第 2 章　基于概率潜在语义的用户模型构造

2.1　问题提出

随着网络技术的飞速发展，Web2.0 技术的逐步成熟，网络资源以其时效性、互动性、开放性、浩瀚性等特点让用户目不暇接。用户主观参与的同时，个性化服务的需求也越来越突出，而个性化服务首先需要建立用户模型，根据用户不同的偏好模型才能为其提供符合其偏好的个性化服务。用户模型的创建与更新是个性化服务的基石。用户建模过程实质是一个知识获取过程，即以某种可计算的方式对用户的需求进行描述，借以将隐含在用户心理的知识或语义意图显式的外化为可运算的模型。用户模型的研究并不是新的研究领域，自 20 世纪 80 年代，人们便在研究中不断探索与用户建模相关的技术，如人工智能技术，人机交互技术，教育学和心理学，认知理论以及其他领域的研究[21,22]，二十多年来人们取得了丰厚的研究成果。

但在取得长足进步的同时，有一个不确定性问题始终没有得以完美的解决，就是查询者所要表达的真实查询意图以及文档作者所要表达的文档主题无法真正获知，即语义不确定性问题。因而，导致无法精确计算哪些文档是用户所要查询的文档，而这恰恰是个性化服务最为核心的问题。

为此，人们想出很多办法，其中，最经典的办法就是向量空间模型（Vector Space Model），这是上个世纪 Salton 提出的用于表示文档语义的向量模型[23,24]。假设文档集包含 n 个词语，VSM 把每个词语看作一个向量，则由这 n 个词语所对应的向量可以生成一个 n 维欧式空间，这个空间被称为检索词空间。文档集中的任一文档（根据它包含的词语）可被表示为检索词空间中的一个向量。同样查询也可被表示为检索词空间中的一个向量。将待检索文档和查询分别用向量来表示是建立向量空间模型的基本前提，这样查询和文档间的匹配问题转化为一个向量空间中的距离计算问题，因此，VSM 有了真正的相关度的概念。但 VSM 模型的问题是，如果用户查询中没有使用文档集中的关键词会怎么样，很显然会导致无法找到与用户查询相关的文档了，会导致检索效率极低。虽然 Salton 后来将传统

的向量空间模型和布尔检索模型结合，提出了扩展的布尔检索模型，即解决了传统的布尔模型没有相关度概念的缺点，又弥补了向量空间模型对包含布尔算符“与”和“或”查询表达式的解决方法，同时进行了词语加权的改进，并且也应用到了著名的 Smart 系统中[24]，但问题始终是出在没有将文档的语义表达出来，只是停留在 Bag of Words 上，即只以关键词的词频和权重作为构造用户模型的基础，检索操作也在这个基础上完成的，而要完成个性化检索时，这些信息就显得无能为力。

为了准确获取用户搜索的语义意图，我们必须先了解用户搜索的过程的心理及认知上的变化，因为搜索的本质是用户将自己的思维进行展开，抽象成关键词。这些关键词是体现在用户头脑中的概念。这些概念都具有内涵和外延，并且随着主观、客观世界的发展而变化。同时，由于用户在搜索之前是处于“知识的非常态”[25]，他们不清楚自己想要的文档具体的存在形式和包含的词语，并且由于词语用法的多样性和词语使用的随意性以及自然语言的二义性常造成检索的多样性，比如，同一个事物在不同用户的头脑中可能反应出的是不同的关键词，如：“笔记本”和“掌上电脑”，“摄像头”和“视频”，“2008 奥运会”和“北京奥运会”等。并且，相同的词汇又具有不同的含义。如一些文本出于修辞的需要，为了避免重复、单调，使用一些同义词替换；另外就是相同的词汇在不同的上下文或者是不同的领域中，所代表的意义也可能是不一样的，比如 Java 在计算机领域中是一种高级语言，而在其他领域表示是“咖啡”，这些问题就造成了用户信息获取方面的困难。

我们只有获取除关键词词频、权重以外的语义信息，并且把这些信息应用到检索模型中，才能完成真正意义上的个性化信息搜索。目前，潜语义分析技术（Latent Semantic Analysis，LSA）[26-28]就是应用最广泛的用来发掘文档潜在的概念的一种新型数据方法。LSA 对词－文档矩阵进行奇异值分解（Singular Value Decomposition，SVD），并提取前 k 个最大的奇异值及其对应的奇异矢量作为文档集合中存在的潜在概念。其基本原理是把每个文档视为以词语为维度的空间中的一个点，一个包含语义的文档出现在这种空间中，它的分布绝对不是随机的，而是服从某种语义结构。同样地，也可将每个词语视为以文档为维度的空间中的一个点。文档的语义是由词语组成的，而词语又要放到文档中去理解，体现了一种“词语－文档”双重概率关系。这种语义结构隐藏于文本当中，潜在地对词语的出现和文档的构成发挥作用，但是由于词语使用的随意性和文档主题的不确定性等因素的存在，这种语义结构被“噪声”所淹没。LSA 利用奇异值分解降秩的方

法达到信息抽取和去除噪声的目的，LSA 不同于向量空间模型（VSM）中文档的高维表示，而是将文档的高维表示投影在低维的潜在语义空间中，缩小了问题的规模，并且使得原本稀疏的数据变得不再稀疏，从而呈现出一些潜在的语义结构。其隐含的思想是，通过语义处理给定词的所有上下文，提取了决定词语语义的相关性的相互限制[29]。虽然 LSA 为挖掘文档的潜在语义和研究人脑认知过程开辟了一条可行的途径，但其不足是无从对空间中的数据作出语境上的说明，缺乏先验信息的植入而使其显得过于机械。而概率潜在语义分析却可以很好的克服 LSA 的不足。基于此，本文提出采用概率潜在语义分析的方法来解决用户的动机分析和文档的潜在语义挖掘的问题。在本章中，主要讨论基于概率潜在语义分析方法 PLSA[30,31]的用户模型的构建。

2.2 用户模型研究综述

用户模型的研究是一个既经典又富有挑战的研究领域，从二十世纪八十年代开始它伴随着个性化搜索服务的研究从没有停息过，到现在的 Web2.0，它始终都将是个性化搜索过程的一个永恒的研究热点。

2.2.1 用户模型的创建技术研究

在个性化搜索中，由于用户处于“知识的非常态”，很难准确地、完整地描述其自身的兴趣偏好，所以，个性化搜索系统通常将用户的兴趣偏好保存下来，依此为用户提供符合其偏好的个性化搜索。用户模型是系统为用户保留的偏好信息描述文件。目前，用户模型的定义、更新及其功能还没有一个统一的标准，但根据现有的研究可以看出，用户模型的创建涉及两个方面内容：①用户模型创建的信息来源；②用户模型的数据结构。

1. 创建用户模型的信息来源

为用户建立其兴趣的描述，就要获取其兴趣所在，通常的获取方式有两种：

其一，系统以显式的方式与用户进行交互获取，这种方式要求用户要积极的配合，避免获取的信息不准确及负反馈的发生；通常也叫显式相关性反馈，可以由系统首先提出一些初始的问题，由用户作初步回答，根据用户的回答，启发式进入后续的一系列问题。如 Google 个性化搜索[32]；或者，让用户注册一些统计信息等，如 Yahoo 个性化搜索[33]。这种方法获取信息的准确度高，耗时少；但是，过多的询问会给用户增加负担，用户往往胡乱选择一些信息，这样，用户模型的质量可想而知。

此外，由于这种方式生成的用户模型是静态的与用户兴趣的动态性形成矛盾。

其二，系统以隐式方式或者说无入侵的方式对用户兴趣进行跟踪、分析、挖掘来获取。隐式收集用户兴趣，通常也叫隐式相关性反馈（也包括伪反馈），是目前个性化搜索研究中具有潜力的研究方向[34]。根据所挖掘信息内容的不同，大致分为如下几种：

（1）针对用户网络使用信息进行挖掘（Web Usage Mining）[5,21]，生成用户模型。通常，对网络使用信息进行挖掘，一般的做法是先将用户会话（Session）进行聚类分析，找出用户所属的类，然后，以类的质心为基础，基于任务级（Task）或网页级（Page）建立用户模型，粒度较粗。

（2）针对用户点击流数据进行分析、挖掘[22,34]，生成用户模型。对于点击流数据的分析，通常做法是将用户、查询、及点击的网页三者作为共现（Co-occurrence）数据来考虑，对于共现数据的分析常用的方法是潜在语义分析和概率潜在语义分析。对建立好的共现矩阵，进行维的约减，找到其潜在的语义空间，用户模型就是建立在这个潜在的语义空间上。

（3）通过对用户查询历史[35,36]或浏览历史[23,24,37]进行分析处理，生成用户模型。这种方式获取用户的兴趣往往是先定义好领域本体知识，然后，利用领域本体知识和一些关键词/短语建立映射，通过用户反馈建立用户兴趣评价，最终建立基于用户兴趣的层次树的用户模型。此种方法建立的用户模型可以细化到加权关键字，所以，用户模型粒度较细。

（4）通过对网络社区信息进行分析[24,26]，生成用户模型。这种方法建立的用户模型可以体现用户兴趣的多样性，用户模型通常是用于信息过滤。

（5）通过对客户端（Cookie）的一些交互信息分析[27,28]，建立用户模型等。

隐式反馈中（包括伪反馈），是无入侵方式获取用户信息无需用户参与，一切全在用户未知的情况下进行，不会对用户造成干扰。但这种方式获取的信息与用户实际兴趣往往有一定偏差，而且，耗时较长。所以，在通常情况下，显式与隐式信息获取方式相结合，才能达到最佳效果。这样，既可以避免隐式反馈的冷启动问题，也可以避免长时间打扰用户，造成负担。

