DOI：10.7652/xjtuxb201304022

个性化搜索中用户兴趣模型匿名化研究

李清华[1,3]，康海燕[1,2]，苑晓姣[3]，XIONG Li[4]，任俊玲[2]
(1.北京信息科技大学网络文化与数字传播北京市重点实验室，100192，北京；2.北京信息科技大学信息管理学院，100192，北京；3.北京信息科技大学计算机学院，100192，北京；4.艾默里大学数学与计算机科学系，30322，美国亚特兰大)

摘要：为了解决个性化搜索技术所潜在的用户隐私信息泄露的问题，提出了用户兴趣模型匿名化方法。首先根据用户兴趣模型之间的相似性将其聚类为满足 p-链接性的等价组，然后计算聚类后兴趣条目的权值。所谓的 p-链接性是指攻击者根据背景知识链接确定某一用户的概率不超过 p。该方法可实现用户兴趣模型匿名化以及兴趣倾向不发生改变，既保护了用户隐私信息，同时也保证了个性化检索性能。实验表明：随着相关结果个数的增多，匿名化后搜索结果的查全率基本能保证在50%以上，另外 p-链接性的减小对于查全率的影响并不是太大。

关键词：个性化搜索；用户兴趣模型；匿名化；隐私信息
中图分类号：TP312 **文献标志码**：A **文章编号**：0253-987X(2013)04-0131-06

User Profile Anonymization in Personalized Web Search

LI Qinghua[1,3], KANG Haiyan[1,2], YUAN Xiaojiao[3], XIONF Li[4], REN Junling[2]
(1. Beijing Key Laboratory of Internet Culture and Digital Dissemination, Beijing Information Science and Technology University, Beijing 100192, China; 2. School of Information Management, Beijing Information Science and Technology University, Beijing 100192, China; 3. Computer School, Beijing Information Science and Technology University, Beijing 100192, China; 4. Department of MathCS, Emory University, Atlanta 30322, USA)

Abstract: To solve potential user privacy leaking in personalized web search, an approach of anonymizing user profiles is proposed. According to the similarity of user profiles, the profiles are clustered as equivalences which meet p-linkability and then calculated the weight of interest items. Here p-linkability means that the probability that a certain user is determined by the attacker on the basis of the background knowledge link is less than p. This approach realizes the anonymization of user profiles and the constancy of the user interest tendency, and sufficiently protects user privacy while the anonymized user profiles are still effective in personalized web search. It is verified that after anonymization the recall ratio of the search result can be guaranteed over 50% as the related items increase. Meanwhile, the reduction of p-linkability does not greatly influence the recall ratio.

Keywords: personalized search; user profile; anonymization; privacy information

实现个性化服务，需要跟踪和学习用户的兴趣和行为，生成用户兴趣模型，根据用户兴趣过滤信息

收稿日期：2012-04-10。 作者简介：李清华(1987—)，男，硕士生；康海燕(通信作者)，男，副教授，硕士生导师。 基金项目：北京信息科技大学网络文化与数字传播北京市重点实验室开放课题资助项目(5026035409)；教育部人文社会科学资助项目(11YJC870011)；北京市教委科技计划面上资助项目(KM201211232014)；国家科技支撑计划资助项目(2012BAH08B02)。
网络出版时间：2012-11-16 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20121116.1039.001.html

以达到准确提供信息的目的。获得用户兴趣模型的方法有服务器端挖掘、用户主动提供、系统被动学习等，其中服务器端的挖掘是现在经常使用的一种方法，因特网中的服务器会生成访问日志文件来记录用户的访问信息。在这样的背景下，如果在服务器端保留原始的用户兴趣模型，开放的网络环境很有可能会造成用户隐私的泄露。另外，为了进行数据分析研究，有些机构会发布这些涉及到个人数据的信息，此时数据发布中的隐私保护技术并不适用于用户兴趣模型这种非结构化的数据。如何在保证用户隐私的前提下，有效提高用户兴趣模型在个性化搜索中的推荐作用以及进行数据共享是一个值得认真研究的问题。

