协同过滤在软件成本估算中的应用①

任雪利

(曲靖师范学院 计算机科学与工程学院，曲靖 655011)

摘 要：准确的成本估算是软件项目管理的重要目标, 但是现有的成本估算方法均有缺点. 协同过滤是一种群体智慧的方法, 已成功的应用于电子商务、影视推荐等多个领域. 本文将协同过滤技术应用于软件项目中成本的估算, 由于传统的协同过滤技术仅能处理数值型数据, 而描述项目特征的属性既有数值型也有非数值型数据, 因此采用不同的策略对属性进行归一化, 使用均值对缺失值进行处理, 余弦相似度用于计算项目间的相似度, 确定近邻集进行成本估算. 将该方法应用于 USP05-FT 数据集, 实验结果表明: 估算结果的准确性可以达到 80%以上.

关键词：协同过滤; 成本; 类比; 归一化

Application in Cost Estimation of Collaborate Filtering

REN Xue-Li

(School of Computer Science and Engineer, Qujing Normal University, Qujing 655011, China)

Abstract: Accurate project cost prediction is an important goal for the software engineering community, but there are some defects in the method to estimate software cost. Collaborative Filtering has been developed in information retrieval researchers successfully which recommends items based on other user's reference in historical data set. Cost estimation based on Collaborative Filtering is researched. Because only numerical data can be handled in traditional collaborative filtering technology, and there are non-numeric numeric data in the attributes that describe the project characteristics, so the different strategies are used to normalize for describing the project's properties. And then the mean values are used for the missing contents. Cosine similarity is used to calculate the similarity between projects. Finnally cost is estimated using the weighted sum of the efforts in k-nearest neighbors. The method is applied in an experimental case to evaluate the effort estimation, and the result shows the accuracy of estimation may arrive to 80%.

Keywords: collaborate filtering; cost; analogy; normalization

成本估算是对完成项目所需费用的估计和计划, 是项目计划中的一个重要组成部分. 要实行成本控制, 首先要进行成本估算[1]. 常用的成本估算方法有: 经验估算法, 算法模型估算及类比估算. 经验估算是由一个被认为是该任务专家的人来控制, 并且估算过程的很大一部分是基于不清晰、不可重复的推理过程, 因此, 专家的个人偏好、经验差异与专业局限性都可能为估算的准确性带来风险. 算法模型估算通过找到软件工作量的各种成本影响因子, 并判定它对工作量所产生的影响程度建立模型进行估算, 该方法客观、高效、可重复, 而且能够利用以前的项目经验进行校准;但是难以用在没有前例的场合, 也不能处理异常情况和成本驱动因子级别的问题[2,3]. 类比(analogy)估算通过对一个或多个已完成的项目与新的类似项目的对比来预测当前项目的成本与进度, 该方法比较直观, 而且能够基于过去实际的项目经验来确定与新的类似项目的具体差异以及可能对成本产生的影响, 但是不能适用于早期规模等数据都不确定的情况及新项目中约束条件、技术、人员等发生重大变化的情况. 协同过滤是一种在历史数据中确定相似用户或物品产生推荐的方法, 已成功的应用于电子商务、影视推荐等多个领域, 因此, 本文将协同过滤技术应用于软件项目

① 基金项目:云南省教育厅基金(2011Y010)
收稿时间:2013-11-02;收到修改稿时间:2013-12-12

成本的估算.

1 协同过滤

协同过滤推荐是当前最成功的推荐技术之一, 它基于这样一个假设: 如果用户对一些项目的评分比较相似, 则他们对其他项目的评分也会比较相似，算法的基本思想是: 目标用户对未评分项目的评分，可以通过其最近邻居对该项目的评分来逼近[4].

为了找出目标用户的最近邻居，需要度量用户之间的相似性. 目前主要有余弦(cosine)相似性、相关(correlation)相似性和修正的余弦(adjusted cosine)相似性 3 种[5]:

1) 余弦相似性: 把用户评分看作 n 维项目空间上的向量,用户间的相似性通过向量间的余弦夹角来度量.设用户 i 和用户 j 在 n 维项目空间上的评分分别表示为向量 $\vec{i}$, $\vec{j}$ ，则用户 i 和用户 j 之间的相似性为

$$sim(i,j)=\cos(\vec{i},\vec{j})=\frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|\|\vec{j}\|} \quad (1)$$

2) 修正的余弦相似性: 由于在余弦相似性度量方法中没有考虑不同用户的评分尺度问题,修正的余弦相似性度量方法通过减去用户对项目的平均评分来改善上述缺陷.设经用户 i 和用户 j 共同评分的项目集合用 I_{ij} 表示, I_i 和 I_j 分别表示经用户 i 和用户 j 评分的项目集合,则用户 i 和用户 j 之间的相似性为

$$sim(i,j)=\frac{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R}_i)(R_{j,c}-\overline{R}_j)}{\sqrt{\sum_{c\in I_i}(R_{i,c}-\overline{R}_i)^2}\sqrt{\sum_{c\in I_j}(R_{j,c}-\overline{R}_j)^2}} \quad (2)$$

$R_{i,c}$ 表示用户 i 对项目 c 的评分, $\overline{R}_i$ 和 $\overline{R}_j$ 分别表示用户 i 和用户 j 对项目的平均评分.