2. 用户模型的描述形式

用户模型的描述形式，即用户模型的存储数据结构，是个性化搜索的重要的环节之一，通用的存储形式大致有如下几种：

（1）基于加权关键词向量的存储形式。

加权关键词向量的具体形式为：$\{(k_1,w_1),(k_2,w_2),\ldots,(k_n,w_n)\}$其中，$k$ 为关键词，

w 为对应关键词的权重。这种表示方式源于向量空间模型（Vector Space Model）。VSM 把每个关键词视作一个向量，如果某文档中有 n 个关键词，则由这 n 个关键词所对应的向量可以生成一个 n 维欧式空间，这个空间被称为检索词空间。文档集中的任一文档均可被表示为检索词空间中的一个向量。同样查询也可被表示为检索词空间中的一个向量。当检索相关文档时，可将待检索文档向量和查询向量利用相似度函数计算两向量间的距离，将距离大于某一阈值的待检索文档返回给用户。著名的 IR 系统 Smart 就是一个 VSM 的典型实例[31]。

利用这一原理，用户模型中的向量可以被认为是查询空间的一个向量，将一切用户模型与文档的匹配问题转化为一个向量空间中的距离计算问题，但同时，它也继承了 VSM 的缺点，即如果某文档与用户模型没有共同的关键词出现，则认为它们之间没有相似度。这实质上是抛弃了语义具有相关性的文档。

（2）基于概率分布的存储形式。

为了表示用户对多领域的多兴趣需求，基于向量空间模型的存储形式无法实现用户的多种兴趣需求的表示，所以，在用户模型中往往用概率分布的形式来表达用户的这种多兴趣需求。其具体存储形式为 $\{p(c_1 \mid u), p(c_2 \mid u), ..., p(c_n \mid u)\}$，其中，$p(c_i \mid u)$ 表示用户在第 i 个领域的兴趣概率分布，文档、查询也可表示成这种形式。

概率分布的存储形式可以大大提高响应时间，提高查询效率。其优点是可以充分表达用户兴趣的多样性，其不足是，在处理用户兴趣时，必须有事先分好的分类，这需要与分类方法相配合或需要专家的参与；如果与分类方法相配合，则在自身查找准确率的基础上必须乘以分类方法的准确率，这样会降低查找的准确率。

（3）基于分类层次树的存储形式。

为了提供一个具有鲁棒性的上下文，文献[38,39]将用户特定兴趣抽取出来，看起来似乎具有完整上下文的兴趣树，这棵兴趣树的不同结点表示不同抽象程度的用户兴趣，越靠近树根的结点，表示的兴趣越具概括性；越靠近叶子结点，表示的兴趣越具体。并且，结点内容完全是通过隐式相关反馈获取的，可以是网页的主题内容提取，也可以用户的书签内容提取等。

基于这种方法的优点是采用传统的聚类方法就可以解决较复杂的概念层次问题，并且，在实际应用中，由于有了 ODP[①]示范作用，人们很容易接受分类层次的存储形式。但缺点是，这种方法在解决用户存在多兴趣问题时，并且，多兴趣

① Open Directory Project, http://www.dmoz.org/

中有可能几个兴趣有交叉的关键词存在，这与聚类算法发生冲突。

（4）基于本体（Ontology）的存储形式。

Ontology 具有的良好的概念层次结构和对逻辑推理的支持，因而在用户模型创建中得到了广泛的应用。基于 Ontology 的用户模型可以表示成具有层次的概念图的形式[40]，并且存储在一般的关系数据库中，采用图的匹配技术来完成信息检索[40]。但在基于概念的信息检索系统中[41,42]，由于要求有较强的推理能力，基于 Ontology 的用户模型一般要用一种描述语言（如 Loom，Ontolingua 等）来表示，用户模型保存在知识库中，通过描述语言的逻辑推理能力来完成信息检索。

基于本体的存储其优点是：能通过概念之间的关系来表达概念语义的能力，所以能够提高检索的查全率和查准率。但是，由于 Ontology 是面向特定领域的，描述的是特定领域的概念模型，所以，必须要有专家的参与，这是最主要不足之处。

此外，基于本体的层次树结构与基于分类的层次树结构是不同的，对于每个结点而言，分类层次树的每个结点都是一个聚类，并且，从树根向叶子，逐渐细化的一种分类形式；即越靠近叶子结点，结点规模越小。而本体层次树的结点只是一个关键词或元数据，从树根向叶子，结点之间只存在概念的细化问题，而不存在结点规模越来越小的问题。因此，二种存储形式有本质区别。

（5）基于语义网的存储形式。

基于语义网的用户模型存储，系统采用的是基于 XML 的 RDF（resource definition framework）作为用户模型的表现形式。典型的系统如 SiteIF[29]和 IFWeb[43]都采用语义网来表示用户模型的。在 SiteIF 项目中，用 RDF 记录了根据请求而建立的用户模型，在用户模型中的每一个结点，都代表一个概念，并且，概念与概念之间的弧，表示一种共现关系，判断文档与用户模型的相关性时，只能用语义网的评估办法来实现。

基于语义网的用户模型的存储，其优点是利用了 XML 的优点，便于阅读且容易理解，数据的含义也很容易表达清楚。同时，又从语义的角度体现了概念与概念之间的共现关系。但由于是基于 RDF 的三元组表示方式，根本无法体现概念结点间的复杂的语义关系。

（6）基于 *N-gram* 的存储形式。

N-gram 是对向量空间模型的一种改进，空间向量模型在构建用户模型时，没有考虑词序及词的上下文关系，所以，无法识别同一个词在不同词组中出现情况，导致向量空间模型的检索精确度下降。而 *N-gram* 正是为了解决这一问题而提出

的，它从根本上体现了一种语境上下文的关系，其基本思想是：如果 n 个词多次同时出现，则认为它们互为上下文关系，在用户模型中存储时，就将这些共现的有序的词，作为一个结点存储，第 n+1 个词与前 n-1 个词出现的概率，可以用概率公式 $p(w_i \mid w_1 w_2 ... w_{i-1})$ 来计算。同时，利用 *K*-lines 解决了词出现的序列和词组的长度问题。典型的系统如 PSUN[30]。

基于 *N-gram* 存储的方式其优势是在语境方面对 VSM 作了改进，使检索精度有所提高。但在选择 N 时，难度很大：如果 N 较大，则提供了更多的语境信息，语境更具区别性，但参数个数多，计算代价大，训练语料需要的多，参数估计不可靠；如果 N 较小时，语境信息少，不具区别性，但参数个数少，计算代价小且训练语料无需太多，参数估计可靠。

2.2.2 用户模型的学习与更新技术研究

用户模型学习和更新的最主要途径是相关性反馈，在反馈中进行查询扩展和检索词权重的调整。本文中，只介绍两种最为典型的用户模型更新技术。

1. 向量空间模型中的相关性反馈及用户模型的学习与更新

向量空间模型中，文档和查询均被表示成为具有相等长度的向量，$\vec{d} = \{d_1, d_2, ..., d_n\}$ $\vec{q} = \{q_1, q_2, ..., q_n\}$，其中，$d_i$，$q_i$ 分别表示相应关键词在文档和查询中的权重。文档与查询的相似度为 $sim(\vec{d}, \vec{q}) = \dfrac{\sum d_i \cdot q_i}{\|d_i\| \|q_i\|}$。

Rocchio 提出了经典的反馈查询：

$$\text{Rocchio：} \quad Q_1 = Q_0 + \beta \sum_{i=1}^{n_1} R_i - \gamma \sum_{i=1}^{n_2} S_i \tag{2.1}$$

其中，β，γ 是可以由人工进行调整的参数。Q_1 是反馈查询向量，Q_0 是原始查询的向量。Rocchio 算法有一个基本的假设：没有被用户判断为相关的文档，就是不相关文档，显然，这个假设存在问题。Ide 提出了对 Rocchio 算法的改进：

$$\text{dechi：} \quad Q_1 = Q_0 + \sum_{i=1}^{n_1} R_i - S \tag{2.2}$$

$$\text{Ide regular：} \quad Q_1 = Q_0 + \sum_{i=1}^{n_1} R_i - \sum_{i=1}^{n_2} S_i \tag{2.3}$$

在 VSM 中，查询扩展可以概括为 Rocchio Term Selection（RTS）方法：

$\delta(t)=idf(t)\times\frac{\sum_{d\in Feedbackdocs} tf_d(t)}{R}$，其中 $idf(t)=\log\frac{N+0.5}{n+1}$

当 $\delta(t)$ 大于一定阈值，t 可作为扩展查询词。关键词权重计算，在 VSM 中使用的是 TFIDF 方法。

VSM 模型中，用户模型的学习与更新过程如下：

（1）用户输入初始查询，系统返回初始文档集，用户判断相关文档与不相关文档。

（2）计算新的查询权值，同时计算文档与查询的相关度，文档按相关度倒序排列，将 top n 返回用户。

（3）根据 $\delta(t)$ 将新产生的关键词，加入到用户模型向量中，同时，将其他关键词的权值进行相应调整，根据一定的规则，调整用户模型中向量排列次序，如果超出最大上限，根据规则，删除权值较低的关键字列表。

（4）用户如不满意，重新查询，反馈过程迭代。

2. 概率模型中的相关反馈及用户模型的学习与更新

经典的概率模型如 Robertson S.E.和 Sparck Jones 提出的二元独立模型 BI 中，只考虑查询词出现在相关文档和不相关文档中的概率分布，而没有查询扩展的过程。通过用户的相关性反馈，只是重新计算查询词在相关文档中和不相关文档中的概率分布。在 BI 中，关键词属于相关文档的概率分布为 $P=\frac{r}{R}$，关键词不属于相关文档的概率分布为 $\overline{P}=\frac{n-r}{N-R}$，其中，$r$ 为关键词出现在相关文档中的次数，$n-r$ 为关键词出现在不相关文档中的次数，N 为文档集合文档总数，R 为相关文档总数，$N-R$ 为不相关文档总数；对一给定的查询词 t，其相关权重为 $w=\log\frac{p(1-\overline{p})}{\overline{p}(1-p)}$，对于给定文档 D，其相关度计算为 $g(D)=\log\frac{P(D\mid rel)}{P(D\mid non\ rel)}$，其中，$p(D\mid rel)$ 表示文档出现在相关文档集的概率分布，$p(D\mid non\ rel)$ 表示文档出现在不相关文档集的概率分布。

