

# 基于混合过滤的地学数据个性化推荐方法设计与实现

王 末<sup>1,2</sup>, 郑晓欢<sup>3</sup>, 王卷乐<sup>4,5,6</sup>, 柏永青<sup>4,5</sup>

(1. 中国农业科学院农业信息研究所, 北京 100081; 2. 农业部农业大数据重点实验室, 北京 100081;  
3. 中国科学院办公厅, 北京 100864; 4. 中国科学院地理科学与资源研究所, 资源与环境信息系统国家重点  
实验室, 北京 100101; 5. 中国科学院大学, 北京 100049; 6. 江苏省地理信息资源开发与利用协同创新中  
心, 南京 210023)

**摘要:** 推荐系统是帮助互联网用户克服信息过剩的有效工具。在地学数据共享领域, 较其他物品的内容属性, 地学数据具有更加丰富的时空属性, 这也给地学数据推荐带来挑战。针对地学数据的特点, 为地学数据共享推荐服务开发了一种动态加权的混合过滤方法。该方法分别采用协同过滤和基于内容过滤算法预测用户对数据的兴趣度, 再以训练模型计算最优加权权重, 计算最终预测评分。在数据获取阶段, 通过用户访问日志数据, 采用 Jenks Natural Break 算法分析用户访问记录获取用户的数据兴趣度。在基于内容过滤部分, 通过数据的空间、时间及内容属性计算数据相似度, 并以用户历史行为为依据计算用户兴趣。在协同过滤和基于内容过滤中分别采用 k-NN 算法计算用户对未访问数据的预测评分, 并进行加权求和。通过训练集, 对理想权重值及用户的共同评价度(co-rating level)进行建模, 拟合二者的关系。该模型被应用于混合过滤的权重调整, 以获得最优的加权方程。测试结果显示, 结合数据时空属性的混合过滤方法的准确度和召回率, 较单一的协同过滤或基于内容过滤方法有显著提高。

**关键词:** 地理空间数据; 推荐系统; 混合过滤; 科学数据共享

DOI: 10.11821/dlyj201804014

## 1 引言

数据是进行科学研究的基本条件<sup>[1]</sup>。当今, 地学领域每天以前所未有的速度产生、收集和储存了海量的科学数据。数据共享是有效利用这些数据重要的途径。资源查找是数据共享服务提供的基本功能之一。然而, 地学数据内在属性包括空间、时间和主题内容信息, 基于传统的检索技术可能不能满足用户对数据属性的需求。面对海量的数据, 科研人员将面临如何发现所需数据的难题。个性化推荐是解决这一信息过剩问题的有效途径。个性化推荐系统已在多个领域得到成功地应用, 包括多媒体内容(音乐、电影等)<sup>[2-4]</sup>、网络教学<sup>[5,6]</sup>、电子商务<sup>[7,8]</sup>、网络搜索<sup>[9,10]</sup>等。但在目前仍缺乏针对科学数据共享服务设计的个性化推荐方法。

个性化推荐系统是一种能够学习用户偏好, 并基于用户偏好预测用户需求, 在大量的可能选项里给出个性化推荐的 Web 应用系统<sup>[11]</sup>。常见的个性化推荐算法类型有协同过

收稿日期: 2017-10-11; 修订日期: 2018-02-01

基金项目: 国家科技基础条件平台建设项目(2005DKA32300); 中国科学院特色研究所培育建设服务项目(TSYJS03); 中国工程科技知识中心建设项目(CKCEST-2017-3-1); 农业科学数据挖掘分析平台研究与建设项目(JBYW-AII-2017-32); 中国农业科学院科技创新工程项目(CAAS-ASTIP-2016-AII)

作者简介: 王末(1987-), 男, 助理研究员, 研究方向为地学数据共享与挖掘。E-mail: wangm.13b@igsrr.ac.cn

通讯作者: 王卷乐(1976-), 男, 博士, 研究员, 主要研究方向为科学数据共享、地理信息系统与遥感应用。

E-mail: wangjl@igsrr.ac.cn

滤算法(collaborative filtering)、基于内容过滤算法(content-based filtering)以及人口统计学过滤算法(demographic filtering)。协同过滤依赖于用户间的共同评分来计算用户间相似度,并将用户喜好项目推荐给与其相似的用户;基于内容过滤则通过项目属性计算项目(item)间相似度,依据用户历史兴趣推荐具有相似属性的项目;人口统计学过滤则是通过用户的社会属性(比如年龄、性别、地域、职业等)来计算用户的相似性,划分用户类型,给出相应的推荐。这些推荐方法有着各自的优缺点,单一地使用某一种推荐算法并不能适应所有的应用场景。在有大量的用户评分数据情形下,协同过滤往往能获得比基于内容过滤更好的效果<sup>[12]</sup>,但协同过滤算法效果容易受到数据稀疏性影响。由于无需其他用户的评分数据,基于内容过滤算法则能避免这种问题。人口统计学过滤算法则易受用户隐私问题的限制,对推荐算法及其重要的信息往往是用户不愿透露的隐私信息。此类推荐算法在实际应用中很少被采用。基于以上考虑,结合多种过滤算法的混合式推荐算法可利用各算法优点,避免其缺点,获得更好的推荐效果<sup>[13]</sup>。

