文章编号: 1001-1595(2011)05-0646-09

面向空间信息智能分发的动态化用户偏好模型研究

李新广, 范明虎, 杜 武
武汉大学 测绘遥感信息工程国家重点实验室, 湖北 武汉 430079

Research on Dynamic User Profile Model for Intelligent Distribution of Spatial Information

LI Xinguang, FAN Minghu, DU Wu
State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: User profile modeling method is the key bottleneck that restricts theory of intelligent distribution of spatial information to progress. Existing algorithms and systems of intelligent distribution of spatial information have drawbacks of inaccurate spatial location and biased utility, etc. and are mostly concerned on the contribution of the user's retrieval behavior to the profile model, but not consider time factors at all, and pay little attention to the role of user feedback. In view of this, the theories and algorithms of the existing literature are extended, by introducing concepts and arithmetic of region number, interest degree, interest degree density, etc., and dynamic factors of weight attenuation function and user information feedback, etc., to make a model able to adjust more accurately and in time with the change of user preference profile. The experimental results show that, compared with traditional static models, the model can more effectively reflect the change of user preference profile.

Key words: intelligent distribution of spatial information; dynamic user profile model; spatial information service; weight attenuation function; information feedback

摘 要: 用户偏好模型的构建方法是制约空间信息智能分发理论取得进展的关键瓶颈。现有的空间信息智能分发算法和系统存在空间范围定位不准确、效用度计算存在偏差等缺陷, 且大多关注用户的检索行为对偏好模型的贡献却均未考虑时间因素的影响, 也很少注意到用户反馈的作用。鉴于此, 对现有文献的理论和算法进行扩展, 通过引入区域数、兴趣度、兴趣度密度等概念和算法, 以及权值衰减函数和用户信息反馈等动态化因素, 使模型能够更为准确、及时地随着用户偏好特征的变化进行修正。试验表明, 相较于传统的静态模型而言, 该模型能够更为有效地反映用户偏好特征的变化。

关键词: 空间信息智能分发; 动态化用户偏好模型; 空间信息服务; 权值衰减函数; 信息反馈
中图分类号: P208 文献标识码: A
基金项目: 国家自然科学基金(40971243)

1 引 言

空间信息智能分发是主动空间信息服务技术的重要组成部分[1], 近年来得到了广泛关注。20 世纪 90 年代以来, 以美国为代表的西方国家就开始研发空间信息智能分发系统。1996 年, 美国开始研发战场警觉及数据分发系统[2-3], 1999 年开始实施全球信息栅格项目[4]。2001 年美国提出"智能节点"的概念, 并于 2003 年结合"网络中心战"的思想开始在军事决策系统中投入使用[5-7]。2004 年, 美国启动战术级作战人员信息网项目[8]。在国内, 近几年也出现了一些相关研究, 但主要集中在分发系统架构层面, 且多是对国外相关研究的一些介绍[9-11]。总体而言, 由于空间信息的独特性及复杂性, 用户偏好模型的构建及其效用度评估算法一直是制约空间信息智能分发研究取得进展的关键瓶颈。

夏宇针对遥感数据的分发, 探索性地采用区间数表达具有区间范围特征的经度、纬度、时间、频谱和空间分辨率等指标的用户检索特征[12], 并通过对 TOPSIS 方法加以扩展进而构建用户偏好模型[13], 较好地解决空间数据各属性特征的表达问题, 不过该方法仍存在空间范围定位不够准确、效用度估计存在偏差、特征值分布过于集中、模型缺少完整的动态化机制等不足[9-11]。为此, 本文作出以下扩展: ①为便于存储和计算, 将不规则子区间进一步分割为基本区间单元, 用于记录频谱、空间分辨率和时间偏好特征; ②引入区域数以准确表达空间范围, 将最小区域范围分割为基本区域单元, 用于记录空间范围偏好特征; ③将

三元组模型扩展为四元组模型, 用以完整记录用户偏好; ④引入兴趣度密度、兴趣度的概念和相应算法, 以便均衡、合理地反映目标区间(区域)范围内用户各次检索的贡献; ⑤增加用户信息反馈、基于权值衰减函数的权值衰减因子等动态化机制, 使模型完全动态化。试验表明, 本文的模型能够随着用户兴趣的转移更为及时、准确地自行调整。

2　模型动态化概述

目前, 一些非空间信息智能分发系统已经考虑到模型的动态化因素, 其原理是通过引入权值衰减函数, 使不同时段的访问信息在表达用户偏好的过程中被赋予不同的权重[14-15]。斯坦福大学的 FAB 自适应文档推荐系统是非空间信息智能分发系统的典型代表, 它通过引入一个简单的权值衰减函数 $h(t)=0.97^t$ 对不同时段用户的信息获取进行加权, 即系统每天晚上均将用户全部特征乘以一个衰减系数 0.97, 从而实现"古老"信息与"最新"信息相比权重较低[16]。

模型的动态化因素没有得到充分考虑是当前空间信息智能分发研究共有的局限。现有的智能分发系统主要通过不断更新用户检索记录, 以建立和修正用户偏好模型的方式使模型动态化。由于用户检索记录的时间跨度往往较大, 同时用户的偏好特征也会随着时间和需要的变化出现一定波动, 因而, 即便对于同一用户, 他在不同时段的检索记录对于其偏好特征的表达也会有不同贡献, 应赋予不同权值。另外, 模型缺少完善的用户信息反馈机制也是现有模型动态化的薄弱环节。仅根据用户的检索记录生成的偏好模型虽然能在一定程度上反映用户的偏好特征, 但这种偏好模型却难以及时反映用户偏好特征的变化。引入用户信息反馈机制能够弥补这一缺陷。由于用户在获得分发结果之后, 会根据自己的判断选择一些较理想的结果, 打开或者下载其中的数据, 而这一行为反映了用户的兴趣偏好[17]。