在概率模型中，查询扩展可采用 RSV[44]（Robertson Selection Value）方法，即 $\delta(t)=(p-\overline{p})\log\frac{p(1-\overline{p})}{\overline{p}(1-p)}$，当 $\delta(t)$ 大于一定阈值，t 可作为扩展查询词。

在概率模型中，用户模型的学习与更新过程如下：

（1）用户输入初始查询，系统返回初始文档集，用户判断相关文档与不相关文档。

（2）计算每篇文档的相关度系数 $g(D)$，将文档集按 $g(D)$ 降序排列，将 top n 文档返回给用户。

（3）根据 $\delta(t)$ 将新产生的关键词及其概率分布加入到用户模型中，同时，将其他关键词的权值进行相应调整，根据一定的规则，调整用户模型中概率分布排列次序，如果超出最大上限，根据规则，删除权值较低的概率分布列表。

（4）如果用户不满意，继续进行查询，反馈过程迭代。

在个性化搜索领域，除基于上述两种典型模型的相关性反馈，研究比较多的还有基于语言模型、布尔模型等的相关性反馈。另外，将基于多种模型的反馈算法结合起来改进用户模型的学习和更新的效率也是研究的热点。此外，用户模型学习和更新与其存储的位置也有关系。

2.2.3 用户模型应用技术的研究

用户模型体现了用户的兴趣偏好，其主要是应用在各种个性化服务过程中，就搜索引擎的个性化服务而言，用户模型的应用分为以下几个方面：

（1）利用用户模型实现信息过滤。

（2）利用用户模型实现对搜索引擎输出结果的重排序。

（3）利用用户模型进行信息导航。

（4）利用用户模型对查询进行修正。

通过对以上相关工作的深入浅出的对比分析，我们不难得出结论：对于用户模型的构建，潜在兴趣的挖掘和获取是当前发展的主流方向，在不侵犯用户隐私的情况下，不知不觉中精确的获取用户的真正搜索意图，对用户心理进行揣测和其认知结构的研究是未来个性化服务中用户模型研究的一个最迷人的方向。

2.3 用户搜索行为的理论分析

2.3.1 从认知角度分析用户的搜索行为

用户的行为分析是用户动机分析的关键，由于以往的工作只是停留在关键词和文档的潜在语义挖掘或分析上，这样做最突出的缺陷就是忽略了用户心理和认知上的变化，事实上，在不断搜索过程中，用户心理状态是不断推进的，由已知向未知再走向已知，同时，心理情绪上的变化也促使他对其搜索的现状进行改变。以往的研究并没有把用户的认知行为状态分析加入到个性化服务的用户模型构建

上。从这一点上说，本文是对传统用户建模的补充与扬弃。在认知基础上进行心理建模是本文的创新点，尤其是利用概率的方法对用户搜索的行为进行分析，对潜在用户动机进行模型的推演。

为了本文后面行文的方便，在本章，我们先将用户心理分析与认知推理的理论基础知识进行梳理。逐步清楚透彻的展示用户动机推演的全部过程。

1. 搜索行为是一种心理需求

（1）信息需求层次结构。

西格蒙·弗洛伊德是公认的 20 世纪最伟大的心理学家，他认为用户的信息搜索行为是指当信息用户有了确定的信息需求时，以各种方式对所需求的信息进行寻求、搜索和使用的行为[38-40]。他把人的整个机体看成是一个能量系统，这个系统遵守能量守恒定理，在心理过程中起主要作用的是本能，即有机体内部先天固有的一种驱动能量，它构成了人的全部精神能量，冲动是一切本能的共有特点，也是本能发挥动力作用的根源所在，在本能中构成冲动的组件是成双成对相互对抗、相互矛盾着的本能方面。用户的搜索行为显然是受本能支配的，本能是引起个体信息行为、维护该行为、并将该行为导向特定信息目标的内在驱动力。促使信息本能形成的原因主要有两个：一是内在条件需要；二是外部条件刺激。当外部条件不变时，内在本能的冲动是一个人产生信息行为的根本原因。所以，要彻底分析用户搜索的全过程，预测用户的未来搜索信息行为，首要完成的核心任务就是深入研究人的信息本能，即其信息搜索的动机。

举例来说，当用户在工作或生活中遇到了问题，需要获得信息来支持该问题的解决时，他具有信息需要。这是一种完全由客观条件决定，不以个人主观意识为转移的需要状态，即信息需要的客观状态。在这种状态下，人们可能并未认识到自己的信息需要。这表面上的原因看起来是现实问题过于复杂和隐蔽而个人的认识能力有限或信息意识淡漠，因此，他并不知道自己的信息需要是什么，这就是潜在的信息需求。人们一旦认识到了自己的信息需要，其信息需求层次也就上升了一级，离信息行为就更近了一步。图 2.1 给出了信息需求层次结构。

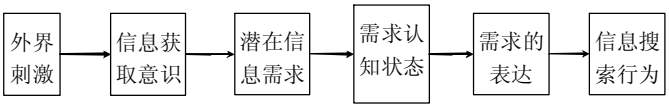


图 2.1　信息需求层次结构

潜在信息需求状态向不同状态转化的过程涉及用户不同的认知过程，紧接下

来，我们将认知心理学的内容进行梳理，对用户信息搜索动机分析机理从认知心理学的角度进行深入剖析，最终展示出对用户动机进行推演的全过程。

（2）个性特征分析是进行用户个性化服务的基础。

弗洛伊德认为人可分成本我、自我和超我三重结构。本我是精神结构中最古老的部分，由于它潜藏在无意识的状态之中而显得隐秘而无法捉摸。本我是人所有精神活动所需心理能的集散地。本我是由原始的本能、先天的欲望所组成。自我是人格结构中理智的、符合现实的部分，它是由本我派生出来的。人要适应环境和支配环境，并从环境中获得所需要的东西，这种人与环境的交互作用便形成了自我。超我是人格中最文明、最有道德的部分，它包括自我理想和良心两部分。进行用户心理研究所必须洞悉的一个基本事实，就是人类的信息获取具有“本我”所有的特征，不同用户的信息需求是具有人类信息需求的“自我”特征，不同的信息取舍具有“超我”特征，也是进行用户个性化服务的基础。

2. 信息搜索的认知过程与问题解决过程

（1）信息搜索的认知过程。

现代认知心理学家奈瑟认为：认知过程实际包括个体对外界刺激产生反应的过程和个体有意识地控制、转换和建构观念与映像的过程。按照认知加工心理学观点，人就是一个信息加工的系统，认知就是一个信息加工的过程，具体包括感觉输入的变换、加工、存储和使用的全过程，也即是信息的获得、加工、贮存和使用的过程，感知、注意、记忆、思维和言语等是其中的行为表现[45]。

感知包括感觉和知觉，它是信息加工处理中重要的认知行为。目前，认知心理学中认为感觉是对刺激的觉察，也是一个信息的获取过程，而知觉是将感觉信息组成有意义的对象，即在已贮存的知识经验的参与下，把握刺激的意义。因此，知觉信息是现实刺激的信息和记忆信息相互作用的结果，是指对感觉信息的组织和解释，也即获得感觉信息的意义的过程，也是对信息的加工处理获得再生信息的过程。

注意是心理活动对一定对象的指向和集中[42]，这主要发生在信息获取阶段，它可以实现对刺激选择的控制和行为调节。表象与知觉有紧密联系。认知心理学将表象看作是已经贮存的知觉象的再现（记忆表象），或经过加工改造而形成的新的形象（想象表象）[43]。按照信息加工观点，前者可以看作是人们对外部刺激感觉到的初级映像，后者可以看作再生信息映像。

记忆一般包括感觉记忆、短时记忆、长时记忆。当外部刺激直接作用于人的感觉器官，会产生感觉象然后实现瞬间贮存，形成感觉记忆；然后经扫描选择后

的感觉记忆信息进入了短时记忆；在复述的条件下短时记忆信息可以长期保持，并可进入长时记忆[43]。在认知过程中感觉记忆主要表现在信息获取阶段产生对外部刺激的映像并存储，在信息处理阶段，人们主要是将过去储存在大脑中的长时记忆信息与现实刺激信息进行比较与分析。

思维是人对客观事物本质特征和规律性联系的间接、概括的反映[41]。它是人类认识的理性阶段，能更深刻、正确、全面地反映客观事物。在认知心理学中，主要探讨的思维过程包括概念形成、问题解决和推理。在信息加工过程中思维活动主要表现在信息处理阶段，通过思维，人们要对外部刺激的本质特征进行认识，最后转换、建构为新的知识。

我们可以将认知活动概括为两个过程：感知思维过程，即信息获取和信息处理的过程，以及知识再生（或信息再生）过程，即作为前两个过程的结果—知识状态改变的过程或产生主观信息的过程。其中，知识状态改变的过程是认知概念的核心，也可以理解为狭义的认知概念，见图 2.2。

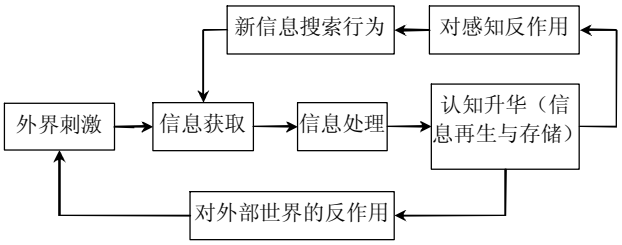


图 2.2　信息搜索的认知过程

（2）信息搜索的问题空间描述与解决理论。

问题解决过程是一个对问题空间（Problem Space）的搜索过程。问题空间是问题解决者对一个问题所能达到的全部认识状态，包含问题的初始状态，一系列中间状态，目标状态以及状态转换算子 Operator[43]。问题解决一般经过三个阶段：