在学术界对推荐系统进行研究以来,提出了多种类型的混合过滤方法。其中,使用最广泛的两种是协同过滤和基于内容过滤<sup>[14,15]</sup>、协同过滤和人口统计学过滤<sup>[16]</sup>。协同过滤和基于内容过滤一般有四种混合模式<sup>[17]</sup>。第一种是分别计算协同过滤和基于内容过滤算法的推荐结果,并将二者结果加权输出<sup>[14,18-21]</sup>。第二种是将基于内容过滤的算法思想集成到协同过滤,以协同过滤的方式作出推荐<sup>[14,22]</sup>。第三种是建立一个新的模型来融合来自协同过滤和基于内容过滤的特征<sup>[23,24]</sup>。第四种是将系统过滤算法思想集成到基于内容过滤,以基于内容过滤的方式作出推荐<sup>[25]</sup>。

混合式过滤算法在不同的应用场景下有不同的目标。最常见的设计目标是提高系统的推荐准确度<sup>[14,16]</sup>。也有些应用场景是为了克服推荐系统的冷启动问题。此外,推荐系统需要处理大量的数据,亦有些混合式推荐算法的目的是提高计算效率。由于混合式过滤算法具有应用潜力,此类算法已在多个领域得以研究应用,如应用书籍<sup>[26]</sup>、电影<sup>[4,27]</sup>、音乐<sup>[28,29]</sup>等。除了上述的商品推荐外,混合式过滤推荐算法也被应用于推荐新闻<sup>[19,30]</sup>、网络教学课程<sup>[31-33]</sup>、数据图书<sup>[34]</sup>、旅游目的地<sup>[35,36]</sup>。然而在地学数据共享领域,缺少专业的数据推荐方法。

相比于传统的推荐应用,地理空间数据用户的需求更为专业和复杂。由于地理空间数据复杂的空间信息和时间属性信息,推荐算法面临更复杂的数据相似度计算问题。地学数据推荐较传统的多媒体内容推荐、商品推荐、文本内容推荐存在更复杂的挑战。此外,由于数据共享网站的设计思路不同于商业网站,往往缺少评价打分系统,用户的行为偏好也更难获取。针对这些挑战,以国家地球系统科学数据共享平台用户行为研究对象,开发一种基于协同过滤和内容过滤的混合式地学数据推荐方法。

## 2 混合式过滤地学数据推荐方法设计

如引言中所述,在有足够的用户评分数据情况下,协同过滤有较好的推荐效果。在协同过滤的实际应用中,往往有大量的项目缺乏足够数量的用户评分,使得项目间相似度计算的置信度较低。为了克服数据稀疏性问题,当用户对两个项目的评分数量不足时,可通过项目内容属性计算项目间相似度。针对地理空间数据,本研究设计一种动态的协同过滤和基于内容过滤混合算法,克服协同过滤情景下的数据稀疏问题。该方法同时利用了协同过滤在预测上的准确性优势和基于内容过滤在计算项目相似度及数据稀疏性上的优势。该方法的工作流程图如下图1所示。

本研究中所指项目 (item) 即为地理空间数据。项目相似度分别通过协同过滤和基于内容过滤计算。在基于内容过滤部分, 分别获取地理空间数据的主题、空间范围和时间范围信息, 用来计算数据的相似度。在协同过滤部分, 则通过用户间的共同评分来计算数据间的相似度。对于每一对用户——数据, 协同过滤和基于内容过滤分别对用户评分作出预测。最终的预测评分对上述两个预测评分采用动态加权的方法计算。这一计算方法的基本原理是基于当用户对某两个项目共同评分的数量越多, 协同过滤的预测能力越强, 则给协同过滤赋予更高的权重; 共同评分的数量越少基于内容过滤的预测能力越强, 则给基于内容过滤赋予更高的权重。

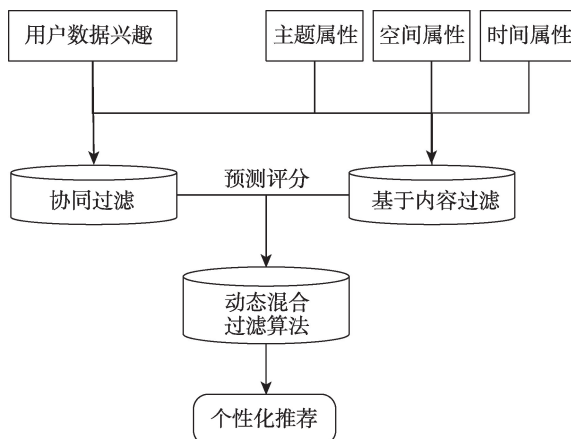


图1 混合过滤地理空间数据推荐方法流程图

Fig. 1 Workflow of the proposed hybrid filtering algorithm

## 2.1 基于内容过滤数据相似度计算

基于内容过滤进行用户评分预测的首要步骤是决定项目的相似度。相似度计算中, 通常使用一定数量的属性值来描述项目。以属性向量来表达项目, 即可通过向量计算的方式来确定项目的相似度。由于地理空间数据的空间和时间属性通常对应一个范围, 较传统应用场景下的商品、电影、音乐等, 其属性更难用值来表达, 相似度的计算也更为复杂。本研究从三个维度来定义地理空间数据的属性, 即空间范围、时间范围、主题内容。三个属性的值域通过数据集的元数据来抽取, 并分别计算三个维度的相似度。数据的总体相似度将通过上述三个维度的相似度加权求和获得。计算方法如下:

$$Sim_{con} = W_c \times Sim_{sub} + W_s \times Sim_{spa} + W_t \times Sim_{time} \quad (1)$$

