建立的任务模型没有出现过这个资源，所以无法利用任务对各个资源的概率分布进行推荐. 这时需要通过用户在之前一段时间的交互行为来预测用户的当前状态，具体做法如下：

Step1. 把用户之前一段时间的交互行为转换为任务建模中的时间片，这个过程可能需要剔除未知的资源.

Step2. 时间片对应的是 LDA 模型中的文章，所以可以将时间片作为新的文章导入已有模型中计算它的主题分布概率.

Step3. 得到主题分布概率，即任务分布概率后，假设其中概率最高的任务为 k，那么可以认为任务模型中任务 k 所对应的概率较高的资源与这个新的资源 w 具有较强的关联.

对于基于内容的主题模型，在经过 LDA 方法学习之后，代表资源内容的文章相当于从维度很高的

① http://www.nlpir.org

词向量降到了维度比较低的主题向量. 所以，计算 2 篇文章之间的相似性可以通过它们的主题概率分布来计算. 衡量 2 个概率分布的差别的标准方法是计算它们的 KL 距离（Kullback-Leibler divergence，KLD）[12]. 假设 2 个概率分布分别是 $\boldsymbol{p}$ 和 $\boldsymbol{q}$，它们的 KL 距离计算公式为

$$D(\boldsymbol{p},\boldsymbol{q}) = \sum_{i=1}^{T} \boldsymbol{p}_i \operatorname{lb} \frac{\boldsymbol{p}_i}{\boldsymbol{q}_i};$$

其中，$\boldsymbol{p}_i$ 和 $\boldsymbol{q}_i$ 分别表示向量 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 的第 i 维.

KL 距离是非对称的，也就是说 $D(\boldsymbol{p},\boldsymbol{q}) \neq D(\boldsymbol{q},\boldsymbol{p})$. 所以，在实际应用中，常常使用一个基于 KL 距离的对称距离，其计算公式为

$$\mathrm{KLD}(\boldsymbol{p},\boldsymbol{q}) = \frac{1}{2}[D(\boldsymbol{p},\boldsymbol{q}) + D(\boldsymbol{q},\boldsymbol{p})].$$

1.5　资源推荐过程

资源推荐的总体流程如图 4 所示.

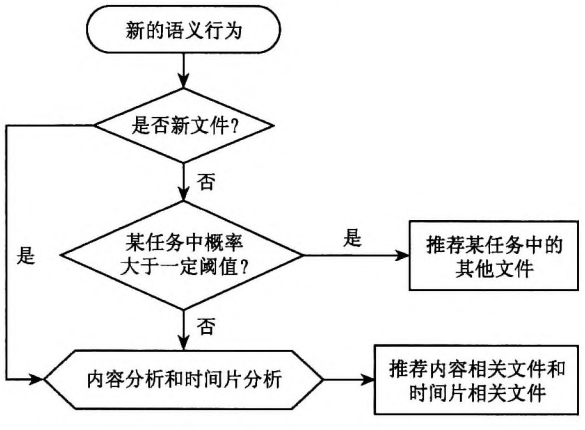

图 4　资源推荐的流程

当新的语义行为到来时，系统首先判断该语义行为所关联的文件是不是新文件；这可以从已经建立的任务模型中查询得知.

对于不是新文件的情况，可以从任务模型中知道该文件在各个任务中出现的概率. 首先找到这个文件的出现概率最高的那个任务，再判断该任务下所对应的概率值是否大于一定的阈值. 这个阈值主要依靠经验来设定，当前设置的阈值是 0.1. 如果大于这个阈值，那么该文件在这个任务出现的概率足够的高，可以把这个任务的其他高概率的文件推荐出来. 因为我们认为，同一个任务下出现概率较高的文件有基于任务的关联关系.

如果这个文件对应的所有任务中的最高出现概率小于设置的阈值或者这个文件是新文件，则系统依靠时间片分析和内容分析来进行推荐. 内容分析会把新文件的标题作为内容 LDA 模型中的新文章来处理，在得到新文章的主题分布概率后，将它与内容 LDA 模型中的所有文章计算 KL 距离，然后把 KL 距离最小的前 N 篇文章所对应的文件推荐出来，一般设置为 $N=5$. 同样，时间片分析也会得到一些推荐文件；最终把时间片分析和内容分析所得到的推荐文件一起推荐出去.