（A）问题表征，即问题解决者要将任务领域转化为问题空间。问题表征依赖于人的知识经验，也受到注意、记忆和思维等心理过程的制约。在搜索引擎的个性化问题解决的过程中，问题表征就是搜索者脑中的概念进行表征。

（B）选择算子或选择操作步骤，也就是选择改变问题初始状态的系列操作步骤，使对问题的认识从初始状态，一步步沿着中间状态逐渐逼近目标状态。问题解决过程中，究竟选择哪些算子，将它们组成什么样的序列，都将依赖于个体采取哪种问题解决的方案或计划，即问题解决策略。在搜索过程中，用户选择什

么信息源、选择什么概念组配进行搜索都会在很大程度上决定趋向目标的路径长和短，对于有经验的用户，只要选择一个信息源，使用一个概念组配式就可能轻而易举地查到所要的信息；而对于无经验的用户，对于同一个信息源，可能花费很长时间也找不到目标信息。

（C）应用算子，即实际运用所选定的算子来改变问题的初始状态或目前的状态。如果达到理想效果算子应用不会改变，但事实情况是信息搜索过程中有很多不确定性，如果没有丰富的经验积累，常会出现错误决策，此时需要变换算子。

问题解决过程是一种高级智力活动过程，是搜索者在特定的问题情境中，认知问题情境，产生初始概念，并在此概念驱动下，随着情境因素和个体经验交互作用（即信息加工过程）增强，推动概念转换，由此逐步地构建出最佳问题表征，并在头脑中对此心理表征执行序列化的认知操作，即产生算子，其中的相关性反馈属于改进算子操作，最终达到目标状态的信息加工过程。然后进行个性化的服务，如个性化推荐等即应用算子的操作了。

问题解决理论是最终我们进行用户动机模型推演的根本出发点，通过问题解决的方式，找到问题空间，以概率的形式定义并建模用户认知问题情境过程中的认知状态，确定问题的初始状态和目标状态是用户动机建模过程中最为关键的。

2.3.2 用户搜索行为的不确定性

Reitman 把现实中的问题分为两大类[43]：一类是有一定规则的、简单而明确的问题，即算子（Operator）和问题空间界定明确的问题，即结构良好问题（Well-Structure Problem）；另一类是规则和条件不明确，算子不清楚、具有很大不确定性的问题，也称结构不良问题（Ill-Structure Problem），即不确定性问题[43]。结构良好的问题有两个明显特征：问题有唯一正确的解决途径；解决问题必须遵循特定的程序、步骤和方法。这种问题只要运用形式运算或形式逻辑即可解决。而结构不良问题经常是给定的目标或者条件没有被清楚地说明，一般需要在经验的作用下将感情和认知高度地整合起来，运用辨证运算或辩证逻辑加以解决[46]。

本文所研究的个性化搜索问题按照上述框架应该属于结构不良问题，即属于不确定性问题。最根本的一点是针对不同的用户，问题的空间是不确定的，不同用户的认知水平也是不确定的，问题的起始和最终状态也是不确定的。我们没有特定方法，一切取决于用户个人经验、偏好，同样的搜索关键词可能对应用户不同的个性化需求，并且，搜索引擎无法获知用户感觉上的信息，如我们常说的这

篇文档基本符合我们的要求，那篇文档有一部分内容有用，对于这样的问题，我们的处理办法只能是通篇全拿来进行分析，将用户忽略的部分也当成是用户的偏好信息，引入噪声的大小根本无法人为控制。这也是个性化搜索迄今为止，始终在研究却仍没有得到较高精度的主要原因。对于个性化研究的难度程度之大，微软公司的文继荣博士曾在 VLDB School2008 panel 上说过：“微软曾经依靠巨大的用户资料去为用户创建用户模型，但将用户模型应用到个性化搜索过程中时，反而降低了搜索的精度”，这也从实践的角度说明了用户模型的构造是个性化服务的基石和最大难点。国内外的研究人员在这一点上基本是达成共识的。日本的 A.I. Kovács 和上野晴树[47]也曾指出，在个性化搜索领域三个著名的不确定性问题：首先，知识获得难题（KAP: Knowledge Acquisition Problem）系统内的信息都按照一定的关联规则或分类原则构成一个信息关联网络，这种信息之间的关联属于系统的隐性知识；由于用户认知的个性化差异，系统很难让用户也能完全按照设计者的意图理解系统内的信息关联网络，这造成了知识获得难题：无论如何设计，只要系统的知识是预先设计的，就不可能同时满足具有不同知识背景的用户。其次，不可知性问题（UP: Unknowability Problem）系统无法预先知道什么样的分类法或术语在未来是最合适的（最适合用户，因为术语或分类与适合的用户群有关，也与时间有关），也不能知道它们是否会被应用，甚至也无法知道它们将来在某个个性化查询中的意义，这里有太多的可能背景，几乎不可能知道或预测在所有这些背景中应该如何处理。再次，明确表达的问题（EP: Explication Problem）无法在一个逻辑形式框架内应用形式语义表达一个模糊的知识，用户事先不知道自己想要的东西应该是什么样，无法用“和某某有点像”、“最好的”、“大一点”之类的含糊的语言来表达他想要的文档或其他东西。

根据以上的分析，用户的搜索动机分析属于不确定性问题的研究范畴，在对用户进行建模的整个过程中，需要将用户的感情和认知过程高度地整合起来，运用辨证运算或辩证逻辑进行解决。

2.3.3 用户搜索行为分析的逻辑框架

信息搜索本身就是一个学习与认知的过程，是一个从搜索体验到用户认知图式改变更新的过程。因此认知因素是我们用户建模作为重点探讨的内容。目前还没有关于用户模型的标准或规则出台，但却有一些成功经验可以吸收，如著名的威尔逊（Wilson）信息搜索行为模型[48,49]和库尔斯奥（Kuhlhtua）信息搜索行为模型[50,51]，以及德尔文 Sense-making 模型[53-55]和埃利斯（Ellis）的信息寻求行为模

型[56]等。受已有模型的启发，本文给出了用户搜索行为分析的逻辑框架，见图 2.3。

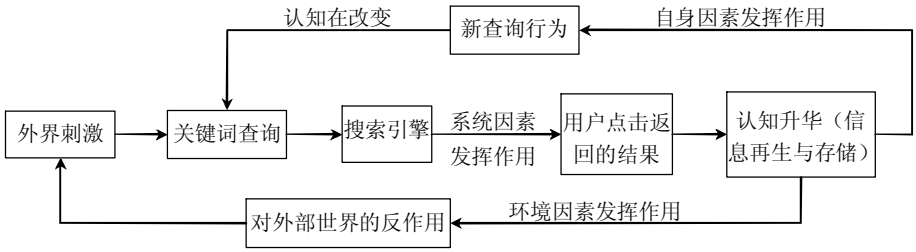


图 2.3　用户搜索行为分析的逻辑框架

由图 2.3 可以看出，信息搜索过程的实质就是用户认知图式与外部输入的交互作用的过程，即搜索是在处理用户主体因素与外部因素过程中完成的。外部输入不仅包括问题情境，更包括由信息搜索所获得的外部知识的输入，大多数情况下外部输入的知识将改变用户原有的认知结构，完成搜索任务中一个问题解决的循环周期。随着第二个问题出现，又会启动第二轮的信息搜索，如此反复循环，直到研究任务中目标预期全部实现。

其中自身因素包括用户的人口统计因素，认知变化因素等。外部因素主要是情境因素，所谓情境就是行为发生时的环境，即环境因素，如课题研究任务与搜索任务可以看作是组织环境中滋生出的重要内容，它对用户的信息搜索行为有着重要影响，也是用户信息活动的重要特征，因为激发用户信息搜索需要和行为的主要情境因素就是课题任务与搜索任务。此外还有其他人已有信息行为，比如其他人的信息搜索习惯、方法、经验，它们都会在不同层面影响信息用户的信息行为。这些都可以看成是信息用户在信息行为发生过程中的环境因素；同时，外部因素也包括系统因素是指用户所选择面对的是哪一个搜索引擎，比如 Google、Baidu、Sogou、Yahoo、Aol 等。

对用户搜索行为分析并对其兴趣偏好进行建模需要收集的因素太多太多，但只有部分因素是对用户个性化搜索起主导作用的。如模型中所涉及的人口学信息因素，包括性别、年龄、社会和经济地位、教育和工作背景等在个性化搜索中的作用其实并不大，还有一些环境变量如立法、经济情况、稳定程度、部门的组织结构这些因素对于个性化搜索可能会起到一定的作用，但作用很显然并不是很大。因此，对用户兴趣进行建模时尽量避轻就重，着重考虑影响个性化搜索的关键因素——自身因素，即搜索动机。

2.4 用户动机分析的两类不确定问题

用户搜索动机分析中经常遇到两类不确定难题：①**一词多义**（polysemy），指词汇间同义性和单个词汇的歧义性，这个问题在任何语言中都存在，是搜索引擎比较难于处理的问题。下面这个例子[57]很好地揭示了这个问题。假设我们要查找美洲虎（jaguar，一种南美的猫科动物）的奔跑速度。在 AltaVista 中的输入关键词 jaguar speed。结果是一些有关美洲虎车、Atari 视频游戏、一支美洲足球队、一个当地网络服务器等。第一个关于动物的页面排在第 183 号，是一则寓言，没有关于速度的信息。在第二次检索时，加入了 cat，结果是关于 Clans Nova Cat、Smoke Jaguar、LMG Enterprises、好车等。仅仅在第 25 位的网页有关美洲虎的信息，但并没有涉及到速度。事实上仅在英文中，多义词也不只 jaguar，如 java、apple、party、chair、ajax 等不胜枚举。对于这类词，如果搜索引擎不能根据用户所需要的主题给出结果列表，而只按照关键词匹配方法，那么返回的网页链接就会毫不相关，是远远不能满足用户的需求的。②**隐喻**是人类思维复杂化的表现，也称为暗喻。到目前为止，隐喻还没有确切的定义，通俗地理解的话，可以认为修辞学角度来解释，诸如“人类灵魂的工程师”“祖国母亲”这类句子就是隐喻，隐喻的研究涉及修辞学、语言学、哲学、心理认知学等各个领域。著名的学者莱可夫和约翰逊[58-61]合著了一本书《我们赖以为生的比喻》（Metaphor we live by）的书中认为，隐喻不仅仅是一种语言现象，更是一种认知现象，隐喻思维是人类认识事物、建立概念系统的根本，与人的思维和判断推理有关，恰当的隐喻有助于人们作出正确的决策，人们用隐喻来表达自我，指导未来行动。雷德曼[62]认为隐喻是思维的工具，它能传送新的认知内容，隐喻可以存在于任何研究领域中，使人们模糊的概念变得清晰。