综上所述, 空间信息智能分发的动态化模型主要包括三个方面的动态化分量: ①通过隐式或显式地获取用户检索记录, 不断地对用户模型进行修正, 这是现有模型都已实现的模型动态化分量; ②是通过引入权值衰减函数, 对不同时段的用户行为进行加权求和的模型动态化分量; ③通过用户信息反馈引入的模型动态化分量。第一种分量已蕴含在用户的每一次检索记录中, 本文重点讨论后两种情况。为便于讨论, ①中所建模型称为静态模型, 考虑②、③因素的模型称为动态模型。

3　用户偏好模型框架

3.1　模型结构

用户模型采用四元组形式

$$C=\{X, W, R, V\} \tag{1}$$

式中, $X=\{x_1, \cdots, x_i, \cdots, x_s\}$; $W=\{w_1, \cdots, w_i, \cdots, w_s\}$; $R=\{\boldsymbol{R}_1, \cdots, \boldsymbol{R}_i, \cdots, \boldsymbol{R}_s\}$; $V=\{\boldsymbol{V}_1, \cdots, \boldsymbol{V}_i, \cdots, \boldsymbol{V}_s\}$。$x_i$ 依次为空间范围、频谱范围、时间范围、空间分辨率范围等元素项, s 为元素项个数(注: 仅考虑具有连续变化范围的元素项, 传统类型的解决方案见文献[9—11]), w_i 为 x_i 的权值。$\boldsymbol{R}_i$ 为 x_i 的分布范围和步长, 分两种情况: 对空间范围, $\boldsymbol{R}_i=[X_{\min}\ X_{\max}\ X_{\text{step}}\ Y_{\min}\ Y_{\max}\ Y_{\text{step}}]$, $i=1$, 其中, $X_{\min}$、$X_{\max}$、X_{step}、$Y_{\min}$、$Y_{\max}$、Y_{step} 分别为检索区域经纬度分量的分布范围的最小值、最大值、步长; ②对频谱范围、时间范围、空间分辨率范围等区间类型, $\boldsymbol{R}_i=[X_{\min}\ X_{\max}\ X_{\text{step}}]$, $i=2, 3, 4$, 其中, $X_{\min}$、$X_{\max}$、X_{step} 分别为分布范围的最小值、最大值、步长。$\boldsymbol{V}_i$ 为反映 x_i 分布特征的数值矩阵或向量, 分三种情况: ①对于空间范围, $\boldsymbol{V}_i=\{\rho_{g,k}\mid g=1, 2, \cdots, m, k=1, 2, \cdots, t\}$, $i=1$, 是数值矩阵, $\rho_{g,k}$ 为空间范围内对应基本区域单元上的分布密度值, m、t 分别为经、纬度方向基本区域单元的个数; ②对于频谱范围, $V_i=\{\rho_g\mid g=1, 2, \cdots, m\}$, $i=2$ 是数值向量, ρ_g 为频谱范围内对应基本区间单元上的分布密度值, m 为频谱范围内的基本区间单元个数; ③对于时间和空间分辨率, $\boldsymbol{V}_i=\{v_{i,g}\mid g=1, 2, \cdots, m_i\}$, $i=3, 4$, 是数值向量, $v_{i,g}$ 为时间、空间分辨率分布范围内对应基本区间单元上出现的频率值, m_i 为相应元素项分布范围内基本区间单元的个数。

3.2　权值衰减函数

用户对空间信息的获取通常在一段时期内反复进行, 其信息的获取记录是时间的函数。假定用户的偏好特征在一定时间内相对稳定, 且随着时间的推移小幅波动, 则有理由认为, 对于一组不同时间内获取的数据, 获取的时间距现在愈近, 愈能反映用户当前的需要, 反之, 亦然。也就是说, 用户检索记录的权值是时间的函数, 时间距现在愈久, 权值愈小, 反之, 则愈大。若将以后的检索

记录也包括在内, 则权值衰减函数的曲线类似于图 1。图中, 权值曲线 $h(t)$ 是一支在当前时间 $(t=t_n)$ 取最大值, 在 t_n 两侧逐渐递减的单峰值曲线, $h(t)=0$ 及 $h(t)=h(t_n)$ 是其渐近线。假设以后 $(t>t_n)$ 的检索记录存在, 则可以认为, 包括过去和将来的所有检索记录的全体整体上反映了用户目前的偏好特征。事实上, 只能得到以前 $(t\leqslant t_n)$ 的记录, 因而, 权值曲线应该取 $t\leqslant t_n$ 时的左半支, 即认为到目前为止的所有检索记录的全体整体上反映用户的偏好特征(图 2)。

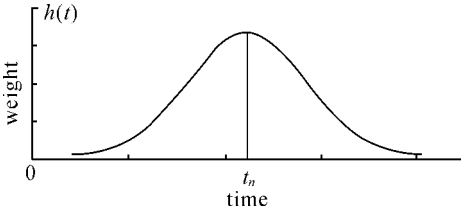

图 1 权值衰减函数曲线特征

Fig. 1 Characteristic of weight attenuation function curve

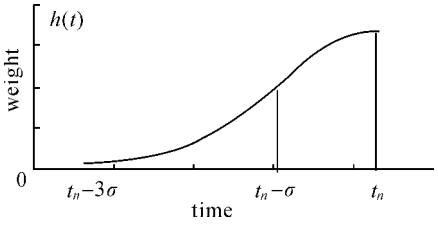

图 2 权值衰减函数曲线

Fig. 2 Weight attenuation function curve

正态分布的密度函数较好地符合了图 1 所示的权值曲线的特征。据此, 本文对其概念加以拓展, 用以定量描述用户检索记录的权值。如下式

$$h(t)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(t-t_n)^2}{2\sigma^2}\right)\quad \sigma>0,\ 0<t\leqslant t_n \tag{2}$$

式中, t_n 代表当前时间; σ 为衡量用户检索记录贡献的时间均方差。假设对于某一用户, 时间段 Δt 以前的检索记录对现在的偏好建模没有贡献, 根据 $h(t)$ 函数的特点, 可以取 $\sigma=\Delta t/3$, 这是因为 $[t_n-3\sigma, t_n]$ 区间内的权值积累达到了总权值的 99.74%, 此时可以忽略 $t<t-3\sigma$ 部分检索记录的贡献(图 2)。