式中:  $Sim_{sub}$ 、 $Sim_{spa}$ 、 $Sim_{time}$  分别表示数据的主题相似度、空间范围相似度、时间范围相似度;  $W_c$ 、 $W_s$ 、 $W_t$  分别表示上述三个相似度的权重。需要指出的是, 从地学数据用户需求的角度考虑, 数据A和数据B的相似度与数据B和数据A的相似度可能并不相同。例如, 数据A在某些维度上包含了数据B, 若用户对数据B有需求, 则数据A可能满足要求。反过来, 若用户对数据A有需求, 则数据B只能部分满足要求。计算所得的数据相似度矩阵是一个非对称矩阵。本研究定义此数据相似度为单向性相似度。三个维度的相似度计算方法将在下文小结中分别介绍。

公式(1)中的权重由领域专家打分确定。权重值采用一项地理空间语义相关度研究成果<sup>[37]</sup>。在咨询多位地学数据共享领域、地理空间语义领域、本体领域专家后, 确定主题内容、空间范围、时间范围的权重分别为0.41、0.35、0.24。基于此, 计算公式为:

$$Sim_{con} = 0.41 \times Sim_c + 0.35 \times Sim_s + 0.24 \times Sim_t \quad (2)$$

**2.1.1 主题内容相似度** 地理空间数据的主题相似度与传统推荐系统应用里的书籍、电影、音乐内容相似度类似, 由内容的描述属性确定。本研究从两个属性确定主题内容相似度: 关键字和分类层级。计算公式为:

$$Sim_c = W_{ck} \times Sim_{ck} + W_{cc} \times Sim_{cc} \quad (3)$$

式中:  $Sim_{ck}$ 、 $Sim_{cc}$  分别表示主题词相似度和分类层级相似度;  $W_{ck}$ 、 $W_{cc}$  分别表示二者的权重, 且  $W_{ck} + W_{cc} = 1$ 。权重的确定取决于领域知识, 本研究中取  $W_{ck} = W_{cc} = 0.5$ 。

每一个项目(地理空间数据)都有一定数量的关键词来描述。若数据*i*和数据*j*的关键词集合分别为*KW<sub>i</sub>*和*KW<sub>j</sub>*,则关键词相似度计算公式为:

$$Sim_{ck}(i,j) = \frac{|KW_i \cap KW_j|}{KW_i} \quad (4)$$

类似地,分类层级相似度以两个数据的分类层级重合度来计算。例如,数据*i*和数据*j*的分类层级分别为:

$$H_i: D_1 \rightarrow E_1 \rightarrow F_1 \rightarrow G_1$$

$$H_j: D_1 \rightarrow E_1 \rightarrow F_2 \rightarrow G_2$$

若分类层级深度表示为 $|H_i|$ ,在*i*和*j*的分类层级重合度为 $|H_i \cap H_j| - 1$ 。分类层级相似度的计算公式为:

$$Sim_{cc}(i,j) = \frac{|H_i \cap H_j| - 1}{|H_i| - 1} \quad (5)$$

则本例中的分类层级相似度为1/3。

**2.1.2 空间范围相似度** 相比于商品、电影、音乐等,地理空间数据的一个显著特征是其空间属性。计算两个地理空间数据集的空间相似度最直接的方法是计算二者的拓扑关系,确定二者的空间范围重合度<sup>[38]</sup>。然而,采用地理信息系统计算拓扑关系开销较大。在地理数据共享平台处理大量地理空间数据的应用场景下,该计算方法实用性较差。地理空间本体则记录了位置名词间的空间关系,能提供快速的空间关系查询,适用于大量空间的空间位置关系查询计算。近年来,有多项的空间信息检索研究应用了地理空间本体作为语义检索工具<sup>[39,40]</sup>。

平台共享的地理空间数据格式主要有栅格、矢量,及表格数据。不论格式,按几何类型所有的空间数据集可划分为面数据、线数据和点数据三种类型。从用户的数据需求角度考虑,三种类型的数据空间范围相似度计算原则为:

(1) 不同几何类型的数据间相似度取决于其空间位置是否有重叠。点状线状数据的面积可忽略。若点状或线状数据空间位置被面状数据包含,则该数据与面状数据相似度为1。而反之面状数据与点状数据或线状数据的相似度为0。计算公式为:

$$Sim_s(i,j) = \frac{|i \cap j|}{|i|} \quad (6)$$

(2) 两个点状数据间相似度取决于其空间位置是否相同,相同为1,不同为0。

(3) 两个线状数据或两个面状数据的相似度由他们之间重叠的程度确定。计算采用公式(6)。

以上空间范围相似度计算是模拟用户对数据需求认知,并基于地理空间名词语义关系计算的近似值。其计算的精确度依赖于元数据记录的空间位置级别(如县级、乡镇级)。基于用户的数据需求考虑,两个空间范围*i*和*j*的相似度是单向的(公式6),即*i*与*j*的相似度和*j*与*i*的相似度不同。最终获得的空间范围相似度矩阵也非对称矩阵。

**2.1.3 时间范围相似度计算** 由于时间的一维性,其相似度的计算较简单。时间范围相似度计算需考虑数据的时间数据类型。数据的时间属性类型有时间点和时间范围两种。时间属性A和B的相似度有二者的重叠程度确定。以 $|A|$ 和 $|B|$ 表示时间A和B的长度,则A和B的相似度计算公式为:

$$Sim_t(A,B) = \frac{|A \cap B|}{|A|} \quad (7)$$



## 2.2 协同过滤数据相似度计算

**2.2.1 项目相似度计算** 协同过滤分为基于用户的 (User-based CF) 和基于项目的 (Item-based CF) 两种。基于用户的协同过滤通过与用户有相同兴趣的用户群来预测用户偏好; 而基于项目的协同过滤则通过用户间共同评分计算项目相似度, 并依据用户历史预测用户偏好。科学数据共享平台提供的是专业性强的服务, 其用户群主要来自高校和科研院所。科研人员在一段时间内将保持其科研兴趣, 对某一主题的科学数据感兴趣。从这一角度考虑, 基于项目的协同过滤更符合本应用场景。

