doi: 10.3969/j.issn.1006-1576.2012.10.014

基于用户偏好的地理计算应用检索

何攀[1]，刘露[1]，陈荦[1]，王祖文[2]
（1. 国防科学技术大学电子科学与工程学院，长沙 410073；2. 北京科技信息研究中心，北京 100036）

摘要：针对现有检索系统对用户过去行为和历史数据关注较少的问题，笔者提出一种基于用户偏好的地理计算应用检索方法。以地理计算领域中的应用检索为背景，基于互联网信息检索中的用户偏好模型，通过分析用户搜索地理计算应用历史，充分考虑用户对地理计算应用的最近访问时间对确定用户偏好的影响，提取用户偏好因子矩阵和时间因子矩阵，算出与用户查询最相关的地理计算应用。实验结果表明，该方法能有效提高用户对地理计算应用的检索效率，为地理计算应用检索优化提供有力支撑。

关键词：地理计算应用检索；用户偏好模型；偏好因子矩阵；时间因子矩阵
中图分类号：TJ861　　**文献标志码：**A

A Retrieval Algorithm of Geographic Computing Applications Based on User Preferences

He Pan[1], Liu Lu[1], Chen Luo[1], Wang Zuwen[2]
(1. *College of Electronic Science & Engineering*, *National University of Defense Technology*, *Changsha* 410073, *China*;
2. *Beijing Science & Technological Information Research Center*, *Beijing* 100036, *China*)

Abstract: Pay less attention to the behavior of users in the past and historical data for existing retrieval system problems, the author presents a geographical computing applications retrieval method based on user preferences. Under the background of geographic computing application retrieval, based on user preference model in internet information retrieval, analyze user searches for geographic computing applications history, and give full consideration to the impact of the last access time of geographical computing applications to determine user preferences, then extract user preference factor matrix and the time factor matrix, and calculate the most relevant geographic computing applications to the user's query. The experiment result shows that this method can availably improve the user geographical computing application retrieval efficiency, and provide strong support for the optimization of geographical computing application retrieval.

Key words: geographic computing applications retrieval; user preference model; preference factor matrix; time factor matrix

0　引言

地理计算是地理信息科学的核心内容之一，主要研究地理信息科学的方法学问题，包括算法、建模和计算体系。地理计算应用是随着高性能地理计算在 Web2.0、云计算和社会计算等基础支撑技术的快速发展而产生的新的地理计算程序与数据的应用模式。传统的地理计算应用通常以地理信息系统（geographic information system，GIS）中地理空间分析工具箱的形式提供，例如 ArcGIS 中的 ArcToolBox，Quantum GIS 中的插件，MapWindow 中的 GIS Tools 等。在传统的应用模式下，GIS 软件套件期望能够提供功能尽可能全面，性能尽可能优越的工具箱。这样往往会导致工具箱的使用方法复杂，学习成本高。事实上，绝大多数用户对于地理计算工具箱的需求往往只有其中一个或者少数几个功能，地理计算应用的个性化定制服务需求越来越迫切。GISCloud 就是在此背景下产生的一个典型应用[1]，但它只实现了地理空间数据集的个性化定制与分享服务。HiGIS Online 是一个在国家 863 计划资助下致力于实现地理计算应用个性化定制与服务的高性能地理计算系统[2]，其中需要解决的一个关键问题就是如何使用户在众多的地理计算应用库中找到适合自己的程序或数据。

在移动计算与应用领域，已经有成熟的类似产品，例如用于苹果 iOS 设备的应用商店（App Store），用于 Android 设备的谷歌应用市场（Google Play），以及移动应用搜索引擎 Chomp、AppExplorer 等。这些在近几年呈现爆炸式发展的产品已经得到了用

收稿日期：2012-06-01；修回日期：2012-07-26
基金项目：国家“十二五”863 项目基金“面向新型硬件架构的复杂地理计算平台”（2011AA120306）
作者简介：何攀（1985—），男，湖南人，硕士，从事信息处理与信息系统技术研究。