1.6　总体流程

综上，本文所述的基于用户交互行为和资源内容的资源推荐算法总体流程如下：

Step1. 采集用户 PC 上的活动窗口切换记录，以此获取用户的交互行为，并以五元组表示，每个交互行为以应用程序和所涉及的资源表征.

Step2. 对一定时间内的用户交互行为进行时间片采样，得到若干个时间片序列，实现语义行为到 LDA 模型中文章词频向量的转换，并通过 LDA 方法进行学习，得到时间片序列的任务分布及任务中语义行为的分布，建立任务模型.

Step3. 对该段时间内用户交互行为涉及的资源内容，经过处理得到词频向量；通过 LDA 方法进行学习，得到文章-主题分布和主题-单词分布，建立主题模型.

Step4. 当用户使用 PC 工作时，对于当前正在使用的资源，采用 1.5 所述的资源推荐方法向用户实时进行资源推荐.

Step5. 当用户所操作的已有模型中不包括的资源达到一定数量时，模型进行自动更新.

2　系统架构

基于任务模型和主题模型的信息推荐系统的总体架构设计如图 5 所示.

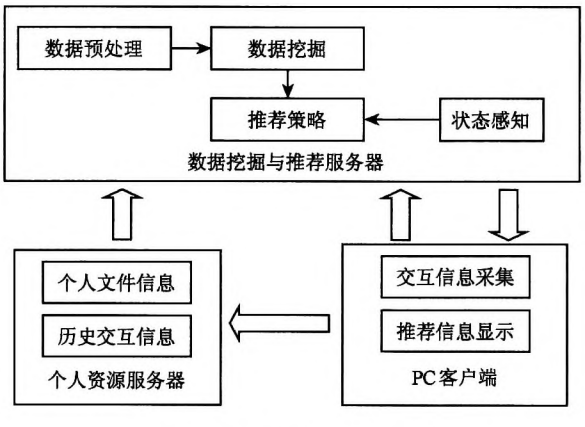

图 5　推荐系统的总体架构

系统主要包括 PC 客户端、个人资源服务器和数据挖掘与推荐服务器. 其中，PC 客户端包括 PC 上的交互信息采集模块和推荐信息显示模块；个人资源服务器主要负责存储用户的个人文件信息和历史交互信息；挖掘与推荐服务器包括交互行为数据

预处理模块、数据挖掘模块、状态感知模块和推荐策略模块，架构图中的箭头表示系统中的信息流向. PC 客户端在采集到用户的文件信息和交互信息之后，以一定的周期把这些信息发送到个人资源服务器中；个人资源服务器把这些信息按照一定的格式保存到相应的数据库中，以供数据挖掘与推荐服务器的查询. 数据挖掘与推荐服务器从个人资源服务器中获取数据，得到数据后，一般是先经过数据预处理模块的处理，然后把特定格式的交互数据传给数据挖掘模块. 数据挖掘模块对交互数据经过预定算法的学习之后生成模型；状态感知模块需要从 PC 客户端获取用户的实时信息；推荐策略模块结合数据挖掘模块生成的模型和状态感知模块提供的实时信息，根据预定的策略生成推荐结果，然后把结果发送到 PC 客户端的推荐信息显示模块. 最后，PC 客户端的推荐信息显示模块把信息显示出来.

3 用户实验

在完成信息推荐系统的设计与实现之后，本文设计了实验对系统的任务建模的功能进行验证.

3.1 实验目的

任务是与人的认知有关的一个概念. 对于一个普通用户而言，把日常在计算机上做的所有事情清楚地划分成不同的任务是是一件十分困难的事情. 但用户在平时的工作中，一般能够知道他主要做了什么事情，而且这些主要的事情对于用户来说才是更重要的.

所以，本文所设计的实验要达到的实验目的是：检验生成的任务模型中是否包含了用户指定的主要任务，在得到的任务模型中主要文件的重要性是否得到了体现，以此验证基于 LDA 方法的任务模型的有效性.

3.2 实验方法

首先需要定义主要任务和主要文件的概念，它们是由用户主观指定的指标.

主要任务是在交互数据的采集过程完成以后，用户要为测试数据指定 K 个主要任务，K 由用户根据自身实际情况指定，每个任务由若干个文件来表示.