在隐私保护研究方面，数据发布中隐私保护的技术主要有文献[1]提出的 k-匿名模型，它要求发布表中的每个元组都至少与其他 $k-1$ 个元组在准标识属性上完全相同。然而，k-匿名模型并不能防止同质性攻击以及背景攻击。针对 k-匿名模型中敏感值的问题，文献[2]提出了 L 多样性，要求每个准标识符(QI)分组中至少包含 L 个多样性的敏感属性取值。Li 等人进一步针对 L 多样性的不足，提出了 t-Closeness 的方法[3]，要求 QI 分组中每个敏感属性取值的分布较为接近。此外，还有学者对轨迹数据进行隐私保护[4]研究。然而，由于在个性化搜索中使用的数据往往是非结构化的，上述方法并不能完全适用于个性化搜索。目前，对于个性化搜索中的隐私保护研究还不是很具体，在文献[5]中提到个性化服务需要收集和集成大量的用户个人信息，需要对个性化数据进行隐私保护。文献[6]提出了使用聚类方法来生成用户兴趣模型簇，使用簇的质心代表每一个用户的兴趣模型。虽然这种方法能够保证用户的隐私不被泄露，但是破坏了用户兴趣的权值，使用簇质心过滤不能准确得到用户感兴趣的搜索结果。

本文的主要贡献有：①提出了用户兴趣模型相似性的计算方法；②提出了用户兴趣模型匿名化和权值计算方法，在保证检索质量的前提下，可防止用户的隐私泄露以及用户兴趣倾向失衡。

1 个性化搜索框架模型

个性化搜索是在普通查询的基础上，通过用户兴趣模型对初始查询结果进行过滤和排序，最终提供给用户符合其兴趣的个性化检索结果。个性化搜索引擎比一般搜索引擎多了 3 个部分：个性化需求分析器、个性化查询过滤器和用户兴趣模型库。其中用户兴趣模型库起着非常重要的作用，存放着用户兴趣模型，是实现个性化搜索的关键。个性化搜索引擎的框架如图 1 所示。

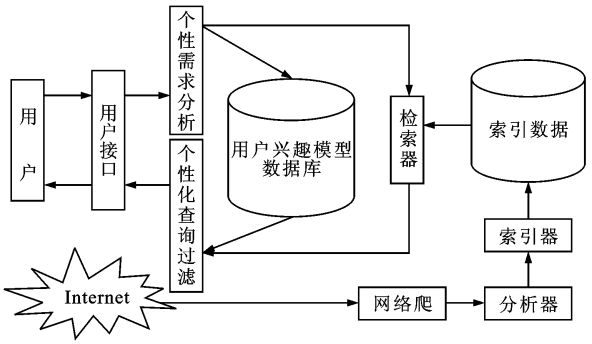

图 1 个性化搜索引擎框架

2 用户兴趣模型匿名化相关定义

定义 1(用户兴趣模型) 用户兴趣模型是基于关键词的向量空间模型，$U_P=\{t_{w1},t_{w2},\cdots,t_{wm}\}$，其中 $t_{wi}=\{t_i,w_i\}$，t_i 表示用户兴趣的关键字，每个 t_i 都有一个权值信息 w_i，权值越高，表明用户对这个关键字方面的信息兴趣越浓厚。例如：$U_P=\{$(体育，3)，(游戏，1)$\}$，表示这个用户可能对体育和游戏比较感兴趣，同时权重 3∶1 意味着这个用户更喜欢体育，更进一步地推测是该用户也会对同时包含体育和游戏的网页感兴趣。

定义 2(用户兴趣模型集) 用户兴趣模型集(User Profile Set)是用户兴趣模型的集合，$U_{PS}=\{U_{P1},U_{P2},\cdots,U_{Pn}\}$，$n$ 为用户兴趣模型个数。模型集示例如表 1 所示。