这两类问题导致的后果有两个，第一，基于传统词汇匹配法的信息检索技术无法弥补这种词汇词义上的不一致所造成的不足，因为文本本身可能并不包含用户提出的查询关键词，或者在选择词时要解决词汇的同义性问题，需要依靠词汇的自动智能扩充以及词表的构建。但词表扩充后，新的问题产生了，某些新进入词表的词汇又具有歧义性。第二，词汇歧义问题仍靠人工转换来确定词义，但该方法不仅成本高，而且并不十分有效。

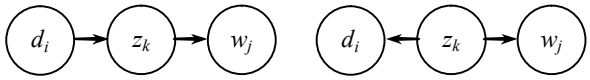
2.5 基于 PLSA 的潜在概念获取与用户模型构建

2.5.1 概率潜在语义分析

为了解决用户搜索中的两类不确定问题，1988 年，来自 Bell Communications Research、University of Chicago 和 University of Western Ontario 的 Susan T.Dumais、Thomas K.Landauer（现为 University of Colorado 教授）、Scott Deerwester 等五位学者共同提出了潜在语义分析（Latent Semantic Analysis，LSA）这一自然语言处理的方法[63-65]。潜在语义分析方法认为在特征词条之间存在潜在的语义关联，而这种语义关联仅仅通过特征词条的词频特性不能很好地描述[66]。潜在语义分析出发点就是文本中的词与词之间存在某种联系，即存在某种潜在的语义结构。这种潜在的语义结构隐藏在文本中词语的上下文使用模式中，其分布不是绝对随机的，而是服从某种语义结构；但 LSA 方法仍存在受限词的问题，于是，于 1999 年，Hofmann[32,67]提出概率潜在语义分析 PLSA（Probability Latent Semantic Analysis）方法，完全克服了 LSA 的不足。PLSA 的最主要特点就是利用潜在因素来进行文档分析，即 $p(d_i,w_j)=p(d_i)p(w_j|d_i)$，这个概率背后，隐藏着潜在的语义空间 $z_k \in \{z_1, z_2, \ldots, z_k\}$，其中：

$$p(w_j \mid d_i) = \sum_{k=1}^{K} p(w_j \mid z_k) p(z_k \mid d_i) \tag{2.4}$$

则 $p(d_i,w_j)=p(d_i)p(w_j|z_k)p(z_k|d_i)$，如图 2.4 所示。由于潜在变量的基数通常比集合中文本或词的数要小很多，所以，潜在变量 $z_k \in \{z_1, z_2, \ldots, z_k\}$将成为瓶颈变量，所以，根据贝叶斯公式[34]，将其箭头方向逆转，如图 2.4（b）所示。



（a） （b）

图 2.4 PLSA 算法图示

PLSA 算法中，潜在变量的估计使用的是期望最大算法（Expectation Maximum，EM）。在 E 步，具体的计算如式（2.5）。

$$p(z_k \mid d_i, w_j) = \frac{p(w_j \mid z_k)p(z_k \mid d_i)}{\sum_{k=1}^{K} p(w_j \mid z_k)p(z_k \mid d_i)} \tag{2.5}$$

M 步骤中，使完备数据的对数似然函数取最大值，其计算过程如式（2.6）和式（2.7）。

$$p(w_j \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)p(z_k \mid d_i, w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, w_m)p(z_k \mid d_i, w_m)} \tag{2.6}$$

$$p(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)p(z_k \mid d_i, w_j)}{n(d_i)} \tag{2.7}$$

E 步与 *M* 步的迭代，最终满足收敛条件时停止。

下面，用矩阵的形式来说明概率潜在语义分析与潜在语义分析的不同之处：$T_0 = (P(d_i \mid z_k))_{i,k}$， $D_0 = (P(w_j \mid z_k))_{j,k}$ $S_0 = diag(P(z_k))_k$， $P = T_0 S_0 D_0^T$

下面解释一下这样分解的意义所在：在 LSA 的潜在空间中，没有空间的方向的解释，而在 PLSA 的潜在空间中，方向很确定的条件概率分布，该分布定义了一个特定的主题的语境。

2.5.2 潜在语义空间的 Zipf 分布

1. Zipf 分布

Zipf 分布是美国哈佛大学语言学教授 George Kingsley Zipf[23]在研究英文单词出现的频率时，发现如果将单词的频率按高到低的次序排序，每个单词出现的频率与它的序号存在简单的反比关系：$F(r)=C/r^{\alpha}$，其中 r 表示单词序号，按照出现频率由大到小排序，C、α为常数，$F(r)$为序号为 r 的单词出现的次数。Zipf 分布表明，少数高频单词出现的次数远远高于占多数的低频单词的出现次数，即在英语单词中只有很少数被经常使用，而绝大多数词很少被用到。

Zipf 分布一经提出，便得到广泛应用，并迅速走出语言学范畴，进入了信息学、计算机科学、经济学、社会学、生物学、地理学、物理学等众多研究领域，比如人们的收入、互联网的网站数量和访问比例、互联网内容和访问比例等[24]，但是，在潜语义空间中如何应用 Zipf 分布选择核心语义，目前还尚未见报道。本

文中，我们应用 Zipf 分布原理，去掉文档冗余部分语义，找到文档的核心语义。

2. 潜语义空间的 Zipf 分布的应用

基于潜在语义分析的一个重要任务是：对于文档 d_i 和关键词 w_j，在 $\{p(z_1|d_i,w_j), p(z_2|d_i,w_j), \ldots, p(z_n|d_i,w_j)\}$ 语义空间中，选择 $z_1, z_2, \ldots, z_k$ 并计算 $p(z_k|d_i,w_j)$ (k=1, 2, …, K)。由于文档的潜在语义与文档显性信息之间的关系是服从 Zipf 分布的，则 k 太大时，不同文档之间的一些重要的语义就会被微小的语义所屏蔽；k 太小时，不同文档之间只能存在一些共性的语义，难以区分不同文档的语义。

在本文中，我们通过实验选择不同的潜在语义维度 k 来分析文档，当 k 小于某一个阈值时，精确度是上升的，而当 k 超过一定的阈值后，精确度则不停地、非线性下降，所以，这个最高的阈值将是我们选择潜在语义空间维的参考因素之一。同时，结合[36]给出的双重概率模型，可以确定初始潜在语义空间的维度，这样，将两者结合起来，来确定潜在语义空间的维数。所以，在确定潜在语义空间的维以后，我们可以认为潜在语义空间中所保留的部分语义可以体现文档的真正语义所在。

2.5.3 基于 PLSA 的用户动机建模

1. 用户兴趣潜在主题获取

在确定潜在语义空间的维以后，我们可以认为潜在语义空间中所保留的部分语义可以体现文档的真正语义所在。同时，也应该注意到，语义空间中不同的维也代表了不同的兴趣主题。主题不仅能够体现用户的兴趣，具体的说，当用户发起一个查询时，他已经形成一个明确的目标主题，可是这个主题不能被查询词完全地表达出来，即用户处于认知的“非常态”。但有一点是肯定的，就是当搜索引擎返回大量繁杂的结果时，用户精心分选的网页，便蕴含着用户兴趣的潜在主题。

聚类技术是获得潜在主题的最佳手段，但传统的聚类方法难以处理主题交叉问题，针对此问题，本文提出了基于概率潜在语义分析的聚类技术，其核心思想是从分析用户所感兴趣的文档中的共现词的分布入手，找到文档的潜在语义空间，对潜在语义空间中的潜语义进行聚类，来实现用户兴趣的聚类，最终生成可以代表用户兴趣层次结构的树。具体内容如下：

首先利用文档中词 w 的概率分布特性 $p(w)$，根据 EM 算法确定不同潜在因素 z_i 下词的分布特性 $p(w|z_i)$，找到潜在因素的概率分布 $p(z_i)$，计算文档 d 包含此潜在因素 z_i 的概率 $p(z_i|d)$，根据此概率值，某一值域范围内的 $p(z_i|d)$被视为具有相同潜在因素，将具有相同潜在因素的文档进行聚类。基于 PLSA 方法进行文档的聚类，

同一文档可能包含多个潜在因素（Latent Factors），所以，同一个文档有可能被划到几个不同的聚类，即用户不同兴趣之间允许交叉，这是传统的聚类方法无法实现的。并且，在潜语义空间中，较大的概率值 $p(w|z_i)$所包含的语义比较丰富，具有更多的共性，较小的概率值 $p(w|z_i)$所包含的语义较具体，具有更多的个性。具体的算法见算法 2.1。

算法 2.1 用户偏好的潜在主题获取

输入：① 从用户感兴趣的网页中提取关键词、短语等的列表，以及根据关键词的出现次数为每个关键词赋予的加权值；② 同一潜在因素下，用 EM 算法得到的关键词的分布概率 $p(w|z_i)$；③ 文档中潜在因素的概率分布 $p(z_i|d)$；④ 初始阈值μ。

输出：① 以潜在因素聚类的用户兴趣层次树；② 产生 top K 个潜在因素下的网页推荐。

描述：

1. 将给定的关键词列表中 $p(w|z_i)$按从大到小降序排列，取出 $p(w|z_i)>\mu$ 的概率分布，并将这些取出的概率分布作为用户兴趣层次树的第一层结点；