主要文件是用户指定的主要任务中所包含的文件.

实验主要步骤如下：

Step1. 采集用户一段时间的交互数据.

Step2. 让用户指定这段时间内的主要任务和主要文件（通常需要让用户观察交互数据帮助回忆）.

Step3. 利用任务建模的方法得到任务模型.

Step4. 把用户指定的主要任务和主要文件与任务建模得到的任务进行对比，计算主要任务和主要文件被发现的百分比.

3.3 实验的评价标准

为了评价实验的结果，需要定义主要文件被发现的依据和主要任务被发现的依据. 在完成任务建模之后，得到的其中一个分布是任务—文件的概率分布，即给定某个任务时出现某文件的概率. 在本文对某个任务下的文件进行推荐时，会把某个任务下出现概率最高的一些文件给推荐出去. 所以，在本文实验中，只要某个文件在某个任务中的出现概率比较靠前，就可以认为它被发现了. 而对于主要任务而言，如果它所对应的主要文件在同一个任务中被发现的数目超过一半，则认为该主要任务被发现. 在这个前提之下，我们对主要文件、主要任务的发现依据定义如下：

主要文件被发现的依据. 如果主要文件在某一任务中概率最高的前 K 个文件中，则认为该主要文件被发现了.

主要任务被发现的依据. 如果用户指定的某个主要任务中的主要文件有超过一半在同一个任务中被发现，则认为该主要任务被发现.

实验的评价指标有 2 个，分别是主要文件被发现的百分比和主要任务被发现出来的百分比.

3.4 实验结果

实验邀请了 8 位被试，都是计算机系的学生. 实验时间的跨度从 3～15 h 不等，在采集数据结束以后马上进行反馈调查，被试应该能够清楚地记得实验期间所进行的任务以及与任务相关的文件. 得到的用户数据的时间跨度与文件数统计如表 1 所示，文件数指的是实验时间内用户所用过的所有文件的数目.

表 1 用户使用时间与文件数统计

用户	使用时间/min	文件数
1	176	96
2	440	214
3	165	56
4	917	291
5	194	75
6	274	103
7	367	189
8	256	153

表 2 和表 3 分别是主要任务和主要文件的统计表. 考虑到短时间内用户的任务数不会太多，在实验过程中，系统中 LDA 模型的主题数都是设置为 5.

表 2　主要任务统计

用户	主要任务数	被发现的主要任务数	被发现的比率/%
1	4	3	75
2	5	3	60
3	3	2	66.6
4	5	2	40
5	2	1	50
6	4	3	75
7	5	4	80
8	4	3	75

表 3　主要文件统计

用户	主要文件数	被发现的主要文件数	被发现的比率/%
1	13	9	69.2
2	14	6	42.8
3	7	4	57.1
4	12	6	50
5	8	6	75
6	11	7	63.6
7	16	10	62.5
8	13	8	61.5

3.5　实验分析

通过实验结果可以看到，本文所提出的使用 LDA 方法建立的任务模型具有比较高的有效性，能够准确识别用户交互过程中的主要任务以及主要任务中所使用的主要资源. 加之基于资源内容的主题模型的辅助，能够反映出最终的资源推荐具有较高的准确率和可用性.

根据实验过程中用户的反馈和对实验结果的分析发现了建模过程中的一些问题，并得到了许多关于如何改进的启发.

1) 任务是主观的事情，不容易衡量和判断. 如说，某个用户认为他有 2 个任务是可以相互区分开的，但是这 2 个任务在进行的过程中会互相交叉，所以对任务数的判断非常不容易.

2) 用户会被不断地打断和切换任务，使用户切换任务的事情可能是经常变化的弹出窗口，如聊天窗口、新闻窗口等，这些数据对任务建模的效果产生了一定的负面影响.

3) 用户专注的任务可以比较容易的发现出来，但是，如果用户的任务比较分散，而且同一个任务使用的文件也比较分散时，就比较难检测用户的任务已经相关的文件.

4) 数据预处理非常重要，采集的交互数据中可能会大量出现没有意义的信息，如桌面的活动信息，必须把这些信息去除，否则会对挖掘结果产生较大的负面影响.