表 1 用户兴趣模型集

编号	兴趣条目
U_{P1}	(公务员，0.8)，(仰泳，0.6)
U_{P2}	(会计，0.5)，(蛙泳，1.0)

定义 3(等价用户组) 兴趣条目相同的用户即为同一等价用户组。根据兴趣条目的相似性，将用户兴趣模型匿名化为等价用户组。

定义 4(用户兴趣模型的相似性) 通过计算用户兴趣模型之间的余弦相似性得到用户兴趣模型的相似性，计算公式如下

$$U_P=\{(t_1,w_1),(t_2,w_2),\cdots,(t_m,w_m)\}$$

$$U'_P=\{(t'_1,w'_1),(t'_2,w'_2),\cdots,(t'_m,w'_m)\}$$

兴趣集合为

$$T=\{t_1,t_2,\cdots,t_m\}\cup\{t'_1,t'_2,\cdots,t'_m\}$$

$U_P(t)$是兴趣条目 t 在用户兴趣模型中的权值

$$U_P(t)=\begin{cases}w_i, & \text{若 } \exists(t_i,w_i)\in U_P,t_i=t\\0, & \text{否则}\end{cases}$$

U_P 和 U'_P 之间的夹角和距离分别为

$$\cos(U_P,U'_P)=\frac{\sum_{t\in T}U_P(t)U'_P(t)}{(\sum_{i=1}^{m}t_i^2)^{1/2}(\sum_{j=1}^{n}t_j^2)^{1/2}}$$

$$d(U_p,U'_p)=\frac{1}{\cos^2(U_p,U'_p)}$$

根据用户兴趣模型之间的相似性进行聚类,得到等价组兴趣模型,这样能够保证等价组内部兴趣倾向的一致性。

定义 5(用户兴趣模型的匿名化)　根据用户兴趣模型之间的相似性聚类成等价用户组兴趣模型,然后重新计算用户兴趣模型的权值,这样就能实现根据背景知识不能确定用户的目的,即保护用户隐私。

3　用户兴趣模型预处理

用户兴趣模型的匿名化依赖于不同用户兴趣条目的相似性,然而用户兴趣模型中的条目是非常随意的,即使两个用户拥有相同的兴趣条目,不同的词法表示都不能发现这种语义上的相似,因为同义词虽然语义上相同,但在词法上仍是不同的词。为了比较兴趣条目语义上的相似性,需要对进行匿名化的用户兴趣模型进行预处理,引入两个概念如下。

定义 6(同义词集合)　同义词集合包含该词及其所有同义词的集合。

定义 7(上位词)　上位词指概念上外延更广的主题词。例如:"花"是"鲜花"的上位词,"植物"是"花"的上位词。

预处理就是首先将用户兴趣模型中的每一个条目用其同义词集合代替,然后进行上位词集合的扩展,添加同义词集合的上位词集合,权重与同义词集合的权重相同。本文引入了同义词词林[7],以获取同义词集合和上位词。表 2 为表 1 数据预处理之后的结果。

根据用户兴趣模型中兴趣条目的同义词集合及上位词,能够使用户兴趣模型的相似性计算得较为准确。文献[8]指出,当用户提交一个查询后,如果对搜索引擎返回的结果不满意时,用户有可能会用语义相近的词代替进行查询。通过同义词集合以及上位词的扩展既能提高搜索的准确率,还能使用户兴趣的相似性计算更为准确。经过预处理后的用户

表 2　用户兴趣模型预处理结果

编号	兴趣条目
U_{P1}	({公务员、勤务员、办事员},0.8) ({仰泳、侧泳、爬泳、蛙泳、自由泳},0.6) ({职工、员工},0.8) ({溜冰、游泳、下棋},0.6)
U_{P2}	({会计、会计师、账房、出纳、出纳员、先生、管账},0.5) ({仰泳、侧泳、爬泳、蛙泳、自由泳},1.0) ({职工、员工},0.5) ({溜冰、游泳、下棋},1.0)