3.3 权值向量

权值向量的确定采用文献[9—11]的方法, 但考虑模型动态化的影响, 步骤如下:

(1) 各元素项检索和反馈的频率统计

$$p_i=d_i/d_t\quad i=1,2,\cdots,s \tag{3}$$

式中, d_i 表示第 i 个元素项各检索和反馈记录的权值之和; d_t 表示全部检索和反馈记录的权值之和; s 为元素项的个数; p_i 表示第 i 个元素项基于权值衰减函数的检索和反馈频率。

(2) 检索和反馈频率归一化

$$p'_i=p_i/\sum_{i=1}^{s}p_i \tag{4}$$

式中, p'_i 表示第 i 个元素项的基于权值衰减函数的归一化的检索和反馈频率。

(3) 权值为归一化的检索和反馈频率

$$w_i=p'_i \tag{5}$$

由于空间范围包含经、纬度两个数据项的信息, 故步骤(2)中计算空间范围的归一化的检索和反馈频率时, p_i 应取经、纬度频率之和, 使计算出的空间范围的权值为经纬度权值之和。

3.4 基本区域(区间)单元

3.4.1 区域数

对区间数的概念加以拓展, 引入区域数用以表达空间范围这一二维区间, 它表示一片沿 X、Y 轴均有一定连续范围的矩形区域。形式如下: $[[a,b],[c,d]]$ 表示分别沿 X、Y 轴方向的区间 $[a,b]$ 和 $[c,d]$ 围成的矩形区域; $[[a,b),[c,d)]$ 表示分别沿 X、Y 轴方向的区间 $[a,b)$ 和 $[c,d)$ 围成的矩形区域; $[(a,b),(c,d)]$ 表示分别沿 X、Y 轴方向的区间 (a,b) 和 (c,d) 围成的矩形区域。其他情况依此类推。

3.4.2 基本区域单元

设用户的检索记录中共有 n 个关于经、纬度的样本值 $Z_k=[[z_k^{x-},z_k^{x+}],[z_k^{y-},z_k^{y+}]]$, $k=1,2,\cdots,n$, 其中, z_k^{x-}、z_k^{x+} 为样本值经度下、上限; z_k^{y-}、z_k^{y+} 为纬度下、上限。令 D 为包含 Z 的所有样本值的最小区域, 即

$$D=[[\min(z_k^{x-}),\max(z_k^{x+})],[\min(z_k^{y-}),\max(z_k^{y+})]]\quad k=1,2,\cdots,n \tag{6}$$

令 $X_{\min}=\min(z_k^{x-})$, $X_{\max}=\max(z_k^{x+})$, $Y_{\min}=\min(z_k^{y-})$, $Y_{\max}=\max(z_k^{y+})$, 且变量 $X_{\text{step}}>0$, $Y_{\text{step}}>0$, 设区域单元

$$U_{i,j}=[[X_{\min}+(i-1)\times X_{\text{step}},X_{\min}+i\times X_{\text{step}}],[Y_{\min}+(j-1)\times Y_{\text{step}},Y_{\min}+j\times Y_{\text{step}}]]$$
$$i=1,2,\cdots,m,\ j=1,2,\cdots,t \tag{7}$$

式中, m、t 分别为 D 沿 X、Y 方向分割的区域单元

的个数。则以下条件恒成立: ① $X_{min} + m \times X_{step} = X_{max}$, $Y_{min} + t \times Y_{step} = Y_{max}$; ② 对于任一样本值 Z_k 的经度的上下限 ξ_g、纬度的上下限 η_k, 均存在唯一的 i、j, 满足 $X_{min} + i \times X_{step} = \xi_g$, $Y_{min} + j \times Y_{step} = \eta_t$。则当 X_{step}、Y_{step} 均取最大值时, 称 $U_{i,j}$ 为空间范围 D 上的基本区域单元。

3.4.3 基本区间单元

设用户的检索记录中共有 n 个区间类型样本值, $Z_k = [z_k^-, z_k^+]$, $k = 1, 2, \cdots, n$, 其中 Z_k^-、Z_k^+ 为样本值区间数的下、上限。令 I 为包含 Z 的所有样本值的最小区间, 即

$$I = [\min(z_k^-), \max(z_k^+)] \quad k = 1, 2, \cdots, n \tag{8}$$

令 $X_{min} = \min(z_k^-)$, $X_{max} = \max(z_k^+)$, 且变量 $X_{step} > 0$, 设区间单元

$$I_i = [X_{min} + (i-1) \times X_{step}, X_{min} + i \times X_{step}]$$
$$i = 1, 2, \cdots, m \tag{9}$$

式中, m 为 I 分割为区间单元的个数。则以下条件恒成立: ① $X_{min} + m \times X_{step} = X_{max}$; ② 对于任一样本值 Z_k 的上下限 ξ_g, 均存在唯一的 i, 满足 $X_{min} + i \times X_{step} = \xi_g$。则当 X_{step} 取最大值时, 称 I_i 为区间范围 I 上的基本区间单元。

4 基于权值衰减函数的模型动态化

4.1 频谱范围元素项的分布特征

频谱范围元素项用区间数表达, 采用文献[9, 18—19]的符号数据分析法, 用户的每一个检索样本 $Z_k = [z_k^-, z_k^+]$, $k \in E = \{1, 2, \cdots, n\}$ 都代表了用户在 t_k 时刻的一次检索意图, 相对于当前时刻 t_n 而言, 可以认为这些样本的权值为 $h(t_k)$。样本区间长度愈短, 用户的检索目标愈集中, 单位区间长度上凝聚用户愈多的检索意图; 反之, 亦然。因此, 若令 $h = \sum_{k=1}^{n} h(t_k)$ 代表所有频谱范围样本的权值之和, 则元素项的分布特征可由经验密度函数式(10)表达。其中, 对应每一个 ξ_j 的函数值 $\rho_z(\xi_j)$, 都代表区间 $[\xi_j, \xi_j + \Delta\xi_j]$, $\Delta\xi \to 0$ 上用户对相应信息的关注程度

$$\rho_z(\xi_j) = \frac{1}{h} \sum_{k \in E} \frac{I_k(\xi_j) h(t_k)}{\| Z_k \|} \tag{10}$$