4　总　　结

为了支持以用户任务为中心的信息关联与信息获取，本文提出了基于时间片方式的用户交互行为处理方法，并将 LDA 方法应用于基于用户行为的任务模型的构建. 实验证明了这种任务建模方法的有效性. 本文还建立了基于资源内容的主题模型，并根据这 2 个模型分析了资源之间的关联关系，最后实现了结合任务模型和主题模型的资源推荐系统.

参考文献(References)：

[1] Feather J. The information society: a study of continuity and change [M]. London: Library Association, 1998

[2] Waddington P. Dying for information? a report on the effects of information overload in the UK and worldwide [OL]. [2013-08-12]. http://old.cni.org/regconfs/1997/ukoln-content/repor~13.html

[3] Bergman O, Beyth-Marom R, Nachmias R. The project fragmentation problem in personal information management [C] //Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2006: 271-274

[4] Dumais S, Cutrell E, Cadiz J J, *et al*. Stuff I've seen: a system for personal information retrieval and re-use [C] // Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2003: 72-79

[5] Shen J Q, Li L D, Dietterich T G, *et al*. A hybrid learning system for recognizing user tasks from desktop activities and email messages [C] //Proceedings of the 11th International Conference on Intelligent User Interfaces. New York: ACM Press, 2006: 86-92

[6] Shen J Q, Geyer W, Muller M, *et al*. Automatically finding and recommending resources to support knowledge workers' activities [C] //Proceedings of the 13th International Conference on Intelligent User Interfaces. New York: ACM Press, 2008: 207-216

[7] Oliver N, Smith G, Thakkar C, *et al*. SWISH: semantic analysis of window titles and switching history [C] // Proceedings of the 11th International Conference on Intelligent User Interfaces. New York: ACM Press, 2006: 194-201

[8] Brdiczka O. From documents to tasks: deriving user tasks from document usage patterns [C] //Proceedings of the 15th International Conference on Intelligent User Interfaces. New York: ACM Press, 2010: 285-288

[9] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3 (Jan): 993-1022

[10] Heinrich G. Parameter estimation for text analysis [OL]. [2013-08-12]. http://faculty. cs. byu. edu/~ringger/CS601R/papers/Heinrich-GibbsLDA. pdf

[11] Hua Bolin. Stop-word processing technique in knowledge extraction [J]. New Technology of Library and Information Service, 2007, 2(8): 48-51 (in Chinese)
(化柏林. 知识抽取中的停用词处理技术[J]. 现代图书情报技术, 2007, 2(8): 48-51

[12] Steyvers M, Griffiths T. Probabilistic topic models [M] // Landauer T, McNamara D, Dennis S, *et al*. Latent Semantic Analysis: A Road to Meaning. New Jersey: Lawrence Erlbaum, 2006

结合用户交互行为和资源内容的资源推荐

作者： 杨智强， 殷钊， 王衡， Yang Zhiqiang， Yin Zhao， Wang Heng
作者单位： 北京大学信息科学技术学院 北京 100871 ;北京市虚拟仿真与可视化工程技术研究中心(北京大学)北京 100871

刊名： 计算机辅助设计与图形学学报 ISTIC EI PKU
英文刊名： Journal of Computer-Aided Design & Computer Graphics
年，卷(期)： 2014,26(5)

参考文献(13条)

1. 查看详情

2. Feather J The information society:a study of continuity and change 1998

3. Waddington P Dying for information? a report on the effects of information overload in the UK and worldwide 2013

4. Bergman O;Beyth-Marom R;Nachmias R The project fragmentation problem in personal information management 2006

5. Dumais S;Cutrell E;Cadiz J J Stuff I've seen:a system for personal information retrieval and re-use 2003

6. Shen J Q;Li L D;Dietterich T G A hybrid learning system for recognizing user tasks from desktop activities and email messages 2006

7. Shen J Q;Geyer W;Muller M Automatically finding and recommending resources to support knowledge workers' activities 2008

8. Oliver N;Smith G;Thakkar C SWISH:semantic analysis of window titles and switching history 2006

9. Brdiczka O From documents to tasks:deriving user tasks from document usage patterns 2010

10. Blei D M;Ng A Y;Jordan M I Latent Dirichlet allocation 2003(Jan)

11. Heinrich G Parameter estimation for text analysis 2013

12. 化柏林 知识抽取中的停用词处理技术 2007(08)

13. Steyvers M;Griffiths T Probabilistic topic models 2006