户的认同，并已成为移动应用的绝对主流模式和事实上的标准。通过使用这些产品，用户能够在大量的应用中较快地定位到自己所需要的程序。然而，在这些应用搜索引擎当中，其查询接口通常仅限于应用名称中的关键字，在实际应用中有较大的局限性。当应用的命名与其功能关联性不强时，这种搜索方式就会失效。

现有的检索系统大多是从检索模型和信息加工过程来提高检索的准确性，并没有给予用户更多的关注。以网络搜索引擎为例，不同背景的用户使用相同的提问来查询，得到的结果没有区别，相同的用户在不同的时间做的相同查询，得到的结果也没有区别，使得用户不容易发现自己的搜索偏好。造成这种结果的主要原因是，在这些服务系统中，没有考虑到用户信息的存在，也就是用户过去的行为和用户的历史数据。地理计算应用的用户通常是具有特定地理计算领域专业背景知识的群体。对于地理计算应用程序的选择更具有较强的专业性和方向性，其搜索历史可作为地理计算领域深度检索的有效样本数据。因此，在地理计算应用的检索过程当中，很有必要引入用户偏好模型来提高检索的效率。

基于用户由稳定兴趣驱动访问 Web 的频率远远高于偶然兴趣的驱动，一定时间段的网络日志中一定蕴含了用户的稳定兴趣[3]的结论，笔者通过用户查询日志对用户偏好进行提取，并且充分考虑时间对于用户偏好的影响，设计实现了一种基于关键词和用户偏好相结合的应用检索算法，将其应用于实际的应用查询中，以提高准确率和召回率。

1　国内外研究现状

1.1　地理计算应用检索

地理计算应用目前主要以集成在地理信息系统计算平台如 ArcGIS 和 MapWindow 上的方式来进行使用。ArcGIS 采用工具箱的方式将众多地理计算应用作为工具集成在一起，用户在使用前必须先了解每种工具的具体功能，才能选择适应自己需求的地理计算工具。MapWindow 4.x 则是将地理计算应用以插件的形式集成到平台当中，这种方式相对工具箱来说更具有灵活性，可以根据自身需要添加地理计算插件；MapWindow 6.x 结合了前 2 种方式，既使用工具箱将常用的地理计算应用组织在一起，同时保留插件机制以便用户添加特定的地理计算应用，并且在工具箱中采用关键字对工具进行查询。

上述对地理计算应用的组织方式对于数量和种类较少的地理计算应用具有一定的可行性，但是实际情况下，地理计算应用的数量和种类会越来越多，因此，仅仅使用关键字对其进行查询是无法满足用户需要的。

1.2　用户偏好提取

用户偏好也可称为用户兴趣，其信息是相对稳定的、时间相对长久的信息需求[4]。对于目前的“信息过载”问题，普通的检索工具仍然无法满足不同背景、不同目的、不同时期的个性化信息需求[5]；因此，需要对不同的用户提供能够尽可能满足用户需求的个性化服务，而支撑个性化服务的一个重要因素即是对用户偏好的有效提取。

随着个性化服务研究的发展，用户偏好提取方法的研究也进一步深入，文献[6]面向“移动信息过载”问题，提出一种基于认知心理学的用户偏好提取方法(cognitive psychology-based approach to user preference elicitation，CPUE)，将加工水平模型和分布式认知理论引入移动用户偏好提取过程，以期提取更全面、精确的用户偏好。该方法突破了传统基于“记忆两过程理论”划分“长期偏好”和“短期偏好”的模式，强调了语义层次认知的重要性，同时根据分布式认知理论，更加强调各种环境因素的认知差异对用户偏好的影响。文献[7]提出了基于领域特征构建个性化垂直搜索算法(PVSA)，其中的领域主题偏好策略为基于域网页日志挖掘行业主题偏好，根据领域兴趣提升域网页权重，该策略是一种相对有效且高效的域兴趣分析、应用方法。