式中, $I_k(\cdot)$ 是示性函数, 表示 ξ 是否存在于 Z_k 中, ξ 为频谱值; $\| \cdot \|$ 表示区间宽度。式(11)表达用户对区间 I_x 上信息的关注程度

$$\psi_z(I_x) = p\{x \in I_x\} = \frac{1}{h} \sum_{k \in E} \frac{\| Z_k \cap I_x \| h(t_k)}{\| Z_k \|} \tag{11}$$

据此, 引入兴趣度、兴趣度密度的概念, 用于所述关注程度的数学表达, 定义如下。

兴趣度密度: 设区间变量 Z 的 n 个观测样本 $Z_k = [z_k^-, z_k^+]$, $k \in E = \{1, 2, \cdots, n\}$ 反映用户的信息偏好, 且各样本区间 Z_k 上用户的关注程度均服从均匀分布, 若各样本反映用户偏好的权重由权值衰减函数式(2)确定, 则由式(10)所定义的 $\rho_z(\xi)$ 函数即为度量用户信息偏好的兴趣度密度函数, 根据该函数求得的函数值, 即为相应区间位置的兴趣度密度。

兴趣度: 根据兴趣度密度式(10)的定义, 由式(11)定义的函数即为兴趣度函数, 在某一给定区间 I_x 上, 由该函数求得的函数值即表达了用户对相应区间信息的偏好程度, 定义为兴趣度。

4.2 空间范围元素项的分布特征

4.2.1 分布特征

空间范围元素项用区域数表示, 用户的每一个空间范围检索样本 $Z_k = [[z_k^{x-}, z_k^{x+}], [z_k^{y-}, z_k^{y+}]]$, $k \in E = \{1, 2, \cdots, n\}$, 都代表了用户的一次检索意图, 相对于当前时刻 t_n 而言, 可以认为这些样本值的权值为 $h(t_k)$。样本区域面积愈小时, 用户的检索目标愈集中, 单位区域面积上凝聚用户愈多的检索意图; 反之, 亦然。因此, 若令 $h = \sum_{k=1}^{n} h(t_k)$ 代表所有空间范围样本观测值的权值之和, 则元素项的分布特征可由经验密度函数式(12)表达。其中, 对应每一个 (ξ_i, η_j) 的函数值 $\rho_z(\xi_i, \eta_j)$, 都代表区域 $[\xi_i, \xi_i + \Delta\xi_i, [\eta_j, \eta_j + \Delta\eta_j]]$, $\Delta\xi \to 0$, $\Delta\eta \to 0$ 上用户对相应信息的关注程度

$$\rho_z(\xi, \eta) = \frac{1}{h} \sum_{k \in E} \frac{I_k(\xi, \eta) h(t_k)}{\| Z_k \|} \tag{12}$$

式中, $I_k(\cdot)$ 是示性函数, 表示 (ξ, η) 是否存在于 Z_k 中, (ξ, η) 为空间范围中的某一点; $\| \cdot \|$ 表示区域面积。式(13)表达用户对区域 $D_{x,y}$ 上信息的关注程度

$$\psi_z(D_{x,y}) = \frac{1}{h} \sum_{k \in E} \frac{\| Z_k \cap D_{x,y} \| h(t_k)}{\| Z_k \|} \tag{13}$$

则表达相应关注程度的兴趣度、兴趣度密度定义如下。

兴趣度密度: 设区域变量 Z 的 n 个样本观测值 $Z_k = [[z_k^{x-}, z_k^{x+}], [z_k^{y-}, z_k^{y+}]]$, $k \in E = \{1, 2, \cdots, n\}$ 反映用户的信息偏好, 且各样本区域 Z_k 上用户的关注程度服从均匀分布, 若各样本反映用户偏好的权重由权值衰减函数式(2)确定, 则由式

(12)所定义的 $\rho(\xi,\eta)$ 函数即为度量用户信息偏好的兴趣度密度函数,根据该函数求得的函数值,即为相应区域位置的兴趣度密度。

兴趣度:根据兴趣度密度式(12)的定义,由式(13)定义的函数即为兴趣度函数,在某一给定区域 $D_{x,y}$ 上,由该函数求得的函数值即表达了用户对相应区域信息的偏好程度,定义为兴趣度。

4.2.2 兴趣度的分解

由式(13)计算的兴趣度包含了经、纬度两个元数据项的信息,需将其沿经、纬度方向进行分解。兴趣度的取值同时受用户偏好模型和待分发数据空间范围的影响,情况较为复杂,很难精确量化,但可以基于以下假设求其近似值:①通常在检索次数足够多的情况下,偏好模型的空间范围因素在经、纬度方向的分量分布特征应相对稳定,兴趣度基本上反映目标区间上的用户偏好程度;②若将空间范围分解为两个独立的沿经、纬度方向的区间变量,则借鉴4.1节式(10)、式(11)的方法,可以计算经、纬度区间变量上的兴趣度 $\Phi_z^0(D_x)$、$\Phi_z^0(D_y)$,由于本模型中此处经纬度的兴趣度通过对目标区间内的兴趣度密度积分求得,它反映用户偏好的分布特征,故 $\Phi_z^0(D_x)$、$\Phi_z^0(D_y)$ 可近似反映空间范围内经、纬度方向兴趣度分量的相对关系;③空间范围的兴趣度沿 X、Y 方向的分量之间的比值可近似由 $\Phi_z^0(D_x)$、$\Phi_z^0(D_y)$ 之间的比值表达。据此,空间范围的兴趣度 $\Phi_z(D_{x,y})$ 在经、纬度方向的分量近似为

$$\Phi_z(D_x)=\sqrt{\Phi_z(D_{x,y})/r}=\sqrt{\frac{1}{n\times r}\sum_{k\in E}\frac{\|Z_k\cap D_{x,y}\|}{\|Z_k\|}} \tag{14}$$

$$\Phi_z(D_y)=\sqrt{\Phi_z(D_{x,y})\times r}=\sqrt{\frac{r}{n}\sum_{k\in E}\frac{\|Z_k\cap D_{x,y}\|}{\|Z_k\|}} \tag{15}$$