余弦相似度 (Cosine similarity) 是基于项目的协同过滤中最常使用的相似度计算方法<sup>[41]</sup>。然而, 余弦相似度忽略了不同用户对项目评分的习惯。一些用户倾向于较轻易地给出高评分, 而一些用户很少给出高评分。修正余弦相似度 (adjusted cosine similarity) 可克服这一问题。令  $U$  为同时对项目  $a$  和项目  $b$  作出评分的用户集合,  $r_{u,a}$  为用户  $u$  对项目  $a$  作出的评分,  $\bar{r}_u$  为用户  $u$  的所有评分的平均值, 余弦相似度  $sim_{cos}$  和修正余弦相似度  $sim_{adj\_cos}$  计算公式分别为:

$$sim(a,b)_{cos} = \frac{\sum_{u \in U} r_{u,a} \times r_{u,b}}{\sqrt{\sum_{u \in U} r_{u,a}^2} \sqrt{\sum_{u \in U} r_{u,b}^2}} \quad (8)$$

$$sim(a,b)_{adj\_cos} = \frac{\sum_{u \in U} (r_{u,a} - \bar{r}_u)(r_{u,b} - \bar{r}_u)}{\sqrt{\sum_{u \in U} (r_{u,a} - \bar{r}_u)^2} \sqrt{\sum_{u \in U} (r_{u,b} - \bar{r}_u)^2}} \quad (9)$$

修正余弦相似度的值域范围为  $-1 \sim 1$ 。但是, 2.1 节中基于内容过滤计算的项目相似度值域范围为  $0 \sim 1$ 。若要对这两种推荐方法进行融合, 则其相似度计算值域范围需统一。为了避免这一问题, 本研究在获取用户评分的过程中修正了用户的评分习惯, 并在此基础上采用余弦相似度 (值域为  $0$  到  $1$ ) 计算项目相似度。具体的用户评分获取方法将在下一小结中介绍。

**2.2.2 项目评分计算** 商业网站通常通过用户评分、浏览、收藏、购买等用户行为获取用户兴趣。科学数据共享平台也可通过用户浏览、下载等行为获取用户兴趣。本研究的目标是所有用户的行为模式, 包括匿名用户和注册用户。部分共享数据用户并不能直接下载, 且网站平台未提供直接的评分系统。因此, 本研究通过用户浏览时间来推算用户评分。受制图学里分级方法 Jenks Natural Breaks 的启发, 本研究通过该方法推算用户对数据的评分。该方法在每一级下将数据差异最小化, 可被看作是一维的  $k$ -means 算法。因此, 该方法能消除用户网络浏览行为习惯的差异。

首先, 通过日志数据获取用户对每个数据集的历史累计时间。然后针对各用户, 使用 Jenks Natural Breaks 法对数据集的浏览时间划分为 5 个等级, 分别代表评分的  $1 \sim 5$  分。表 1 中以用户浏览时间为例, 用户对数据集的累计时间从  $1 \sim 30$  min 不等, Jenks Natural Breaks 划分的 5 个等级为  $[1,2]$ 、 $(2,5]$ 、 $(5,7]$ 、 $(7,13]$ 、 $(13,30]$ 。

## 2.3 动态加权混合过滤模型

本研究提出的混合过滤模型对协同过滤和基于内容过滤的预测结果进行动态加权, 对不同的用户产生不同评分预测模型。假定协同过滤和基于内容过滤预测用户  $u$  对数据集  $i$  的评

表 1 Jenks Natural Breaks 划分用户浏览时间示例  
Tab. 1 Jenks Natural Breaks for item rating assignment

数据集	A	B	C	D	E	F	G
浏览时间 (min)	13	2	1	7	30	25	5
评分	4	1	1	3	5	5	2

分分别为  $pred_{CF}$  和  $pred_{CBF}$ , 基于内容过滤的权重为  $\beta$ 。则协同过滤的权重为  $(1-\beta)$ 。混合过滤模型的评分预测可表示为:

$$pred_{weighted}(u, i) = \beta \times pred_{CBF} + (1 - \beta) \times pred_{CF} \quad (10)$$

模型中协同过滤和基于内容过滤预测的评分得范围应一致。在协同过滤和基于内容过滤中, 分别采用  $k$  最邻近 ( $k$ -NN) 方法计算预测评分。该方法预测用户  $u$  对数据集  $i$  的评分公式为:

$$pred(u, i) = \frac{\sum_{r \in ratedItems(u)} sim(r, i) * r_{u, r}}{\sum_{r \in ratedItems(u)} sim(r, i)} \quad (11)$$

式中:  $r$  为用户  $u$  产生过评分的数据集;  $sim(r, i)$  为数据集  $r$  和数据集  $i$  之间的相似度。如前文所述, 在协同过滤中, 两个数据集间的相似度计算依赖于用户同时对这两个数据集产生的评分。对两个数据集同时评价的用户越多, 则相似度计算的置信度越高。当对两个数据集共同评价的用户越多, 则协同过滤应被赋予更高的权重。但共同评价的数量应如何影响权重的变化, 从而获得最优的模型是未知的。本研究采用回归模型拟合权重和共同评价数量的关系, 再将该模型应用到动态推荐模型中。从式 (10) 中可知, 权重  $\beta$  可表示为:

$$\beta = \frac{pred_{weighted}(u, i) - pred_{CF}}{pred_{CBF} - pred_{CF}} \quad (12)$$