兴趣模型表示出的是不同用户兴趣模型在语义上的相似性,而不是词法上的相似性。这样使得用户兴趣模型的匿名化变得更加可行和高效。经过预处理后的用户兴趣模型的相似性计算结果如表 3 所示。

表 3　预处理后的用户兴趣模型相似性

模型状态	$\cos(U_P,U'_P)$	$d(U_P,U'_P)$
原始的用户兴趣模型 U_{P1}、U_{P2}	0.000	$+\infty$
同义词集合替换后的结果	0.536	1.865
上位词扩展后的结果	0.714	1.401

4　用户兴趣模型匿名化算法设计

4.1　用户兴趣模型匿名化过程

文献[6]将用户兴趣模型聚类为簇,用簇质心代表每一个用户兴趣模型(簇质心方法)。表 1 数据得到的簇质心见表 4。

表 4　簇质心模型

t_i	w_i		
	U_{P1}	U_{P2}	平均值
公务员	1.0	0.0	0.5
会计	0.0	0.6	0.3
仰泳	0.8	0.0	0.4
蛙泳	0.0	1.0	0.5

从表 4 可以看出,聚类后的簇质心权值的平均值已经和原始的用户兴趣模型的权值有很大的不同,例如平均值中公务员的权值为 0.5,会计的权值为 0.3,而 U_{P2} 公务员权值为 0,会计的权值为 0.6。簇质心模型已经在很大程度上破坏了 U_{P2} 用户的兴趣倾向。

本文提出的方法首先根据同义词集合替换后的

结果生成等价组兴趣模型 EUP(Equivalence User Profile)的兴趣条目集合 $T'=U_{P1}(T)\cup U_{P2}(T)\cup\cdots\cup U_{Pm}(T)$,然后根据原始用户兴趣模型计算权值,计算公式如下

$$U_{Pi}(t)=\begin{cases}w, & \text{若 } t\in U_{Pi}\\ \dfrac{\sum_{i=1}^{n}U_{Pi}(t)}{n}, & \text{否则}\end{cases}$$

计算结果如表 5 所示。

表 5 等价组兴趣模型

EUP	{公务员、勤务员、办事员}	{仰泳、侧泳、爬泳、蛙泳、自由泳}	{会计、会计师、账房、出纳、出纳员、先生、管账}
U_{P1} 的权值	1.0	0.8	0.3
U_{P2} 的权值	0.5	1.0	0.6

从表 5 可以看出,等价组兴趣模型保留着原始用户兴趣模型的大部分内容,而且兴趣条目的权值与原来的兴趣倾向没有太大改变。另外,等价组兴趣模型相对于原始的用户兴趣模型添加了一些噪声数据,例如对于 U_{P2},{公务员、勤务员、办事员}为噪声数据,保证了用户兴趣模型的隐私。

定义 8(p-链接性) 根据相似性将用户兴趣模型匿名化成不同的等价用户组,攻击者根据背景知识链接确定某一用户的概率不超过 p。

定义 9(满足 p-链接性的匿名化) 用户兴趣模型匿名化的目标是实现用户兴趣模型集满足 p-链接性的隐私保护需求。当且仅当对于等价组兴趣模型中的每一个兴趣条目,根据该条目链接到某一用户的概率不超过 p 时,该组兴趣模型满足 p-链接性的隐私需求。当且仅当所有的等价组兴趣模型满足 p-链接性隐私需求时,用户兴趣模型集才满足 p-链接性隐私需求。