鉴于用户偏好提取对于缓解“信息过载”问题，以及个性化服务深入研究的重要性，同时，地理计算应用检索具有明显的领域特征；因此，用户偏好提取对于地理计算应用检索将具有重要意义。

1.3　信息检索的评测方法

信息检索领域存在着几个已经被广为接受的性能指标。最基本的性能指标定义是准确率和召回率。借鉴网页搜索中准确率和召回率的定义，应用检索的准确率和召回率定义如下：

假设给定一个查询 q，在某数据集合上有 N_{q} 个应用是符合 q 所隐含的信息需求的。某个检索系统返回的结果中，有 N_{correct} 个是符合要求的，也就是说在正确的 N_{q} 个应用的范围之内；另有 $N_{\mathrm{Incorrect}}$ 个不符合要求，即落在 N_{q} 个应用的范围之外。检索系

统一共返回了(N_{correct}+$N_{\text{Incorrect}}$)个结果。

定义 1 准确率(precision)，定义为

$$\text{precision}=\frac{N_{\text{Correct}}}{N_{\text{Correct}}+N_{\text{Incorrect}}} \tag{1}$$

定义 2 召回率(recall)，定义为

$$\text{recall}=\frac{N_{\text{Correct}}}{N_{\text{q}}} \tag{2}$$

准确率反映了检索系统对某个查询返回结果中正确结果的比例。召回率反映了检索系统对某个查询返回的结果中正确结果占全部正确结果的比例。

2 基于用户日志的偏好提取

文献[3]指出网络日志挖掘旨在通过对网络日志进行有效的数据挖掘，发掘隐藏在日志数据背后的 Web 用户访问模式。这个目标基于这样的假设：网络日志中确实蕴含了用户访问 Web 的某些规律特性，这些特性通常表现为一些特定的模式。如果能将这些模式挖掘出来并加以利用，那将大大提升与用户偏好相关的网络应用的服务质量。绝大多数的网络日志挖掘研究都是基于这一假设发掘出了各种有用的 Web 用户访问模式[8-9]。例如，很多研究旨在找出用户的频繁访问路径或用户的频繁访问模式，认为找出的频繁路径或模式一定能够反映出网络日志中蕴含的 Web 用户的偏好[10-11]。基于以上网络日志信息能反映出用户偏好的结论，用户搜索、安装与使用地理计算应用的记录，对分析用户对某领域或某类地理计算应用的偏好同样具有指导意义。

为了获得用户对地理计算应用的使用记录，如程序名称、使用次数、最近使用时间等信息，笔者设计了用户日志表用以专门存储这些信息，表的结构如图 1 所示。

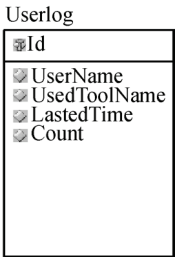

图 1 用户日志表

如果用户 U 使用过应用 P，则说明用户 U 对应用 P 有所偏好，但偏好的强度与用户使用该应用的频率高低有关[12-13]。因此笔者将偏好因子定义如下。

定义 3 偏好因子指用户对某个特定应用的使用频率。用户 U 对应用 P 的偏好因子定义为：设 C_{p} 为用户 U 使用应用 P 的次数，C_{U} 为用户 U 使用应用的总次数，则偏好因子 α 为

$$\alpha=C_{\text{p}}\Big/C_{\text{U}} \tag{3}$$

此外，用户的偏好是随着时间的推移而不断发生变化的；因此时间对于用户的偏好因子的影响必须予以考虑。通常情况下，越新的选择越能表现用户此段时间的偏好方向，故时间因子的定义如下。

定义 4 时间因子。设用户 U 使用应用 P 的最近时间 t_{last}，当前时间为 t_{now}，则时间因子 τ 为

$$\tau=\frac{1}{t_{\text{now}}-t_{\text{last}}} \tag{4}$$

定义 5 兴趣矢量假设对于应用 P，有 n 个用户对其进行过使用，则其偏好矢量表示为 $\boldsymbol{A}=(\alpha_1,\alpha_2,\cdots,\alpha_n)$，矢量中的 α_n 为用户 U_n 对应用 P 的偏好因子。