3) 相关相似性: 设经用户 i 和用户 j 共同评分的项目集合用 表示,则用户 i 和用户 j 之间的相似性 $sim(i,j)$ 通过 Pearson 相关系数来度量:

$$sim(i,j)=\frac{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R}_i)(R_{j,c}-\overline{R}_j)}{\sqrt{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R}_i)^2}\sqrt{\sum_{c\in I_{ij}}(R_{j,c}-\overline{R}_j)^2}} \quad (3)$$

$R_{i,c}$ 表示用户 i 对项目 c 的评分, $\overline{R}_i$ 和 $\overline{R}_j$ 分别表示用户 i 和用户 j 对项目的平均评分.

根据相似度确定用户的邻居，常用的挑选邻居的原则可以分为固定数量的邻居和基于相似度门槛的邻居两类. 然后依据邻居的相似度权重以及他们对项目的偏好，预测新项目的评分.

2 协同过滤在成本估算中的应用

协同过滤推荐是当前最成功的推荐技术之一，成本估算是软件项目管理的重要内容，本文将协同过滤技术应用于软件项目成本的估算. 具体过程如图所示:

图 1 成本估算过程

为了叙述的方便，首先给出以下定义:

定义 1.将具有 m 个项目和 n 个属性的项目-属性数据库表示为一个 $m\times n$ 的矩阵，矩阵的行表示项目，列表示属性. 其中: $p_i\in\{p_1,p_2...p_m\}$ 表示第 i 个项目， $a_j\in\{a_1,a_2...a_n\}$ 表示项目的第 j 个属性，矩阵中的值 $r_{ij}\in\{r_{11},r_{12}...r_{mn}\}$ 表示项目 p_i 的第 j 个属性 a_j 的取值. 如果属性的值缺失，那么为空. 归一化数据: 将属性的取值统一在相同的取值范围[0,1]中. 由于传统的协同过滤技术仅能处理数值型数据，而描述项目的属性类型可能是数值、集合、范围、模糊值或布尔类型等，例如: 项目的规模是一个具体的数值，项目的开发语言可能是{html,php,sql},团队的开发经验是[1，10]，项目的复杂程度等级是低，使用到了 CASE 集成工具，因此该值的 1，因此，首先需要将不同类型的属性进行归一化. 具体情况如下:

a) 若属性是数值，采用公式(4)将其归一化.

$$nor(r_{ij}) = \frac{r_{ij} - \min(P_j)}{\max(P_j) - \min(P_j)} \quad (4)$$

其中, P_j 表示所有项目的第 j 个度量值，$\max(P_j)$ 和 $\min(P_j)$ 分别表示第 j 个属性值中的最大值和最小值.

b) 若属性是集合类型，采用公式(5)将其归一化.

$$nor(r_{ij}) = \frac{|r_{ij}|}{|\bigcup_{i=1}^{m} r_{ij}|} \quad (5)$$

其中,r_{ij}|表示该属性集合种元素的个数，$|\bigcup_{i=1}^{m} r_{ij}|$表示所有项目的第 j 个属性集合的并集中元素的个数.

c) 若属性是范围，采用公式(6)将其归一化.

$$nor(r_{ij}) = \frac{\lfloor \max(r_{ij}) - \min(r_{ij}) \rfloor + 1}{\lfloor \max(P_j) - \min(P_j) \rfloor + 1} \quad (6)$$

其中, $\max(r_{ij})$ 与 $\min(r_{ij})$ 分别表示第 i 个项目的第 j 个属性值中的最大值和最小值， P_j 表示所有项目的第 j 个度量值，$\max(P_j)$ 和 $\min(P_j)$ 分别表示第 j 个属性值中的最大值和最小值.

d) 若属性是模糊型，首先将该模糊值转化为数字，在此，将模糊值按照程度轻重依此排序，并对应数字 1 开始的递增数值，然后采用公式(4)将其归一化.

e) 若属性的取值仅有两个，当逻辑型处理，采用公式(7)进行归一化.

$$nor(r_{ij}) = \begin{cases} 1 & \text{与待评估项目取值相同} \\ 0 & \text{与待评估项目取值不同} \end{cases} \quad (7)$$