若令  $pred_{weighted}(u, i)$  为用户  $u$  对数据集  $i$  的实际评分, 则计算出的  $\beta$  为理想的权重。本研究定义协同过滤中用来预测评分的  $k$  个最邻近数据集的平均共同评价数量为 CL (co-rating level), 则 CL 可表达为:

$$CL(u, i) = \frac{\sum_{j \in kNN} cn_j}{k} \quad (13)$$

式中:  $cn$  为对数据集  $i$  和数据集  $j$  的共同评价数量。通过上述方法, 可从训练样本中计算出相应的  $\beta$  和  $CL$  值。理论上  $\beta$  的值域范围为  $[0, 1]$ 。但由于用户行为的不确定性, 计算出的  $\beta$  范围可能超出  $[0, 1]$ 。本研究视超出此范围的  $\beta$  值为无效值。通过样本的有效  $\beta$  和  $CL$  值, 可拟合出二者关系:

$$\beta = f(CL) \quad (14)$$

将该拟合方程代入式 (14) 即可得出动态预测模型。

在协同过滤中, 通过用户历史浏览时间推算的用户评分为 5 个级别。该方法需要足够的数据集浏览数量 (至少 5 个) 来区分用户兴趣度的差异。若用户数据集浏览数少于 5 个, 则只采用基于内容过滤进行评分预测。此外, 随着  $CL$  的增加, 协同过滤的权重相应增加, 直到 1。以  $thre$  表示协同过滤的权重为 1 时  $CL$  阈值,  $n_u$  表示用户  $u$  访问过的数据集数量。则最终预测模型可表示为:

$$pred(u, i) = \begin{cases} pred_{CBF}(u, i) & : n_u < 5 \\ pred_{weighted}(u, i) & : n_u \geq 5, CL < thre \\ pred_{CF}(u, i) & : CL > thre \end{cases} \quad (15)$$

### 3 数据来源与实验设计

#### 3.1 数据来源

**3.1.1 服务器日志数据** 服务器日志数据是本研究用户行为数据的来源。本研究获取了

2015年的服务器日志数据进行试验,共12062607条。该日志数据以NCSA ECLF格式储存,每天日志信息里包含了用户IP、访问时间、方法、访问URL地址、状态、访问来源链接、客户端信息等。

**3.1.2 数据集元数据** 地空间数据集的元数据描述了数据的主题内容、空间范围、时间范围等信息,是基于内容过滤中计算数据集相似度的信息来源。在地球系统科学数据共享平台共享的数千个数据集中,本研究随机选择了200个样本数据集进行试验,并分别通过元数据提取了样本数据集的分类、关键词、空间范围、时间范围信息。

**3.1.3 地理空间本体** 本研究采用了王东旭等针对地学数据共享开发的地理空间本体<sup>[42]</sup>。通过本体查询工具,可获取不同地理名词间的空间拓扑关系,并用于数据间空间相似度计算。

### 3.2 数据预处理

原始数据存在大量的冗余信息,并不能直接用于实验。本研究进行了大量的数据预处理工作。对于服务器日志数据,预处理步骤包括数据清洗、用户识别、会话识别、用户访问时间计算。数据清洗是为了消除与挖掘任务无关的记录项,包括浏览器对图片、样式文件等的请求,网络爬虫的请求,以及错误的请求。用户识别是为了区分不同的用户。会话识别则在此基础上将不同用户的访问划分为单独访问时间段,作为一个完整的访问流程。本研究的数据预处理采用作者针对地学数据共享平台开发的预处理方法<sup>[43]</sup>。该方法在实际应用研究中表现出优秀的数据预处理效果<sup>[44]</sup>。会话识别后,以相邻两个访问记录的时间戳来计算访问时间。

此外,地理空间数据的元数据需进行提取和转换。元数据表以文本形式记录了数据的空间、时间和内容主题信息。本研究针对数据表格格式,开发了数据提取转换程序,获取了数据的空间描述词、时间范围描述,以及主题描述关键词。

### 3.3 实验设计

本研究随机选取了平台共享的200个数据集。根据用户历史访问,计算出用户对这200个数据集的评分。经过数据预处理,共得到7287个活跃用户的117375个评分。然后将这些评分中的70%作为训练集用于推荐系统中相似度计算,10%用于权重和CL关系的建模(建模集),剩下20%用于测试推荐效果(测试集)。推荐算法编程语言为Python。此外,基于内容过滤中数据相似度计算过程中采用Java Jena框架查询地理空间本体。

对训练集分别采用协同过滤和基于内容过滤中相似度计算方法计算数据集相似度,获得协同过滤数据集相似度矩阵和基于内容过滤数据集相似度矩阵。在计算协同过滤数据集相似度矩阵的同时,同时获取数据集的CL矩阵,用于记录协同过滤数据集相似度是基于多少共同评价而计算的。使用k-NN算法分别计算获得协同过滤和基于内容过滤对建模集中用户——数据集的预测评价。然后,通过公式(10)可得理想的权重计算方程:

$$\beta = \frac{r(u,i) - pred_{CF}}{pred_{CBF} - pred_{CF}} \quad (16)$$

式中: $r(u,i)$ 为用户的真实评价; $pred_{CF}$ 为协同过滤预测的评价; $pred_{CBF}$ 为基于内容过滤预测的评价。得到理想权重 $\beta$ 后,即可通过对应的CL来拟合 $\beta$ 和CL的关系。获得该拟合关系后,将该拟合方程应用于公式(10),获得动态的权重混合模型,并将该模型应用于测试集,检验推荐效果。