定义 10(背景知识) 攻击者从其他渠道获得的一些目标对象的信息[9],例如用户兴趣模型集的大小,每一个用户兴趣模型中条目的个数等等。本文中,等价组兴趣条目的大小以及用户的原始兴趣被认为是背景知识。经过用户兴趣模型匿名化处理后,根据背景知识确定用户的概率为

$$p_b(t)=\frac{|U_{Pi}|}{|C|}$$

式中:$|U_{Pi}|$ 为原始兴趣模型中兴趣条目个数;$|C|$ 为匿名化后兴趣条目个数。

4.2 用户兴趣模型匿名化算法

用户兴趣模型匿名计算法如下。

```
|UPS|为集合中用户兴趣模型的个数
result ← 最终的返回结果,初始值为空
while |UPS|>0
  up ← 从 UPS 中选取第一个用户
  UPS ← UPS — up
  equiUserGroup ← 创建 up 所属的等价组对象
  while equiUserGroup 不满足 p-链接性 && |UPS|>0
      mostSimilarUser ← 从 UPS 中选取和 up 相似度最大的用户兴趣
      UPS ← UPS — mostSimilarUser
      equiUserGrou ← equiUserGroup ⋃ mostSimilarUser
    end while
    if equiUserGroup 满足 p-链接性 then
    计算等价组兴趣模型对应的用户的权值
    result ← result ⋃ equiUserGroup
    end if
end while
return result
```

5 实 验

为简单起见,本次实验只对兴趣条目进行了上位词的扩展。实验数据集采用的是 Sogou 搜索引擎发布的 2008 年 6 月部分网页查询需求及用户点击情况的网页查询日志,选取了其中的 3 500 条查询记录进行实验。数据格式为:“用户 ID\t[查询词]\t 该 URL 在返回结果中的排名\t 用户点击的顺序号\t 用户点击的 URL”。其中,用户 ID 是根据用户使用浏览器访问搜索引擎时的 Cookie 信息自动赋值,即同一次使用浏览器输入的不同查询对应同一个用户 ID,数据集如图 2 所示。由于仅仅根据 Cookie 信息不能作为用户标识的标准,本文选取 Cookie 信息对应查询条目大于 8 的记录作为实验对象。

5.1 实验步骤

首先按行读取数据,对所有用户查询记录进行分词,去除“啊、的、是”等停用词,形成用户原始兴趣模型,初始权值均为 1.0。对于分词采用的是 Lucene 中文分词组件 JE-Analysis,例如“北京音响的出租报价”对应的兴趣为(北京,1.0)(音响,1.0)(出租,1.0)(报价,1.0)。其次,进行上位词的扩展,上

```
1011517038707826 [诺基亚手机主题下载] 14 62 zhuafeng.com/
1011517038707826 [主题]　　4 39 mskin.qq.com/?from=qq
1011517038707826 [手机主题]8 46 www.ownskin.com/
1011517038707826 [主题]　　6 40www.mytopic.cn/
1011517038707826 [主题]　　2 37www.ownskin.com/
1011517038707826 [诺基亚手机主题下载] 2 49mskin.qq.com/
1011517038707826 [n73手机主题] 4 551860sj.com.cn/s/N73
1011517038707826 [手机主题] 10 47desktop.shouji56.com/
```

图 2　实验数据集

位词扩展是按照前面所述的根据同义词词林进行扩展。这里借用了哈工大(HIT)信息检索研究中心(CIR)语言技术平台(LTP)提供的 LTP WebService 来获取上位词。例如"北京"对应上位词的分类号为 Di03(首都_省会),对于没有上位词的情况返回值为-1。然后,形成用户兴趣模型,按照相似度对用户兴趣模型进行匿名化。最后,根据原始兴趣模型和匿名化的兴趣模型分别进行查询,分析兴趣模型匿名化对于个性化搜索的影响。