式中,$r=\Phi_z^0(D_y)/\Phi_z^0(D_x)$,$\Phi_z(D_x)$、$\Phi_z(D_y)$ 分别为空间范围兴趣度在经、纬度方向的分量。对于 $\Phi_z^0(D_x)$ 或 $\Phi_z^0(D_y)$ 为0的情况,因为此时 $\Phi_z(D_{x,y})$ 必为0,此时可直接令 $\Phi_z(D_x)=0$,且 $\Phi_z(D_y)=0$。尽管在分解的过程中 $\Phi_z(D_x)$、$\Phi_z(D_y)$ 会出现此消彼长的部分误差,但两者的加权和在一定程度上减小这种影响。

4.3 时间、空间分辨率元素项的分布特征

和频谱范围一样,在根据时间和空间分辨率进行数据检索的过程中,用户也常用区间数来表达需求范围,但元数据表达上有所不同:①时间元数据虽然也表现为一个区间范围,但由于遥感成像几乎是瞬时完成的,故在进行效用度计算时,时间更适合作为一个点来处理;②空间分辨率则直接表现为一个或几个离散点。有些遥感数据一景影像中的各个波段空间分辨率是一致的,也有些波段较多的影像中,一景影像存在着几个不同的分辨率。据此,在建立偏好模型时,可用区间数来表达用户检索中时间和空间分辨率的偏好情况,而在进行效用度估计时,则应当作为一个或几个离散点来处理。

对于时间和空间分辨率,采用文献[9,18—19]的符号数据分析法。用户的每一个检索样本值 $Z_k=[z_k^-,z_k^+]$, $k\in E=\{1,2,\cdots,n\}$ 都代表了用户在 t_k 时刻的一次检索意图,相对于当前时刻 t_n 而言,可以认为这些样本值的权值为 $h(t_k)$。若令 $h=\sum_{k=1}^{n}h(t_k)$ 代表所有时间范围或空间分辨率范围样本观测值的权值之和,并采用 v_g(各基本区间单元的基于权值衰减函数的加权频率)表达各基本区间单元的分布情况,以此表达用户的偏好特征,则对于给定遥感数据的元数据项,其成像时间和空间分辨率所对应的用户偏好模型中相应元素项的值,客观上反映了用户对相应数据的关注程度,即兴趣度。则兴趣度函数为

$$\Phi_z(x)=v_g=\frac{1}{h}\sum_{k\in E}I_k(Z_k\cap I_g)h(t_k)\quad \text{iff}\ x\in I_g \tag{16}$$

式中,x 为时间或空间分辨率元素项的属性值,且 x 在基本区间单元 I_g 上。对于一景影像存在数个空间分辨率的情况,$\Phi_z(x)$ 取分辨率对应的最大频率值。

5 空间信息的分发决策

5.1 构造决策矩阵

采用文献[9—11]的方法,从决策理论角度出发,将待分发信息集作为方案集,其决策矩阵如表1。

表1 决策矩阵

Tab.1 Decision matrix

	x_1	x_2	…	x_s
S_1	y_{11}	y_{12}	…	y_{1s}
S_2	y_{21}	y_{22}	…	y_{2s}
⋮	⋮	⋮		⋮
S_t	y_{t1}	y_{t2}	…	y_{ts}

其中,S_j 为备选方案;x_i 为元素项;y_{ji} 为 S_j 相应

于 x_i 的属性值。对于空间范围, γ_{ji} 为区域数 $[[\gamma_{ji}^{x-}, \gamma_{ji}^{x+}], [\gamma_{ji}^{y-}, \gamma_{ji}^{y+}]]$, 其中, γ_{ji}^{x-}、γ_{ji}^{x+} 为备选方案的经度下、上限, γ_{ji}^{y-}、γ_{ji}^{y+} 为纬度下、上限。对于频谱范围, γ_{ji} 为区间数 $[\gamma_{ji}^{-}, \gamma_{ji}^{+}]$。对于时间或单一空间分辨率, γ_{ji} 为一数值, 对多分辨率数据, γ_{ji} 为空间分辨率的集合。

5.2　计算备选方案的效用度

根据式(11)、(14)、(15)和(16), 分别计算频谱范围、空间范围、时间和空间分辨率的兴趣度。

备选方案 S_j 的效用度 μ_j 为各个元素项的兴趣度的加权和, 由下式计算

$$\mu_j = \sum_{i=1}^{s}\left(w_i \times \phi_i\right) \tag{17}$$

式中, w_i 为元素项 x_i 的权值; ϕ_i 为 x_i 的兴趣度。

5.3　备选方案的分发

在获取效用度之后, 便可以根据效用度的大小对信息进行分发。具体操作上分推送模式和拉取模式两种情况, 其思路略有不同[16]。

(1) 推送模式。可以根据用户设定的阈值 T_r 进行分发, 即当 $\mu_j \geqslant T_r$ 时, 将数据分发给相应的用户, 否则不分发。或者, 首先对最近的某一个时期内用户检索的数据进行效用度计算, 找出其中最小的效用度, 以此作为默认阈值, 决定是否分发。

(2) 拉取模式。首先, 根据用户的检索条件, 获取满足要求的方案集。比如, 对于区间数或区域数类型的元素项, 可以检索与用户输入的区间数或区域数相交的备选方案, 而对于点(或点集)类型的元素项, 可以检索出落入检索区间内的备选方案。然后, 计算上一步结果集各方案的效用度, 根据效用度大小排序, 并以此作为数据推荐的优先次序。

6　用户反馈引入的模型动态化

检索记录并不能完整地反映用户的真实需求, 同时由于检索记录相对较少, 这导致所生成的用户模型较为粗糙和相对滞后, 难以准确、及时地反映用户的偏好特征及其变化。通过收集用户的信息反馈能够弥补这一不足。本文通过记录用户打开、下载的数据的特征, 并据此对原有模型进行修正, 从而使模型能够根据用户兴趣的转移及时做出调整。鉴于用户检索和信息反馈是一种反复迭代的动态过程, 对两者一并考虑, 算法如下:

(1) 在信息分发过程中, 记录用户每次打开、下载数据的各元素项的值。

(2) 将每次的记录结果反馈给系统, 修正原有模型, 步骤如下。

对于空间范围。首先, 根据式(12)计算用户检索数据的偏好模型。然后, 根据式(12)并结合上一步的计算结果, 计算用户反馈数据的空间范围对用户偏好模型的贡献(注: 由于用户反馈数据的空间范围有可能不是规则的矩形, 在此情况下, 反馈数据的空间范围并不参与基本区域单元的分割, 而只参与计算偏好模型的兴趣度密度值)。反馈数据的空间范围和此前各次检索的空间范围的全体的兴趣度密度函数, 作为空间范围元素项修正后的用户偏好模型。此时, 兴趣度密度公式为

$$\rho_2(\xi,\eta) = \frac{1}{h'}\sum_{k\in E'}\frac{I'_k(\xi,\eta)\,h'(t_k)}{\|Z'_k\|} \tag{18}$$

式中, E' 为用户检索记录和反馈数据空间范围样本的集合; h' 为 E' 中各样本时间权值之和; Z'_k, $k\in E'$ 为检索或反馈的空间范围样本; $h'(t_k)$ 为 t_k 时刻样本的权值; $I'_k(\cdot)$ 为示性函数, 表示 (ξ,η) 是否在 Z'_k 中, (ξ,η) 为空间范围中的某一点, $\|\cdot\|$ 表示区域面积。考虑用户反馈后的兴趣度计算方法类似于式(13), 其原理为对给定的空间范围目标区域内的兴趣度密度(见式(18))进行积分求和。兴趣度沿 X、Y 方向的分量类似于式(14)和式(15), 分解原理同 4.2 节。

对于频谱范围。反馈的频谱范围采用与原模型用户检索的频谱范围相同的方式参与基本区间单元的分割和兴趣度密度的计算。此时, 兴趣度密度公式为

$$\rho_2(\xi) = \frac{1}{h'}\sum_{k\in E'}\frac{I'_k(\xi)\,h'(t_k)}{\|Z'_k\|} \tag{19}$$

式中, E' 为用户检索记录和反馈数据频谱范围样本的集合; h' 为 E' 中各样本的权值之和; Z'_k, $k\in E'$ 为检索或反馈的频谱范围样本; $h'(t_k)$ 为 t_k 时刻样本的权值; $I'_k(\cdot)$ 为示性函数, 表示 ξ 是否在 Z'_k 中, ξ 为频谱值, $\|\cdot\|$ 表示区间宽度。考虑用户反馈后的兴趣度计算方法类似于式(11), 其原理为对给定的空间范围目标区间内的兴趣度密度(见式(19))进行积分求和。

对于时间。首先, 根据式(16)计算用户检索数据的偏好模型。然后, 根据式(20)计算用户反馈数据的时间属性对用户偏好模型的贡献(由于用户反馈数据的时间属性为点值, 故反馈数据的时间值并不参与基本区间单元的分割, 而只参与模型修正)。

$$\nu'_g = \frac{1}{h'}\sum_{k\in E'} I'_k(T_k \cap I_g)\,h'(t_k) \qquad (20)$$

式中, E' 为用户反馈的时间样本集合; I_g 是基本区间单元; h' 为时间元素项的反馈数据和原模型中检索记录的权值之和; T_k, $k \in E'$, 为第 k 次反馈数据的生成时间; $h'(t_k)$ 为 t_k 时刻样本的权值; $I'_k(\cdot)$ 为示性函数, 表示 T_k 是否在 I_g 中。式(16)与式(20)之和即为修正后的时间元素项的偏好模型, 如下

$$\nu''_g = \frac{1}{h'}\Big[\sum_{k\in E} I_k(Z_k \cap I_g)h(t_k) + \sum_{k\in E'} I'_k(T_k \cap I_g)\,h'(t_k)\Big] \quad \text{iff } x \in I_g \qquad (21)$$

式中, h' 为时间元素项的反馈数据和原模型中检索记录的权值之和, 其他参数含义同上。取 $\phi_2(x) = \nu''_g$ 为考虑用户反馈后的时间元素项的兴趣度, 它体现了用户对目标时间点的空间信息的关注程度。

对于空间分辨率。其偏好模型的计算方法类似于时间, 区别在于若空间分辨率不唯一时, 需要遍历计算所有分辨率值的贡献。

(3) 根据修正后的模型, 计算备选方案的效用度, 根据效用度的大小对方案进行优劣排序, 并据此进行下一次的信息分发。

(4) 重复执行以上步骤。

7 实例分析

试验数据源于文献[9]: ① 根据文献[9]中4.5.2.4节的用户检索数据生成用户模型, 为了生成动态模型, 对原始数据增添了检索时间; ② 选取文献[9]中5.2.2.3节方案1~8的数据作为备选方案。

根据数据①, 分别建立静态用户模型和动态用户模型。两种模型的元素项的权值见表2。根据两种模型, 分别计算②中备选方案的效用度: 根据静态模型算得的备选方案的效用度见表3; 根据动态模型算得的备选方案的效用度见表4; 文献[9]中备选方案的效用度见表5。

表 2 用户偏好模型元素项的权值

Tab. 2 Element weights of user profile model

	空间范围		频谱范围	时间	空间分辨率
	经度	纬度			
静态模型	0.229 2	0.229 2	0.145 8	0.208 3	0.187 5
动态模型	0.230 4	0.230 4	0.129 3	0.196 8	0.213 1

表 3 基于静态模型的待分发方案的效用度及各元素项的兴趣度

Tab. 3 Utility degrees of items to be distributed and interest degrees of each element based on static model

方案	兴趣度					效用度
	经度	纬度	频谱范围	时间	空间分辨率	
1	0.45	0.44	0.69	1.00	0.33	0.58
2	0.24	0.14	0.69	1.00	0.33	0.46
3	0.12	0.04	0.69	1.00	0.33	0.41
4	0.36	0.22	0.68	0.60	0.67	0.48
5	0.04	0.27	0.68	1.00	0.67	0.50
6	0.05	0.21	0.68	0.20	0.67	0.32
7	0.00	0.00	0.68	0.10	0.67	0.24
8	0.00	0.00	0.68	0.10	0.67	0.24