对于测试集中每个用户,本推荐方法将产生5个预测评分最高的数据集。采用准确度(Precision)和召回率(Recall)两个指标来评价推荐效果。实验分别测试了在不同的k值(k-NN算法中)情况下,协同过滤、基于内容过滤以及二者混合模型的推荐效果。

## 4 结果分析

理想权重 $\beta$ 计算结果显示, 建模集 11738 个评价中很大部分 (62%) 的 $\beta$ 在 $[0,1]$ 值域范围之外。因此, 使用剩下 38% 的评价用来拟合 $\beta$ 和 $CL$ 的关系。拟合结果显示二者关系最接近某一对数函数 (图2) 所示。拟合方程如公式 (17) 所示, 拟合的方程的决定系数 ( $R^2$ ) 为 0.328。

$$\beta = 0.581 + 0.059 \times \ln(CL) \quad (17)$$

令 $\beta=1$ , 则可得 $CL$ 的阈值为 1211。说明在 $CL>1211$ 时, 将仅采用协同过滤推荐结果。此实验结果获得的评分预测方程为:

$$\text{pred}(u,i) = \begin{cases} \text{pred}_{CBF} & : n_u < 5 \\ (0.581 + 0.059 \times \ln(CL)) \times \text{pred}_{CBF} + (0.419 - 0.059 \times \ln(CL)) \times \text{pred}_{CF} & : n_u \geq 5, CL \leq 1211 \\ \text{pred}_{CF} & : CL > 1211 \end{cases} \quad (18)$$

该推荐方法流程为: 程序首先检查用户的历史评价记录, 若历史评价记录数量 $<5$ , 则只启用基于内容过滤推荐算法; 若用户评价数量 $>5$ 且预测评分对象数据集的 $CL < 1211$ , 则启用混合推荐算法; 若 $CL > 1211$ , 则只启用协同过滤算法。

图3展示了推荐效果的对比结果。结果显示混合推荐模型的准确度和召回率指标较协同过滤和基于内容过滤有较大提高。当 $k=10$ , 混合推荐模型获得最佳的推荐效果, 准确率为 0.271, 召回率为 0.424; 协同过滤在 $k=15$ 时获得最佳推荐效果, 准确率为 0.216, 召回率为 0.338; 基于内容过滤则在 $k=10$ 时获得最佳推荐效果, 准确率为 0.153, 召回率为 0.239。

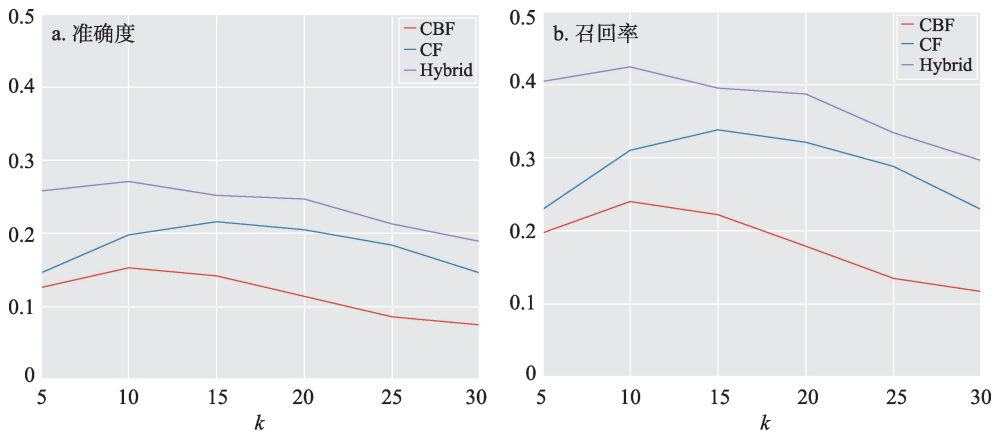


图3 准确度和召回率评价结果

Fig. 3 Precision (left) and Recall (right) evaluation of CBF, item-based CF and proposed Hybrid approach

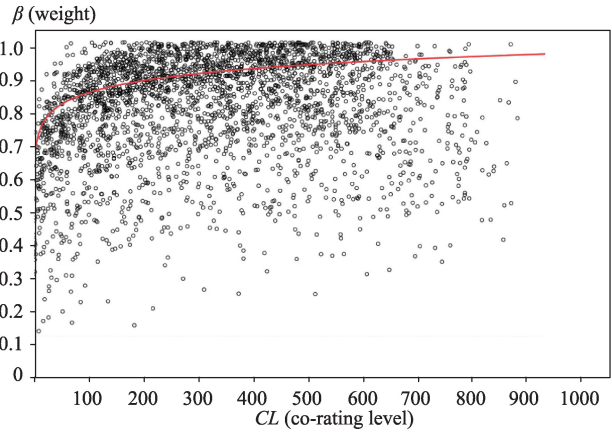


图2 理想权重 $\beta$ 和 $CL$ 散点及拟合图

Fig. 2 Scatter plot of ideal weight and co-rating level ( $CL$ )



## 5 结论与讨论

本研究提出了一种面向地理空间数据推荐应用的混合式推荐算法。从数据的空间范围相似度、时间范围相似度、内容主题相似度三个方面来解决基于内容过滤中的相似度计算问题。实验结果表明,本研究提出的动态加权混合式过滤算法较单纯的协同过滤或基于内容过滤的推荐效果有明显提高。将地理空间数据的时空属性作为推荐系统的输入变量,提高了推荐效果,可应用于地理空间数据网络服务。研究中提出的以 Jenks Natural Break 来区分用户兴趣度的方法,亦可用于其他领域用户行为研究。