2. 将新生成的第一层的每个结点视为初始结点，以递归的方式计算 $p(w|z_i)$，然后，同样取出 $p(w|z_i)>\mu$ 的概率分布，并将这些取出的概率分布作为用户兴趣层次树的第 n+1 层结点；

3. 直到对所结点计算的概率值 $p(w|z_i)\leq\mu$ 时，递归过程终止；

4. 通过人为参与，判断所得到的潜在语义是否理想，如不理想，调整μ，重新执行算法步骤 1～3；如潜在语义较为理想，则转为 5；

5. 针对最终生成的叶子结点，计算概率分布 $p(z_i|d,w)$，统计不同潜在因素下文档内出现的所有的共现词的权值之和作为相应文档的权值；

6. 计算不同潜在因素下所有文档的权值之和作为相应叶子结点的权值，将叶子结点按权值由大到小建一个链表；

7. 生成聚类层次树，算法结束。

算法中，步骤 1～3 操作的理论依据是 Zipf 分布原理，其根本意义在于找到文档的核心语义，并在核心语义基础上实现兴趣的聚类。

2. 用户兴趣模型生成

用户的搜索过程可以看作是认知的过程，而人的认知活动同时可以看成一个信息传递系统，把人们对外界的知觉、记忆、思维等一系列认知过程看成信息的传播接受和加工的过程，并对人的思维活动作出定量的分析，而在建立用户模型的过程中，即只须将用户所用的关键词的概率、潜动机概率、文档出现的概率分别用向量的方式表达出来，并将用户潜在搜索动机表示成一个层次树的形式。具

体操作如下：

（1）潜在用户兴趣主题的获取，请详见“用户兴趣潜在主题获取”小节。

（2）利用 ODP 的层次化“主题”来描述用户的兴趣。

ODP[①] 即开放式分类目录搜索系统，是目前网上最大的人工编制的分类检索系统。它创建于 1998 年 6 月，目的是为了解决最广泛地收集、最便捷地检索、最普遍地利用资源的理念与少数参与者无法处理急剧膨胀的网络信息之间的矛盾，使之成为一个完全开放的、由网民共建的、网络共享的网络分类目录。ODP 以层次结构组织的，根节点下面有 15 个一级主题节点，如图 2.5 所示，每个一级主题节点下有若干个二级主题节点。本文中将其看成一个可免费可得的层次结构清楚的本体概念集合。利用 ODP 的层次化“主题”来描述用户的兴趣，因为主题能够从释义上和概念上把兴趣更加准确地定位，能够为用户的兴趣模型建模提供一个规范的分类模型。

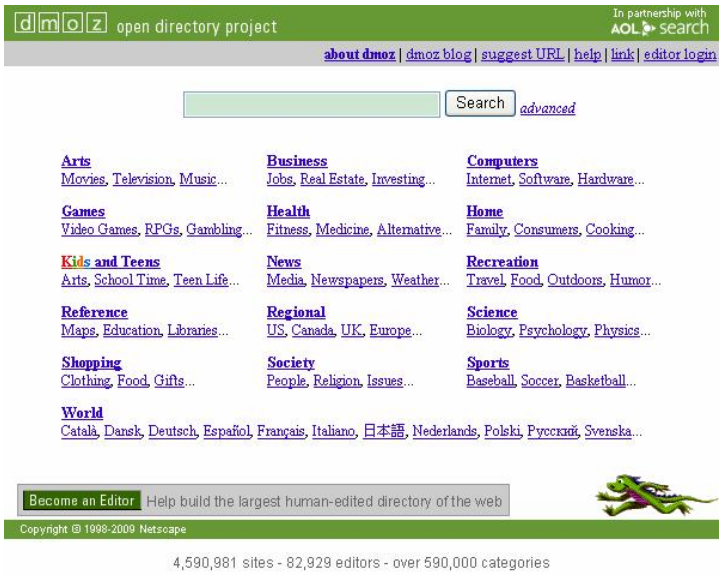


图 2.5　ODP 一级目录结构

本文主要是利用 ODP 直观地建立用户模型的层次主题结构，利用知识本体较为紧密的语义关联产生延伸概念，更准确的分析用户对每个网页的偏好，如果将

① Open Directory Project：http://www.dmoz.org

ODP 看成是一个大的本体库的话，用户模型可以看成是一个小的面向不同用户个性化本体库，利用这个小的本体库可以将与用户无关的的概念及网页排除在外而达到简化个性化分析过程的目的。本文这么做也是受到 Pitkow 等人提出的 Outride System 的启发，他们分析用户的浏览记录同时结合开放目录的 ODP 系统建立每一个用户的用户模型，当用户查询时，系统会自动计算该查询与用户模型之间的相关程度，对用户的搜索进行推荐。

（3）用户兴趣模型生成。

为了阐述方便，下面给出相关定义。

定义 2.1 兴趣向量 $T=\{t_1, t_2, \ldots, t_m\}$。其中 t_i（i=1, 2, …, m）为不同的主题。

定义 2.2 兴趣偏好 $PT=\{P(t_1), P(t_2), \ldots, P(t_m)\}$，其中 $P(t_i)$ 表示用户在主题 t_i 上的偏好程度，并且 $\boldsymbol{PT}$ 满足归一化条件：$\sum_{i=1}^{m} P(t_i)=1$。

定义 2.3 若用户的查询集合为 $\boldsymbol{Q}=\{q_1, q_2, \ldots, q_l\}$，点击网页集合为 $\boldsymbol{P}=\{p_1, p_2, \ldots, p_n\}$，则①$P(p_j)$为用户访问网页 p_j 的先验概率；②$P(z_i|p_j)$为当用户访问 p_j 时有意图 z_i 的条件概率；③$P(q_k|z_i)$为当用户有意图 z_i 时，会发起查询 q_k 的条件概率。

用户查询是一个认知变化的过程，为了将这个过程完全概率化，必须了解其查询发起时的潜在信息需求是什么，本文用潜变量 Z 来表示，则 $P(q_k|z_i)$则表示在此需求之下用户给出的查询，并且，在“用户潜在主题获取”小节中潜变量 Z 的概率分布已经确定，现只需给兴趣偏好向量 $\boldsymbol{PT}$ 赋值，并找出 $\boldsymbol{T}$ 和 $\boldsymbol{Z}$ 之间的关系。具体实现算法见算法 2.2。

该算法基于以下两个假设：

假设 1：如果 $P(q_k|z_i)=\mathrm{Max}\{P(q_k|z_{i'})|\, z_{i'}\in Z\}$，那么 q_k 更能代表 z_i；

假设 2：如果 q_k 在目录 c 中出现次数越多，那么 q_k 与目录 c 的关系越密切。

算法 2.2 用户兴趣模型生成

输入: $P(q_k|z_i)$, $P(z_i)$ 以及查询 q_k 在 ODP 的目录 c 中出现的次数。

输出: $\boldsymbol{T}=\{t_1, t_2, \ldots, t_m\}$; $\boldsymbol{PT}=[P(t_1), P(t_2), \ldots, P(t_m)]$。

方法:

1) $\boldsymbol{QS}=\{QS_1, QS_2, \ldots, QS_m\}$，$QS_1=QS_2=, \ldots, =QS_m=\Phi$;
2) for each q_k and each z_i
3) $q_k\in QS_i$ where $P(q_k|z_i)\geq P(q_k|z_{i'})$ $i'=1, 2, \ldots, n$ but $i'\neq I$;
4) end for
5) for i=1 to m
6) for each q_k in QS_i and each category c in ODP
7) $freq((q_k, c))$= matchings(q_k, c)/ matchings (q_k) ;

---

8) $freq(QS_i,c)$+=$freq\ (q_k, c)$;
9) end for
10) end for
11) for i=1 to m
12) t_i=c where $freq\ (QS_i,c)$>$freq\ (QS_i,c')$;
13) $P\ (t_i)$= $P(z_i)$
14) end for

算法 2.2 中只考虑到查询的主题问题，但用户的认知过程是渐进的过程，在这个渐进的过程中，用户的访问和发起会话是阶段性的，不可能是持续的，因此，下一节中将讨论用户模型的学习与更新问题。

2.5.4 用户模型的学习与更新

上面的部分通过潜在语义分析方法找到了用户的潜在动机，并以概率的形式进行了模型的构建，那么，又如何在用户模型中体现用户兴趣是随着时间而变化的，随着搜索过程中，认知状态的变化，人的搜索兴趣也会随之迁移，为此，用户模型的学习与更新同用户模型的构造同样重要。

人的认知过程是对输入信息的编码、贮存和提取的过程。在这一过程中，人的记忆分为短时的记忆和长时的记忆两种。德国心理学家艾宾浩斯[68]（Hermann Ebbinghaus）指出输入的信息在经过人的注意过程的学习后，便成为了人的短时的记忆，但是如果不经过及时的复习，记住过的东西就会遗忘；而经过了及时的复习，短时的记忆就会成为人的长时的记忆，在大脑中保持很长的时间。这一点对于用户模型的学习与更新是相当重要的，因为用户兴趣除主观点击之外，还存在客观的遗忘因素在内必须考虑进去。在这一点上，与著名的威尔逊（Wilson）信息搜索行为模型[48,49]和库尔斯奥（Kuhlhtua）信息搜索行为模型[50,51]，以及德尔文 Sense-making 模型[53-55]和埃利斯（Ellis）的信息寻求行为模型[56]侧重点均不同，这也是本文的创新点所在。

本文中采用依据用户查询进行增量形式的用户模型更新算法，即首先将用户提出的查询进行类别鉴定，然后，在用户模型中找到长期没有关注的关键词进行删除操作。同时，对于用户关心的主题也通过对遗忘指数的计算，实现对其权重的加或减。其中，遗忘指数采用斐波那契函数实现，这一点虽缺乏论证，但由于指数函数 2^{-x} 与艾宾浩斯遗忘曲线几乎完全符合。所以，给出斐波那契函数：

$$Fibonacci[i]=\begin{cases}0 & i=0\\ 1 & i=1,2\\ Fib[i-1]+Fib[i-2] & i\geq 2\end{cases} \tag{2.8}$$