表 4 基于动态化模型的待分发方案的效用度及各元素项的兴趣度

Tab. 4 Utility degrees of items to be distributed and interest degrees of each element based on dynamic model

方案	兴趣度					效用度
	经度	纬度	频谱范围	时间	空间分辨率	
1	0.50	0.55	0.68	1.00	0.35	0.60
2	0.26	0.08	0.68	1.00	0.35	0.44
3	0.09	0.00	0.68	1.00	0.35	0.38
4	0.40	0.26	0.66	0.84	0.59	0.53
5	0.00	0.33	0.66	1.00	0.59	0.48
6	0.00	0.21	0.66	0.18	0.59	0.30
7	0.00	0.00	0.66	0.00	0.59	0.21
8	0.00	0.00	0.66	0.19	0.59	0.25

表 5 文献[9]中方案的效用度

Tab. 5 Utility degrees of items in the literature [9]

方案	1	2	3	4	5	6	7	8
效用度	0.96	0.90	0.84	0.93	0.96	0.87	0.40	0.65

表2显示, 两种算法所生成的权值有一定的差异, 这是由于权值衰减函数引入前后, 模型的动态化因素对权值有着不同程度影响, 后者中模型的时效性得到了进一步加强。显然, 相较于前者, 动态模型更能反映用户当前的信息偏好。

表3和表5均基于静态模型算得。对比两表可以看出, 两种方法中效用度计算结果相差较大, 但总体趋势相近, 原因如下三种: 文献[9]以特征值代替频率值作为计算效用度的依据, 人为地增大了取值较小的各元素项的相似度值; 文献[9]以与频谱范围的目标区间相交的各不规则子区间的特征值中的最大值作为相似度, 而不考虑同时与其相交的其他子区间的影响, 这也在一定程度上

增大了效用度的取值; 经、纬度范围在存在第二种误差的同时, 也存在一些不相关的纬、经度样本值的影响。因为, 一些在二维空间上根本不相交的空间范围, 其在经、纬度上的分量却可能是相交的, 这会对效用度的计算产生影响。从表 3 可以看出, 空间范围的兴趣度普遍较小, 这是因为这些待分发的数据在二维经纬度空间上与用户检索频繁的空间位置重叠较少。

对比表 3、表 4 可以发现, 相较于前者, 后者的兴趣度和效用度值大多存在不同程度的变化, 变化方向也不一致, 这是由不同原因造成的。权值衰减函数的引入对发生在不同时段的用户检索和反馈信息的建模贡献进行了不同程度的拉伸或抑制, 因此, 相对于文献[9] 和本文的静态模型而言, 引入了动态化因素的建模算法, 使得模型更能体现用户近期的行为特征。较于表 3、表 4 中元素项的兴趣度存在以下特征: ① 频谱范围兴趣度均有一定程度的微幅减小, 这主要是因为用于用户偏好建模的频谱范围检索行为主要集中在较早时段, 而在近期较少发生(用户检索时该元素项缺省) , 这导致频谱范围元素项的兴趣度密度经验函数取值整体偏低; ② 空间分辨率的兴趣度则同时存在小幅度的增大和减小两种情况, 这是因为, 相较于用户检索行为的发生时间而言, 用于偏好建模的空间分辨率范围检索数据的各区间分布较为均匀, 在根据权值衰减函数进行加权建模时, 虽然部分检索行为的建模贡献被抑制, 但另外一部分却得到了拉伸, 这使得在各备选方案的分布位置的兴趣度经验密度值波动不大, 但同时也存在一定的此消彼长; ③ 时间元素项的兴趣度同时存在增大、减小和不变三种情况, 其中, 方案 1、2、3 和 5 兴趣度大小不变, 这是因为这些方案的值均介于用户检索的时间范围建模数据的分布区间之内, 基于权值衰减函数的加权建模对兴趣度取值无影响, 方案 4、6、7 和 8 均位于建模数据分布区间两端, 且仅受部分检索记录不同程度、不同方向的影响, 故同时存在增大和减小的情况; ④ 空间范围元素项中的经度和纬度指标的分布存在较多情况, 方案 7 和 8 在表 3 和表 4 中均取值为 0, 这是因为两方案的空间范围取值与用户所有的检索建模数据分布范围均不相交, 这种情况在文献[9—11] 的算法中无法得到体现。方案 1~ 6 则均同时存在增大和减小的情况, 原因类似于 ② 和 ③ 中分析的情况, 区别在于空间范围的经度和纬度之间存在着相关性, 两者是以区域数(基本区域单元) 的形式作为整体参与运算的; ⑤ 在 ①~ ④ 中各因素的综合作用下, 各方案的效用度均出现不同程度增减。

表 3、表 4 的对比结果表明, 各动态化因素对于各元素项的兴趣度以及最终的效用度都有比较明显的影响, 该影响基本上能够更为真实地反映用户当前的偏好特征。

8　结束语

现有的空间信息智能分发的理论研究和系统实现存在着空间范围定位不够准确、效用度计算存在偏差等缺陷, 且大多仍停留在非完全动态化的层面, 它们往往只关注用户的检索行为对偏好建模的贡献, 很少注意到用户反馈的作用, 且均未考虑时间因素的影响, 从而导致用户偏好模型难以准确、及时地反映用户兴趣特征的变化。鉴于此, 本文对现有文献的理论和算法进行扩展, 通过引入区域数、兴趣度、兴趣度密度等概念和算法, 并引入权值衰减函数和用户信息反馈等动态化因素, 对以上问题予以解决。

试验表明, 相较于静态模型, 本文模型能够更为有效地反映用户偏好特征的变化。本文算法为空间信息智能分发的用户建模提供了一个可行的解决方案。

参考文献:

[1] WANG Zegen, HUA Yixin. Research on Technology of Active Spatial Information Service[J]. Acta Geodaetica et Cartographica Sinica, 2006, 35(4): 379-389. (王泽根, 华一新. 主动空间信息服务技术研究[J]. 测绘学报, 2006, 35(4): 379-389.)