地学数据个性化推荐较传统的文本内容、多媒体内容推荐具有更复杂的空间信息计算问题。且用户对推荐内容的要求更高,用户需求的替代性差。实验结果发现即便将数据的空间范围和时间范围考虑进相似度的计算,相比协同过滤和混合推荐方法,单纯基于内容过滤的推荐效果依然较差。这反映了预测用户数据需求的难度。可能是由于用户在获取地理空间数据过程中,会考虑众多难以计算的因素,如数据来源,数据质量等。

**致谢:** 感谢国家科技基础条件平台——地球系统科学数据共享平台为本研究提供数据支持。

## 参考文献(References)

- [1] Tenopir C, Allard S, Douglass K, et al. Data sharing by scientists: Practices and perceptions. *Plos One*, 2011, 6(6): e21101.
- [2] Kaššák O, Kompan M, Bielíková M. Personalized hybrid recommendation for group of users: Top-N multimedia recommender. *Information Processing & Management*, 2016, 52(3): 459-477.
- [3] Lee S K, Cho Y H, Kim S H. Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations. *Information Sciences*, 2010, 180(11): 2142-2155.
- [4] Wei S, Zheng X, Chen D, et al. A hybrid approach for movie recommendation via tags and ratings. *Electronic Commerce Research & Applications*, 2016, 18(C): 83-94.
- [5] Bobadilla J, Serradilla F, Hernando A. Collaborative filtering adapted to recommender systems of e-learning. *Knowledge-Based Systems*, 2009, 22(4): 261-265.
- [6] Zaiane O R. Building a recommender agent for e-learning systems. *Computers in Education*, 2002. Proceedings. International Conference on. IEEE, 2002: 55-59.
- [7] Huang Z, Zeng D, Chen H. A comparison of collaborative-filtering recommendation algorithms for e-commerce. *IEEE Intelligent Systems*, 2007, 22(5): 68-78.
- [8] Jcastroschez J, Miguel R, Vallejo D, et al. A highly adaptive recommender system based on fuzzy logic for B2C e-commerce portals. *Expert Systems with Applications*, 2011, 38(3): 2441-2454.
- [9] He Q, Jiang D, Liao Z, et al. Web query recommendation via sequential query prediction. *Data Engineering*, 2009. ICDE'09. IEEE 25th International Conference on. IEEE, 2009: 1443-1454.
- [10] McNally K, Coyle M, Briggs P, et al. A case study of collaboration and reputation in social web search. *Acm Transactions on Intelligent Systems & Technology*, 2011, 3(1): 1-29.
- [11] De Campos L M, Fernandezluna J M, Huete J F, et al. Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks. *International Journal of Approximate Reasoning*, 2010, 51(7): 785-799.
- [12] Burke R. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 2002, 12(4): 331-370.
- [13] Porcel C, Tejeda-Lorente A, Martínez M A, et al. A hybrid recommender system for the selective dissemination of research resources in a technology transfer office. *Information Sciences*, 2012, 184(1): 1-19.
- [14] Li Q, Kim B M. Clustering approach for hybrid recommender system. *Web Intelligence*, 2003: 33-38.
- [15] Spiegel S, Kunegis J, Li F, et al. Hydra: A hybrid recommender system [cross-linked rating and content information]. *Conference on Information and Knowledge Management*, 2009: 75-80.
- [16] Alejandro Bellogín, Castells P, Chavarriaga E. An empirical comparison of social, collaborative filtering, and hybrid recommenders. *Acm Transactions on Intelligent Systems & Technology*, 2013, 4(1): 1-29.