同时，设 $d=Fibonacci[i]$，其中，i 为未被访问天数，如某主题 5 天未被访问则遗忘指数为 $d=Fib[5]$。用户模型更新时，用户兴趣衰减函数设定为：

$$\boldsymbol{PT}=\sum_{d=0}^{h}2^{-(d+1)}\times \boldsymbol{PT}_d \tag{2.9}$$

具体用户兴趣模型更新见算法 2.3。

算法 2.3 用户兴趣模型更新

输入：

1）待更新的用户模型。

2）查询 Q，即一系列关键词；预先定义好的阈值 θ。

输出：更新过的用户模型。

方法：

1）对于每个查询 Q，包含一系列关键词 $\{q_1, q_2, \ldots, q_n\}$，进行预处理。

2）利用 PLSA 判断用户查询所属主题：

E-Step：

$$\begin{aligned}P(topic_k\mid Q)&=\sum_{q_i}P(q_i)P(topic_k\mid Q,q_i)=\frac{\sum_{q_i}f(q_i,Q)P(topic_k\mid q_i,Q)}{|Q|}\\&=\sum_{q_i}f(q_i,Q)\frac{P(topic_k)[P(Q\mid topic_k)P(q_i\mid topic_k)]^{\beta}}{|Q|\sum_{topic_{k'}}P(topic_{k'})[P(Q\mid topic_{k'})P(q_i\mid topic_{k'})]^{\beta}}\end{aligned} \tag{2.10}$$

其中，$P(q_i|topic_k)$表示用户模型中特定主题 $topic_k$ 下，产生查询 q_i 的概率。

M-Step：

$$P(Q\mid topic_k)=\frac{\sum_{q_i}f(Q,q_i)[P(topic_k\mid Q,q_i)]^{\beta}}{\sum_{Q',q_i}f(Q',q_{i'})[P(topic_k\mid Q',q_{i'})]^{\beta}} \tag{2.11}$$

$$P(q_i\mid topic_k)=\frac{\sum_{Q}f(Q,q_i)[P(topic_k\mid Q,q_i)]^{\beta}}{\sum_{Q,q_{i'}}f(Q,q_{i'})[P(topic_k\mid Q,q_{i'})]^{\beta}} \tag{2.12}$$

E-step 和 M-step 迭代，最终得到 $P(topic_k|Q)$。

3）根据概率 $P(topic_k|Q)$将查询 Q 进行分类。

如果 $P(topic_k|Q)>\theta$，表示用户正在查询此主题，则用户模型中

$$weight\ \ topic_k=weight\ \ topic_k+k \quad (k \text{ 为常数}) \tag{2.13}$$

如果 $P(topic_k|Q)\leqslant\theta$，表示用户此时没有关注此主题，则将其权重乘以遗忘系数：

$$Weight\ Topic_k = \sum_{d=0}^{h} 2^{-(d+1)} \times Weight\ Topic_k \tag{2.14}$$

4）当模型中某主题权值小于阈值 δ 时，将用户模型中此主题删除。（δ 为人工设定）

5）返回更新过的用户模型。

2.6 实验及评价

由于用户模型的评价手段和评价方式到目前为止，还没有统一的标准。本文为了评估基于潜在语义分析方法所建立的用户模型的有效性，采用两种评估手段：第一，评价该用户模型在个性化搜索过程中的有效性，看看通过潜在语义分析方法构建的用户模型是否可以改善搜索的精度；第二，根据用户模型来分析和预测用户将要进行的查询，将用户真正提出的查询与预测的查询相比较，看看预测精确度为多少，是否与不用用户模型相比会有所改善，因为没有以往的查询预测实验作为对比，所以，第二个评估手段只是凭借预测的精确度，人为来评价其是否有效。

2.6.1 数据集

（1）我们选用的数据集是在 Depaul 大学的学术网站上下载的[①]，它就是著名的个性化搜索专用数据集“CTI 数据”。这个数据集是 Depaul 大学计算机科学与通讯、信息系的学生访问 CTI 服务器的随机抽象。本数据集中包含 2002 年 4 月中的两周的 web 日志文件，没经处理的原始数据包含 20950 个会话，5446 个用户。由于在数据集中存在过于简单的会话，这样的会话我们认为是噪声，将其去掉。最后得到处理过的会话及用户如表 2.1 所示。

为了实验中数据更充分，在这处理过的数据中我们抽取 20 个会话数据最为丰富的用户作为实验对象。抽取结果见表 2.2。

① http://www.cs.depaul.edu

表 2.1 正确会话的统计

userID	2885	4219	2126	3872	1367	795	4322	2937	4370	1095
Sessions	58	46	45	41	38	37	34	32	31	31
UserID	2127	2630	4202	888	3919	4714	3129	4211	838	4280
Sessions	28	28	28	27	26	25	25	25	25	23
userID	3778	4220	4230	3864	4471	2697	1948	3529	1242	1916
Sessions	22	22	22	22	21	21	21	20	20	20
userID	4199	4693	4892	4458	404	3899	1735	5299	4944	877
Sessions	20	20	19	19	19	19	19	18	18	18
userID	921	3186	3566	4176	4954	4235	3132	3371	3600	1210
Sessions	18	18	18	18	17	17	17	17	17	17
userID	885	202	3126	3593	4185	2683	2635	3723	3320	1217
Sessions	17	16	16	16	16	16	15	15	15	15
userID	3903	3331	166	1577	3218	1877	3249	1940	824	2642
Sessions	15	15	15	15	15	15	15	15	15	14
userID	3798	2940	3907	3910	3915	4019	3159	3406	1569	1663
Sessions	14	14	14	14	14	14	14	14	14	14
userID	2130	4415	4447	4470	3727	3140	3941	1124	4204	3956
Sessions	14	14	14	14	14	13	13	13	13	13
userID	2492	4236	3904	413	4400	4032	4111	3237	3141	1945
Sessions	13	13	13	13	13	13	13	13	13	13

表 2.2 第一个数据集统计结果

userID	2885	1095	2736	4400	4407	877	1874	624	4947	4176
Sessions	220	190	136	130	124	119	109	106	105	103
UserID	4370	4280	1039	3600	3864	3454	4437	2659	4414	3954
Sessions	101	98	92	91	89	89	82	80	77	77

在这个数据集中，为每个抽取的用户建立兴趣模型，并预测用户兴趣点所在，来检验用户模型的有效性。

（2）第二个数据集是从 KDD Cup 网站上下载的，KDD Cup 是由 SIGKDD

（ACM Special Interest Group on Knowledge Discovery and Data Mining）组织，每年一次的 KDD 竞赛，和 SIGKDD 国际会议同期举行。同时面向学术界和业界。我们下载的是 KDD Cup2005 年的数据集，这个数据集在 KDD Cup 网站上是任意下载的[①]。2005 年的数据集提供了 800,000 个查询和 67 个预定义的查询主题。由于数据集对于我们的实验来说，过于庞大，我们仅抽取其中的 111 个查询词，这些查询词在样本中已经分好了类别。为了方便起见，最终选中的这 111 个查询词被称为实验中的 KDDCUP data。

2.6.2 评价标准

（1）用户模型创建的评价。

我们设计了两个实验，第一个实验要考察潜在概念获取时聚类算法的效果，评价办法是将聚类结果与人工分类的结果对比，与人工分类结果相同的那些网页被视为“被正确获取主题的网页”，采用式（2.15）计算主题获取精确度 *TGP*（Topic Getting Precision）.

$$TGP = \frac{\text{被正确聚类的网页数}}{\text{参与聚类的网页总数}} \tag{2.15}$$

第二个实验根据建立的用户模型向用户推荐网页。如果用户下一步点击的网页是我们预测的网页，则称之为预测成功一次。以式（2.16）作为评估标准。

$$PRP = \frac{\text{成功预测的网页数}}{\text{预测出的网页总数}} \tag{2.16}$$

（2）用户模型更新的评价。

在本文中，用户模型的更新采用查询扩展的方式进行模型的更新，即对用户提出的新的查询首先进行分类，然后，将用户模型中相应的主题的权值进行调整或者将过于陈旧的主题予以删除，以及新的主题加入模型中来实现用户模型的更新，并保持与用户兴趣的一致性。所以，在评测用户模型更新的性能主要评测点在用户查询关键词的分类精确度如何。由于 KDD Cup 2005 数据集过于巨大，所以实验中只选择了 KDD Cup 2005 实验数据集中给出的 111 个样本查询关键词作为实验对象，这些样本是已经由专家标注好类别的，因此，我们实验中分类结束后，可以与专家标注的类别进行相应对比，来评价分类后的精确度。在评价过程中，使用的评价标准是 *F*1 测度。

---

① http://www.acm.org/sigs/sigkdd/kddcup/index.php

$$Precision = \frac{\sum_i \#\ of\ \ queries\ are\ \ correctly\ tagged\ as\ c_i}{\sum_i \#\ \ of\ \ queries\ \ are\ tagged\ as\ c_i} \tag{2.17}$$

$$Recall = \frac{\sum_i \#\ \ of\ \ \ queries\ \ \ are\ correctly\ \ \ tagged\ \ \ as\ \ \ c_i}{\sum_i \#\ \ of\ \ \ queries\ \ \ whose\ category\ is\ labeled\ by\ experts\ \ \ as\ \ \ c_i} \tag{2.18}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.19}$$

2.6.3 实验结果及分析

实验 1：实验过程中，由于每一个用户都需要建立一个用户兴趣描述文件，所以，*TGP* 和 *PRP* 对每个用户都是不同的，这里仅以 UserID 为 2885 的用户为例，将其结果展开分析。将用户全部会话中的网页提取关键词，共提取出 18 个关键词，按 PLSA 算法，算出关键词分布的概率 $p(w|z_k)$，然后，将大于阈值μ的关键词分布，取出形成第一层结点，依次以第一层结点为初始结点，算法递归执行，直到满足结束条件。对于 UserID 为 2885 的用户，其最后生成的兴趣层次树如图 2.6 所示。图中生成第一层结点的概率分布如表 2.3 所示。