5.2　实验结果

对于不同的 p 值,生成的等价组个数如图 3 所示。可以看出,随着 p 值的减小,生成的等价组个数逐渐减小。这是因为 p 值越小,链接确定该用户的概率越小,故等价组的兴趣条目增多,总数固定的兴趣条目生成的等价组个数相应的会逐渐减少。另外,经过上位词扩展后和未经过上位词扩展的等价组个数稍有区别,这是因为经过上位词扩展后得到的相似度最大的用户兴趣模型和未经过上位词扩展所得到的不同,导致匿名化的过程和结果不同。

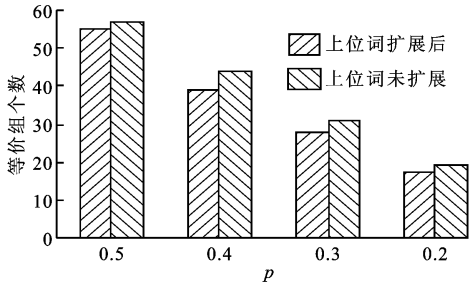

图 3　p 对应的等价组个数

图 4 显示了本文提出的匿名化兴趣模型、文献[6]提出的簇质心模型以及普通搜索的查全率比较(取搜索的前 30 条进行比较)。可以看出,本文提出的匿名化方法稳定性要优于簇质心模型。这从另一个角度说明了簇质心并不能非常准确地代表用户的兴趣倾向。另外,匿名化后的个性化搜索查全率仍然高于普通搜索的查全率,匿名化的兴趣模型对于个性化搜素影响不大。

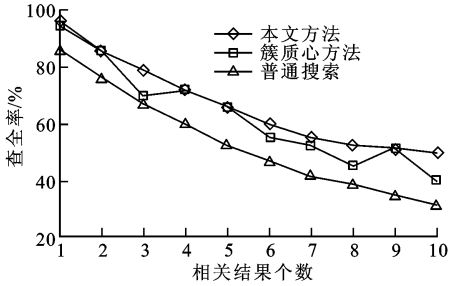

图 4　几种方法查全率的比较

图 5 显示了不同的 p 值对应的查全率,可以看出,随着相关结果个数的增多,匿名化后搜索结果的查全率基本能保证在 50%以上。同时可以看出,p-链接性的减小,对于查全率的影响并不是太大。这是因为虽然匿名化根据相似度添加了其他用户的兴趣条目,但是该用户的用户倾向即兴趣条目的权值仍能代表原来的兴趣倾向。

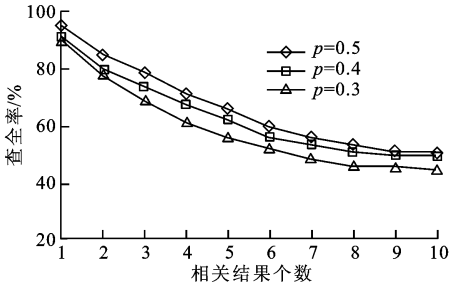

图 5　不同 p 值对应的查全率

图 6 为几种方法搜索准确率的比较。搜索准确率采用平均准确率评价标准,计算公式为

$$\bar{P}=\frac{\sum_{i=1}^{s}P(i)\mathrm{rel}(i)}{r}$$

式中:r 表示为相关结果的数目;s 表示选取前 s 条搜索结果和相关结果进行比较;$P(i)$表示截止到第 i 个搜索结果,相关结果所占的比例;$\mathrm{rel}(i)$表示第 i 个结果是否为相关的结果,值为 0 或 1。从图 6 中可以看出,本文提出的用户兴趣模型匿名化方法得到的搜索准确率要高于普通搜索。同时,随着相关结果个数的增多,本文方法得到的搜索准确率变化幅度不大,这也说明了本文方法得到的用户兴趣倾向没有遭到破坏。与簇质心方法相比,本文方法更能准确地代表用户的兴趣倾向。