- [17] Bobadilla J, Ortega F, Hernando A, et al. Recommender systems survey. *Knowledge-based Systems*, 2013, 46: 109-132.
- [18] Billsus D, Pazzani M J. User modeling for adaptive news access. *User Modeling and User-Adapted Interaction*, 2000, 10 (2): 147-180.
- [19] Claypool M, Gokhale A, Miranda T, et al. Combining content-based and collaborative filters in an online newspaper. *Proceedings of ACM SIGIR Workshop on Recommender Systems*, 1999: 60.
- [20] Marx P, Hennigthrau T, Marchand A, et al. Increasing consumers' understanding of recommender results: A preference-based hybrid algorithm with strong explanatory power. *Conference on Recommender Systems*, 2010: 297-300.
- [21] Tran T, Cohen R. Hybrid recommender systems for electronic commerce. *National Conference on Artificial Intelligence*, 2000.
- [22] Melville P, Mooney R J, Nagarajan R, et al. Content-boosted collaborative filtering for improved recommendations. *National Conference on Artificial Intelligence*, 2002: 187-192.
- [23] Campos L, Mfernandezluna J, Fluete J, et al. Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks. *International Journal of Approximate Reasoning*, 2010, 51(7): 785-799.
- [24] Fouss F, Pirotte A, Renders J, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(3): 355-369.
- [25] Mooney R J, Roy L. Content-based book recommending using learning for text categorization. *Proceedings of the Fifth ACM conference on Digital libraries*. ACM, 2000: 195-204.
- [26] Vaz P C, Matos D M D, Martins B, et al. Improving a hybrid literary book recommendation system through author ranking. *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. Washington, DC, USA, ACM. 2012: 387-388.
- [27] Lommatzsch A, Kille B, Kim J W, et al. An adaptive hybrid movie recommender based on semantic data. *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*. Lisbon, Portugal, Le Centre De Hautes Etudes Internationales D'informatique Documentaire. 2013: 217-228.
- [28] Aureliodomingues M, Gouyon F, Mariojorge A, et al. Combining usage and content in an online music recommendation system for music in the long-tail. *World Wide Web*, 2012: 925-930.
- [29] Yoshii K, Goto M, Komatani K, et al. Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. *International Symposium/Conference on Music Information Retrieval*, 2006: 296-301.
- [30] Wen H, Fang L, Guan L. A hybrid approach for personalized recommendation of news on the web. *Expert Systems with Applications*, 2012, 39(5): 5806-5814.
- [31] Cobos C, Rodriguez O, Rivera J, et al. A hybrid system of pedagogical pattern recommendations based on singular value decomposition and variable data attributes. *Information Processing & Management An International Journal*, 2013, 49 (3): 607-625.
- [32] Salehi M, Kamalabadi I N. Hybrid recommendation approach for learning material based on sequential pattern of the accessed material and the learner's preference tree. *Knowledge-Based Systems*, 2013, 48: 57-69.
- [33] Zhuhadar L, Nasraoui O, Wyatt R, et al. Multi-model ontology-based hybrid recommender system in e-learning domain. *Web Intelligence*, 2009, (3): 91-95.
- [34] Vellino A, Zeber D. A hybrid, multi-dimensional recommender for journal articles in a scientific digital library. *Web Intelligence*, 2007: 111-114.
- [35] Al-Hassan M, Lu H, Lu J. A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system. *Decision Support Systems*, 2015, 72: 97-109.
- [36] Chen J, Chao K, Shah N, et al. Hybrid recommendation system for tourism. *International Conference on E-business Engineering*, 2013: 156-161.
- [37] 赵红伟, 诸云强, 杨宏伟, 等. 地理空间数据本质特征语义相关度计算模型. *地理研究*, 2016, 35(1): 58-70. [Zhao Hongwei, Zhu Yunqiang, Yang Hongwei, et al. The semantic relevancy computation model on essential features of geospatial data. *Geographical Research*, 2016, 35(1): 58-70.]
- [38] Schneider M. Computing the topological relationship of complex regions. *Database and Expert Systems Applications: 15th International Conference, DEXA 2004, Zaragoza, Spain, August 30-September 3, 2004 Proceedings*. Berlin, Heidelberg; Springer Berlin Heidelberg, 2004: 844-853.
- [39] Bowers S, Lin K, Ludascher B. On integrating scientific resources through semantic registration. *Scientific and Statistical Database Management. Proceedings. 16th International Conference on*. IEEE, 2004: 349-352.
- [40] Fox P, McGuinness D L, Cinquini L, et al. Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial

- Observatory experience. *Computers & Geosciences*, 2009, 35(4): 724-738.
- [41] Jannach D, Zanker M, Felfernig A, et al. *Recommender Systems: An Introduction*. Cambridge: Cambridge University Press, 2010.
- [42] 王东旭, 诸云强, 潘鹏, 等. 地理数据空间本体构建及其在数据检索中的应用. *地球信息科学学报*, 2016, 18(4): 443-452. [Wang Dongxu, Zhu Yunqiang, Pan Peng, et al. Construction of geodata spatial ontology and its application in data retrieval. *Journal of Geo-information Science*, 2016, 18(4): 443-452.]
- [43] Wang M, Wang J. A data preprocessing framework of geoscience data sharing portal for user behavior mining. 23rd International Conference on Geoinformatics, IEEE, 2015: 1-5.
- [44] 王末, 王卷乐. Web 环境下地学数据共享用户行为模式分析. *地球信息科学学报*, 2016, 18(9): 1174-1183. [Wang Mo, Wang Juanle. A study on the user behavior of geoscience data sharing based on web usage mining. *Journal of Geo-information Science*, 2016, 18(9): 1174-1183.]

## A hybrid personalized data recommendation approach for geoscience data sharing

WANG Mo<sup>1,2</sup>, ZHENG Xiaohuan<sup>3</sup>, WANG Juanle<sup>4,5,6</sup>, BAI Yongqing<sup>4,5</sup>

(1. Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081, China; 2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture, Beijing 100081, China; 3. Office of General Affairs, Chinese Academy of Sciences, Beijing 100864, China; 4. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China; 5. University of Chinese Academy of Sciences, Beijing 100049, China; 6. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China)

**Abstract:** Recommender systems are effective tools helping Internet users mitigate information overloading. In geoscience data sharing domain, items (datasets) are more informative in terms of spatial and temporal attributes compared to regular item (e.g. books, movies, music). Thus, high-performance recommendation algorithms for geoscience data are more challenging. This study proposed an approach that combines content-based filtering with item-based collaborative filtering using dynamic weights. The approach examines merits of both collaborative filtering in its predictive ability and item content information to mitigating data sparsity and early ratter problem. Users' ratings on items were first derived with their historical visiting time by Jenks Natural Breaks. In the CBF part, spatial, temporal, and thematic information of geoscience datasets were extracted to compute item similarity. Predicted ratings were computed with k-NN method separately using CBF and CF, and then combined with dynamic weights. With training dataset, we attempted to find the best model describing ideal weights and users' co-rating level. A logarithmic function was identified to be the best model. The model was then applied to tune the weights of CF and CBF on user-item basis with test dataset. Evaluation results showed that the dynamic weighted approach outperformed either solo CF or CBF approach in terms of Precision and Recall.

**Keywords:** recommender system; geoscience data; hybrid filtering; science data sharing