表 2.3 UserID 为 2885 用户的兴趣树第一层结点的概率值

P(schedule|course)=0.0092	*P*(default|news)=0.2748
P(syllabilist|course)=0.1339	*P*(login|authenticated)=0.1067
P(syllabus|course)=0.0624	*P*(facultyinfo|people)=0.0693
P(searchcourses|course)=0.0092	*P*(evalgrid|people)=0.0378
P(studentprofile|CTI)=0.089	*P*(gradassist|CTI)=0.002
P(darsinput|CTI)=0.0711	*P*(core|CTI)=0.0178

从所生成用户兴趣层次树可见，/people/ search.asp 及/program/course.asp 被错误聚类，按照人为判断，总参与聚类的网页数为 450 个，被正确聚类的网页数为 412 个，准确率 *TGP*=412/450=0.915，对于其他用户的聚类算法评估与此相同，不再一一陈述。若想得到总体聚类的精确度，可以对选择的 20 个用户分别求聚类精确度，然后，求出其平均值即为算法总体聚类精确切度。本文按照所取出的 20 个实验用户计算，最终求得的总体 *TGP* 为 0.7776。

实验 2：为了评价这个兴趣描述是否符合用户的真实兴趣，我们按照算法最后生成的链表，将权值最大的前 2 项潜因素内的网页推荐给用户。仍以 UserID 为 2885 的用户为例，将其 58 个会话平均分为两组，一组作为训练生成用户兴趣描述文件中的层次树，另一组用于测试。分组生成的用户兴趣层次树与图 2.6 类似，只是网页个数不同，权值最大的 2 个的潜因素仍分别为 News 和 Courses，将这两个潜因素内的 6 个网页推荐给用户，看预测成功否。结果，在测试集中用户点击网页总数为 238 个，潜因素 News 下的网页为 84 个，潜因素 Course 下的网页为 99 个，计算点击率 *PRP*=(84+99)/238=0.7689。也就是说，这个基于潜语义空间的用户兴趣描述符合用户 UserID=2885 的真实兴趣。为了证实对于其他用户也是适用的，我们对取出的 20 个用户分别做了实验，得出他们各自的 *PRP*，最后，求出平均值，即总 *PRP*=0.7001。实验证明，潜语义空间的兴趣描述与用户真实兴趣相符合。

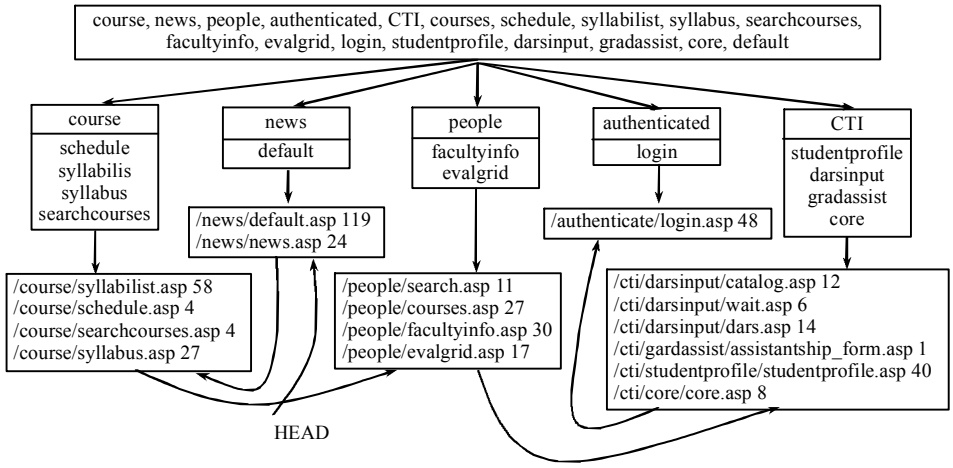


图 2.6　UserID 为 2885 用户的兴趣层次树

实验 3：我们将 KDD Cup 2005[①] 数据集中查询词根据预先给定的一些类别进行分类。由于 KDD Cup 2005 数据集过大，在此数据集中包含三个部分：Categories.txt, CategorizedQuerySample.txt 和 Queries.txt，在实验中我们只选取了 CategorizedQuerySample.txt 文件中给出的那些查询词作为实验对象。这里包含的

① http://www.acm.org/sigs/ sigkdd/kddcup/index.php

111 个查询词已经被专家标注好了类别，这也为实验证明我们所提出算法的准确度提供了依据。在进行分类之前，首先看一下在 KDD Cup 2005 数据集中的目标分类的层次结构（即 Categories.txt），从图 2.7 可以看到分类的结果应该是一棵分类层次树。

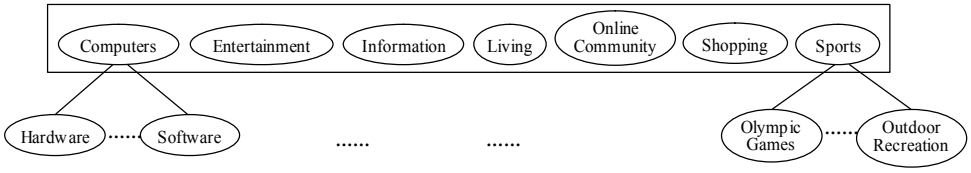


图 2.7 目标分类的层次结构

分类这些没有上下文的查询词，在分类过程中需要面对如下个问题：第一，很多查询词都过短并且没有任何意义，比如单一的数字，像 1939，10k，用户在搜索中用了这样的查询词是很让人头疼的事情。第二，查询词的一词多义问题，比如 apple 可以表示一种水果，也可以表示一种品牌计算机，或表示计算机公司等，java 可以表示一种咖啡，也可以表示一种计算机编程语言。在此种情况下，多义词本身就应该分别被分类到其所具有的意义的几个类别中，但具体分类到哪个类别中，还需要具体情况具体判断。比如：apple 就得被分类到\Computer\Hardware 这个类别和 Living\Food & Cooking 这个类别中。第三，有些查询词必须到确切的语境中才能准确判断其意义，比如 pod cast 指 Web2.0 中的播客，但在这个词刚产生的时候，在现实生活中如果用户在搜索引擎中输入这个词，是令人无法理解的。

为了解决这些困难，我们的解决方案是根据的搜索行为进行判断，判断用户的潜在搜索动机，根据用户的潜在兴趣来准确判断用户输入的搜索词具体应该属于哪种意思。同时，在这一过程中，结合 ODP 分类目录，可以更准确的生成用户兴趣的层次结构。在此，仅列出一些查询词分类后的例子，见表 2.4。

同时，我们也做了对比实验，Baseline 方法选择了 SVM 分类器，SVM 实现代码从 Http://svmlight.joachims.org/ 网站上下载。实验对比结果见图 2.8。从实验对比结果可以看出，基于 PLSA 方法对单一的、无上下文的查询词的主题类别获取仅次于 SVM，甚至某些查询词的主题类别获取还较 SVM 精准，这是因为，在用户搜索过程中，我们使用了基于概率潜在语义分析的用户搜索动机分析，所以，当新的查询词出现时，计算相应的主题与查询词的匹配度大小，就可以判断查询词的主题类别。同时，在图 2.8 中还可以看出，基于 SVM 方法的查询词分类明显

不如对某一文本集的分类精度高，因为，查询词是单一的，仅计算查询向量与主题向量的相关度来判断查询词所隶属的主题，仍存在一定的问题。

表 2.4　查询样本的分类

Query	Categories
baby stores	shopping\Stores & Products Living\Family & Kids
cross pendant	Living\Gifts & Collectables Living\Fashion & Apparel Living\Religion & Belief Shopping\Stores & Products Shopping\Buying Guides & Researching
eBay Electronics	Shopping\Auctions & Bids Computers\Multimedia Shopping\Buying Guides & Researching

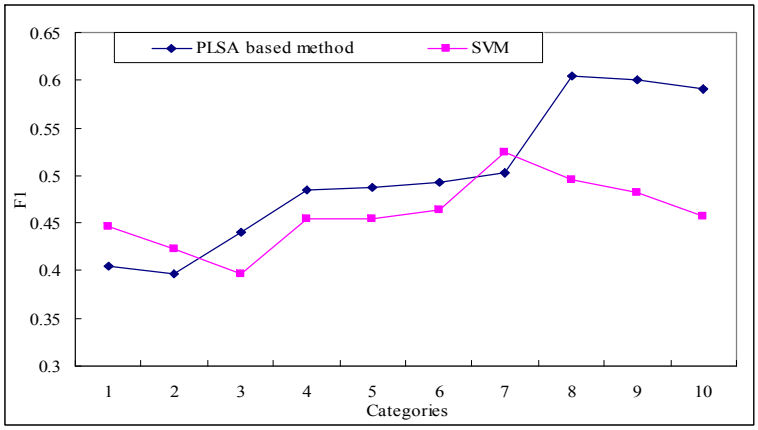


图 2.8　KDD Cup 样本数据集上的 F1

2.7　本章小结

本章对用户搜索行为从认知心理学的角度进行了系统的分析，认为用户的信

息搜索过程可以看成是一个用户思维活动的过程，它是对搜索过程中一系列问题解决的过程。但这一认知过程中，存在太多不确定因素，如系统因素，环境因素，用户自身因素，其中用户自身因素在用户搜索过程中起主要作用，因此，将此认知过程模型化的同时，必须将用户自身因素概率化，因此，本文提出了基于概率潜在语义分析方法建立用户模型，并利用查询反馈作为学习与更新用户模型的操作，实验结果表明，用户模型的构建可以准确捕获用户兴趣的主题所在，与用户真正语义意图是相吻合的。同时，通过查询的增量方式进行用户模型的更新，实验中也验证了对查询词进行主题分类的准确性。这也证明了，自身因素概率化过程与认知过程是完全相吻合的。