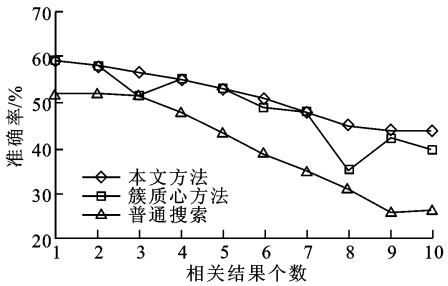

图 6 几种方法搜索准确率的比较

6 总 结

本文基于 k-匿名的匿名化思想以及引入同义词词林中的同义词集合和上位词，对个性化搜索中的关键部分用户兴趣模型进行了匿名处理，生成满足 p-链接性的等价组兴趣模型，从而在保证用户隐私的情况下提高了个性化搜索质量。个性化搜索中隐私保护的研究有非常广泛的应用前景，但其今后的发展也面临越来越多的挑战。随着信息技术的发展、个性化系统的广泛应用，个性化服务中隐私保护技术也会得到更多的重视和研究。

参考文献：

[1] SWEENEY L. K-anonymity: a model for protecting privacy [J]. Int'l Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[2] MACHANAVAJJHALA A, GEHRKE J, KIFER D. l-Diversity: privacy beyond K-anonymity [C]// Proc of the 22nd Int'l Conf on Data Engineering. Piscataway, NJ, USA: IEEE Computer Society, 2006: 24-35.

[3] LI N, LI T, VENKATASUBRAMANIAN S. t-Closeness: privacy beyond k-anonymity and l-diversity [C]// Proc of the 23rd Int'l Conf on Data Engineering. Piscataway, NJ, USA: IEEE Computer Society, 2007: 106-115.

[4] 霍峥，孟小峰. 轨迹隐私保护技术研究 [J]. 计算机学报，2011，34(10)：1820-1829.
HUO Zheng, MENG Xiaofeng. A survey of trajectory privacy-preserving techniques [J]. Chinese Journal of Computers, 2011, 34(10): 1820-1829.

[5] 朱青，赵桐，王珊. 面向查询服务的数据隐私保护算法 [J]. 计算机学报，2010，33(8)：1315-1323.
ZHU Qing, ZHAO Tong, WANG Shan. Privacy preservation algorithm for service-oriented information search [J]. Chinese Journal of Computer, 2010, 33(8): 1315-1323.

[6] YUN Zhu, LI Xiong, CHRISTOPHER V. Anonymization of user profiles for personalized web search [C]// Proceedings of the 19th International Conference on World Wide Web. New York, NY, USA: ACM, 2010: 1125-1126.

[7] 田久乐，赵蔚. 基于同义词词林的词语相似度计算方法 [J]. 吉林大学学报：信息科学版，2010，28(6)：602-608.
TIAN Jiule, ZHAO Wei. Words similarity algorithm based on Tongyici Cilin in semantic web adaptive learning system [J]. Journal of Jilin University: Information Science Edition, 2010, 28(6): 602-608.

[8] 余慧佳. 基于大规模日志分析的搜索引擎用户行为分析 [J]. 中文信息学报，2007，21(1)：109-114.
YU Huijia. Research in search engine user behavior based on log analysis [J]. Chinese Journal of Information, 2007, 21(1): 109-114.

[9] FUNG B C M, WANG K, CHEN R, et al. Privacy-preserving data publishing: a survey of recent developments [J]. ACM Computing Surveys, 2010, 42(4): 1-14, 53.

（编辑 杜秀杰 赵大良）

（上接第 117 页）

[J]. IEEE Transactions on Computers, 2007, 56(3): 344-357.

[8] 舒怀林. PID 神经元网络多变量控制系统分析 [J]. 自动化学报，1999，25(1)：105-111.
SHU Huailin. Analysis of pid neural network multivariable control systems [J]. Acta Automatica Sinica, 1999, 25(1): 105-111.

[9] SAATY T L. The analytic hierarchy process [M]. New York, USA: McGraw-Hill, 1980: 170-205.

（编辑 管咏梅 武红江）