[2] DOUGLASS R J, MORK J, SURESH R. Battlefield Awareness and Data Dissemination (BADD) for the Warfighter[C] // Proceedings of Digitization of the Battlefield II. Orlando: SPIE, 1997: 18-24.

[3] STEPHENSON T P, DECLEENE B T, SPECKERT G, et al. BADD Phase II: DDS Information Management Architecture[C] // Proceedings of Digitization of the Battlefield II. Orlando: SPIE, 1997: 49-58.

[4] WU Wei. Development Assumption on Chinese Army's New Generation Communication Network[J]. Journal of CAEIT, 2007, 2(5): 445-449, 463. (吴巍. 我军新一代通信网络发展设想[J]. 中国电子科学研究院学报, 2007, 2(5): 445-449, 463.)

[5] DAWIDOWICZ E. Performance Evaluation of Network Centric Warfare Oriented Intelligent Systems[C] //

Proceedings of the Second International Workshop on Performance and Intelligence of Intelligent Systems. Mexico: NIST, 2001: 73-79.

[6] DAWIDOWICZ E, RODRIGUEZ A, LANGSTON J. Intelligent Nodes in Knowledge Centric Warfare[C]// Proceedings of the 7th International Command and Control Research and Technology Symposium. Monterey: [s. n.], 2002.

[7] DAWIDOWICZ E, JACKSON V. The Right Information and Intelligent Nodes[C]// Proceedings of 8th International Command and Control Research and Technology Symposium. Washington: [s. n.], 2003.

[8] GLOBALSECURITY. WIN-TCapabilities[EB/OL]. [2011-6-18]. http://www.globalsecurity.org/military/systems/ground/win-t-cap.htm.

[9] XIA Yu. The User Profile Model for Intelligent Delivery of Spatial Information[D]. Wuhan: Wuhan University, 2009. (夏宇. 面向空间信息智能分发的用户偏好模型研究[D]. 武汉: 武汉大学, 2009.)

[10] XIA Yu, ZHU Xinyan, LI Deren, et al. A User Profile Model for Intelligent Delivery of Spatial Information[C]// Proceedings of Geoinformatics 2008 and Joint Conference on GIS and Built Environment. Guangzhou: SPIE, 2008.

[11] XIA Yu, ZHU Xinyan, ZHANG Chunlin, et al. Towards Intelligent Spatial Information Dissemination Based on User Profile Model[C]// Proceedings of International Conference on Earth Observation Data Processing and Analysis. Wuhan: SPIE, 2008.

[12] MOORE R E. Methods and Applications of Interval Analysis[M]. Philadelphia: Society for Industrial and Applied Mathematics, 1979.

[13] HWANG C, YOON K. Multiple Attributes Decision Making: Methods and Applications[M]. Berlin: Springer-Verlag, 1981.

[14] ASNICAR F A, TASSO C. ifWeb: a Prototype of User Models Based Intelligent Agent for Document Filtering and Navigation in the World Wide Web[C]// Proceedings of 6th International Conference on User Modeling, Sardinia: [s. n.], 1997.

[15] ZHANG Bingqi. The Representation, Acquisition and Inference of Personalized Requirements: A Case Study[D]. Beijing: Graduate University of Chinese Academy of Sciences, 2005. (张丙奇. 个性化需求的描述、获取与推断—案例研究[D]. 北京: 中国科学院研究生院, 2005.)

[16] YAN Duanwu, WANG Yuefen. Information Acquisition and User Service[M]. Beijing: Science Press, 2010. (颜端武, 王曰芬. 信息获取与用户服务[M]. 北京: 科学出版社, 2010.)

[17] ZANG Cheng. Research on Key Techniques of Privacy Preservation in Personalized Search[D]. Hangzhou: Zhejiang University, 2008. (臧铖. 个性化搜索中隐私保护的关键问题研究[D]. 杭州: 浙江大学, 2008.)

[18] BILLARD L, DIDAY E. Symbolic Data Analysis: Definitions and Examples[EB/OL]. [2011-6-18]. http://aaron.stat.uga.edu/people/faculty/BILLARD/tr_symbolic.pdf.

[19] BILLARD L. Symbolic Data Analysis, What is It?[C]// Proceedings of Computational Statistics: 17th Symposium. Rome: IASC, 2006: 261-269.

(责任编辑: 宋启凡)

收稿日期: 2010-12-20
修回日期: 2011-05-27

第一作者简介: 李新广(1977—), 男, 博士生, 主要从事空间信息服务、数据挖掘方面的研究。
First author: LI Xinguang(1977—), male, PhD candidate, majors in spatial information service, data mining.
E-mail: emailoflxg@163.com

(上接第 645 页)

and Globally Convergent Pose Estimation from Video Images[J]. IEEE Transactions on PAMI, 2000, 22(6): 610-622.

[12] SCHWEIGHOFER G, PINZ A. Fast and Globally Convergent Structure and Motion Estimation for General Camera Models[J]. Proc of 17th British Machine Vision Conference. 2006: 147-156.

[13] HORN K P. A Closed-form Solution of Absolution Orientation Using Unit Quaternion[J]. Journal of the Optical Society of America, 1987, 4(4): 629-642.

[14] KEN S. Animating Rotation with Quaternion Curves[J]. Siggraph, 1998, 19(3): 245-254.

[15] HORN K P. Closed-form Solution of Absolution Orientation Using Orthonormal Matrices[J]. Journal of the Optical Society of America: Series A, 1988, 5(7): 1127-1135.

[16] LUENBERGER D G. Linear and Nonlinear Programming[M]. 3rd ed. Berlin: Springer, 2008.

(责任编辑: 丛树平)

收稿日期: 2009-07-20
修回日期: 2009-10-16

第一作者简介: 龚辉(1982—), 男, 博士生, 研究方向为数字摄影测量和遥感图像处理。
First author: GONG Hui(1982—), male, PhD candidate, majors in digital photogrammetry and remote sensing image processing.
E-mail: gonghuip@163.com