定义 6 时间因子矢量对于应用 P 和 n 个使用过它的用户，其时间因子矢量表示为 $\boldsymbol{T}=(\tau_1,\tau_2,\cdots,\tau_n)$，矢量中的 τ_n 为用户 U_n 对应用 P 的时间因子。

定义 7 相关度判定模型中的相关度判定是根据应用的总偏好度来进行计算的。因为时间因子对用户偏好的影响，所以将时间因子作为偏好因子的权值，加权求和计算应用 P 总的偏好度，即相关度：

$$r=\boldsymbol{A}\cdot\boldsymbol{T}=\sum_{i=1}^{n}\alpha_i\times\tau_i \tag{5}$$

相关度 r 越大，说明这 n 个用户对应用 P 的偏好度越高；因此，应用 P 也更有可能与用户的查询需求相符。

3 基于用户偏好的地理计算应用检索算法

该算法需要较大规模的用户日志数据，以保证用户对某个应用的使用频率能够正确反映用户的偏好。为了提高查询的效率，笔者在查询之前就建立好偏好因子矩阵和时间因子矩阵；同时只对与查询关键字相关的地理计算应用计算相关度，在一定程度上减少了计算量。算法的具体描述如下：

1) 统计用户日志中的用户总数和应用总数，建立用户偏好因子矩阵 $\boldsymbol{A}(m\times n)=\begin{bmatrix}\alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n}\\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n}\\ \vdots & \vdots & & \vdots\\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn}\end{bmatrix}$

和时间因子矩阵 $\boldsymbol{T}(n\times m)=\begin{bmatrix}\tau_{11} & \tau_{12} & \cdots & \tau_{1m}\\ \tau_{21} & \tau_{22} & \cdots & \tau_{2m}\\ \vdots & \vdots & & \vdots\\ \tau_{n1} & \tau_{m2} & \cdots & \tau_{nm}\end{bmatrix}$，其中：$m$ 为应用总数；n 为用户总数。用户偏好因子矩阵的元素 α_{ij} 为用户 j 对于应用 i 的偏好因子，计算方法如公式 (3)；时间因子矩阵的元素 τ_{ij} 为用户 i 对于应用 j 的时间因子，计算方法如公式 (4)。

2) 输入查询关键字 q，用户希望查找到与 q 相关的应用。

3) 在偏好因子矩阵中查找出与关键字 q 相关的应用，并从中提取出其偏好因子矢量，同时从时间因子矩阵中提取出相应应用的时间因子矢量。

4) 对第 3 步得到的偏好因子矢量和时间因子矢量求点积，计算得到与关键字 q 相关的应用的相关度 r，计算方法如公式 (5)。

5) 按照相关度从大到小的顺序对应用进行排序，选取排在前面的 10 个[14]应用作为输出显示给用户。

该算法通过计算对时间因子进行加权的用户总的偏好度，以判断查询结果集中的地理计算应用是否与用户的查询需求相关。算法的计算量主要集中在偏好因子矩阵和时间因子矩阵的建立上，还包括对所有地理计算应用的遍历、相关度的计算以及对查询结果集中的地理计算应用进行排序。假设用户总数为 n，地理计算应用总数为 m，查询结果集中的地理计算应用为 $k(k<m)$，采用快速排序算法对查询结果集进行排序，则算法总的时间复杂度为