处理缺失值: 由于在收集数据的过程中各种因素导致的数据缺失是不可避免的，因此，需要对项目中存在的缺失值进行处理. 常用的缺失值处理方法可分为删除存在缺失值的个案和缺失值插补[6]. 删除存在缺失值的个案是将存在缺失值的个案删除，造成了信息的极大浪费; 缺失值插补使用可能值对缺失值进行插补，常用的插补方法有: 均值插补，利用同类均值插补，极大似然估计及多重插补. 本文使用均值插补对缺失值进行处理，即使用数据的均值对缺失值进行填充.计算项目 p_a 和 p_i 的相似度 $sim(p_a, p_i)$: 将项目的属性取值表示为向量，通过计算向量的余弦夹角来表示项目的相似度，具体的方法是采用公式(8)进行.

$$sim(p_a, p_i) = \frac{\sum_{j \in M_a \cap M_i} ((nor(r_{aj}) \times nor(r_{ij}))}{\sqrt{\sum_{j \in M_a \cap M_i} (nor(r_{aj})^2)} \sqrt{\sum_{j \in M_a \cap M_i} (nor(r_{ij})^2)}} \quad (8)$$

其中, M_a 和 M_i 分别表示项目 p_a 和 p_i 的所有度量值，$sim(p_a, p_i)$ 是一个介于 0 到 1 闭区间的数值，表示项目 p_a 和项目 p_i 的相似程度.

估算成本: 计算出项目之间的相似度以后，根据待评项目的近邻集 k-nearest 估算该项目的成本. 在此假定项目 a 的第 b 个属性为成本，其评估值表示为 $\hat{r}_{ab}$，本文采用公式(9)进行估算.

$$\hat{r}_{ab} = \frac{\sum_{i \in k-nearest} (r_{ib} \times sim(p_a, p_i))}{\sum_{i \in k-nearest} sim(p_a, p_i)} \quad (9)$$

3 实例

本文从 USP05-FT 中选择了 4 个项目来说明基于协同过滤的成本估算过程. 通过公式 4-7 对数据进行归一化的结果如表 1 所示:

表 1　归一化结果

ID	Func	IC	DF	DE	DO	UFP	Lang	Tools	Texpr	Aexpr	TS	DBMS	Mtd	SAT
403	1	1	1	0	1	0.19	1	0.6667	0.84	1	0.3	1	1	1
716	0.86	0	1	0	0	0	0.5	0.1667	0.52	1	0.7	0	1	1
802	0.91	1	0	0.6	1	0.71	0.5	0.1667	0.52	1	1	1	1	0
810	0.91	1	0	1	1	1	0.5	0.1667	0.56	1	0.7	1	1	0

在此假定 403 是待评估的项目，则根据公式(9)计算相似度的结果如下:

$$sim(403,716) = 0.7979$$
$$sim(403,802) = 0.78841$$
$$sim(403,810) = 0.77495$$

在本文中选择 2 个项目作为近邻，则 403 的成本估算结果为:

$$\hat{w}_{ab} = \frac{0.7979 * 0.5 + 0.78841 * 8}{0.7979 + 0.78841} = 4.227563$$

该项目的实际成本为 3.

将该方法应用于 USP05-FT 中的其它项目，实验结果表明: 该方法的准确率可以达到 80%以上.

4 总结

成本估算是软件项目管理的重要内容，本文将协同过滤技术应用于软件项目的成本估算，通过将历史

项目集中不同类型的数据进行归一化，采用余弦相似度进行计算，确定待评估项目的近邻集，根据近邻集中项目的成本确定待评估项目的成本. 将该方法应用于 USP05-FT 中的项目实例，结果表明: 该方法的评估准确性可以达到 80%以上.

参考文献

1 Boehm BW, Abts C, Chulani S. Software development cost estimation approaches—A survey. Annals of Software Engineering, 2000,10(1-4):177–205.

2 Boehm BW, Clark B, Horowitz E, Westland C. Cost models for future software life cycle processes: COCOMO 2.0. Annals of Software Engineering, 1995, 1: 57–94.

3 Boehm BW, Valerdi R, Lane J, Brown A. COCOMO suite methodology and evolution. CrossTalk: The Journal of Defense Software Engineering, 2005, 18(4): 20–25.

4 许海玲，吴潇，李晓东，阎保平.互联网推荐系统比较研究. 软件学报,2009，20(2) : 350–362.

5 赵晨婷,马春娥.探索推荐引擎内部的秘密,第 2 部分: 深入推荐引擎相关算法,2011.3

6 数据缺失值的 4 种处理方法. http://www.itongji.cn/article/100311B2012.html,2013.9