$$\boldsymbol{T}=O(n\times m)+O(m)+O(k\log k) \tag{6}$$

由式 (6) 可以看出，算法的时间复杂度主要由第 1 项决定，因此算法的时间复杂度为 $O(n\times m)$。

算法的执行过程可以通过流程图的形式描述，如图 2。

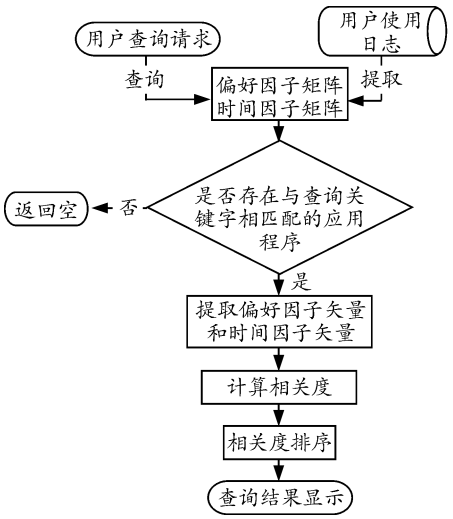

图 2　算法流程

4　基于用户偏好检索的应用实现

笔者实现了前面提出的基于用户偏好的地理计算应用检索算法，并将其集成于自主开发的地理信息系统平台 HiGISDesktop 中。利用该算法实现对工具箱中的地理计算应用进行检索，将最有可能符合用户需求的 10 个地理计算应用显示在查询结果中，并按照算得的偏好度进行排序。图 3、图 4 为不同用户输入相同查询内容的查询结果。

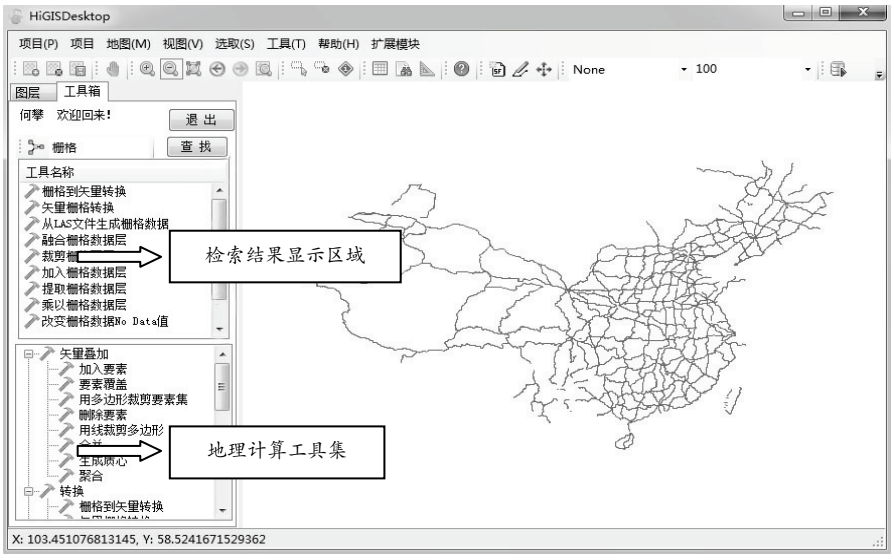

图 3　检索结果

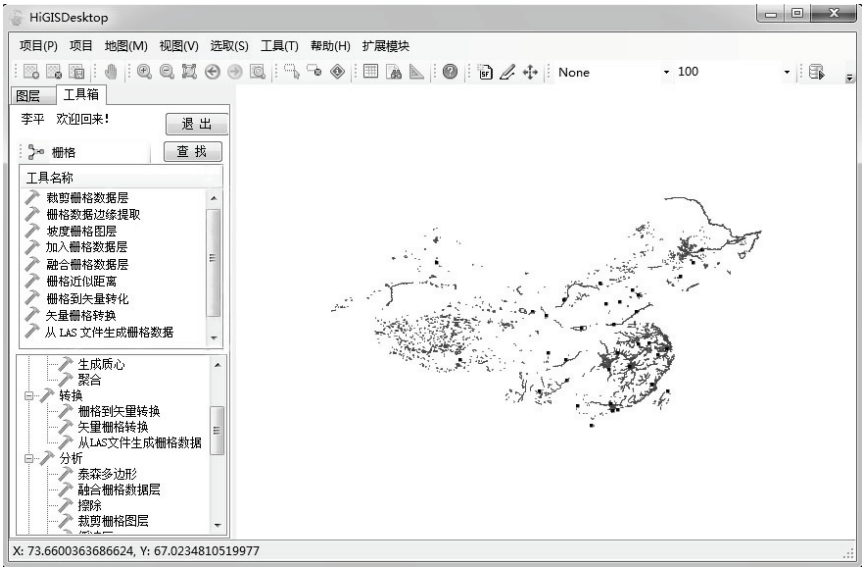

图 4　检索结果

图 3 为用户何攀输入“栅格”进行查询后所得到的查询结果；图 4 为用户李平同样输入“栅格”进行查询后所得到的查询结果。由用户记录所提取的用户偏好得出，用户何攀平时更多的是进行数据转换，而用户李平处理得更多的是图层数据的操作。

从图 3、图 4 的检索结果显示区域可以看出，图 3 中栅格数据转换的地理计算应用显示在前面，图 4 中图层处理的地理计算应用显示在前面，这与提取的用户偏好是一致的。由此可见，笔者提出的基于用户偏好的地理计算应用检索算法，在结合关键字查询的同时，利用用户日志中提取的用户偏好信息对查询结果进行筛选和排序，能够更好地满足用户的实际查询需求。

5　总结与展望

笔者针对地理计算应用检索的需要，设计实现了基于用户偏好的地理计算应用检索算法，在结合关键字查询的基础上，将检索结果按照用户偏好进行筛选和排序，以提高检索的准确性。实验结果表明，该算法能够有效提高地理计算应用检索的准确性，向用户呈现其更有可能需要的检索结果。该算法已经成功应用于 HiGISDesktop 地理计算平台，在今后的工作中，需要根据地理计算应用领域的特殊性，考查更多的影响用户偏好的因素，进一步提高算法的检索性能。

参考文献：

[1] The GIS Cloud Team. GIS Cloud[OL]. [2012-05-31]. http://www.giscloud.com/.

[2] HiGIS. [OL]. [2012-05-31]. https://twitter.com/highgis.

[3] 郭岩. 网络日志中用户兴趣挖掘及其利用[D]. 北京: 中国科学院计算技术研究所, 2004.

[4] Craig Silverstein, Monika Henzinger, Hannes Marais, et al. Analysis of a very large Web search engine query log[C]. New York: In SIGIR Forum, 1999, 33(1): 6-12.

[5] Zeng C, Xing CX, Zhou LZ. A survey ofpersonalization technology[J]. Journal of Software, 2002, 13(10): 1952-1961.

[6] 张磊, 陈俊亮, 孟祥武, 等. 基于用户偏好的垂直搜索算法[J]. 电子科技大学学报, 2010, 39(1): 91-96.

[7] 王立才, 孟祥武, 张玉阶. 移动网络服务中基于认知心理学的用户偏好提取方法[J]. 电子学报, 2011, 39(11): 2547-2553.

[8] 邢东山, 沈钧毅, 宋擒豹. 从 Web 日志中挖掘用户浏览偏爱路径[J]. 计算机学报, 2003, 26(11): 1518-1523.

[9] 崔航, 文继荣, 李敏强. 基于用户日志的查询扩展统计模型[J]. 软件学报, 2003, 14(9): 1593-1599.

[10] 刘永利, 欧阳元新, 闻佳, 等. 基于概念聚类的用户兴趣建模方法[J]. 北京航空航天大学学报, 2010, 36(2): 188-192.

[11] 王继民, 陈钟, 彭波. 大规模中文搜索引擎的用户日志分析[J]. 华南理工大学学报: 自然科学版, 2004, 32(l): 1-5.

[12] Mark Grechanik, ChenFu, Qing Xie, et al. A Search Engine For Finding Highly Relevant Applications[C]. ICSE ’10, Cape Town: ACM Press, 2010: 475-484.

[13] Jung SY, Hong JH, KimTS. A statisticalmodelforuser preference[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 834-843.

[14] 余慧佳, 刘奕群, 张敏, 等. 基于大规模日志分析的网络搜索引擎用户行为研究[J]. 中文信息学报, 2007(1): 109-114.