Research on Similarity in Software Cost Estimation

Xueli Ren[1, a], Yubiao Dai[1, b]

[1] School of Computer Science and Engineer, Qujing Normal University, Yunnan 655011, China

[a]oliveelave@126.com, [b]abiaodai@163.com

Keywords: Cost estimation; similarity; collaborative filtering.

Abstract. The collaborative filtering techniques is applied to cost estimation of software projects in the paper, and four methods to calculate similarity in collaborative filtering are introduced which are cosine similarity, Euclidean distance, adjusted cosine similarity and correlation correction similarity. These methods are applied in USP05-FT projects to estimate cost, and the experimental results show that the adjusted cosine similarity and correlation correction similarity are better than other similarity calculation methods whose accuracy has reached more than 75%.

Introduction

Cost estimating and cost management ar one of the core tasks of software project management. In developing the project plans, it is necessary to make an estimate of the duration of the project , cost , manpower and other resources what does project need .The common methods to estimate cost for software projects fall into three main categories: algorithmic cost estimation, expert judgment and estimation by analogy. Algorithmic estimation involves the application of mathematical models, such as COCOMO [2-4], which is mainly a data-driven method, so it is objective and repeatable; accuracy of result estimated depends on the model constructed. Expert judgment relies on the experience and the judgment of experts, so it is subjective and unrepeatable, accuracy of expert based prediction is erratic. The idea of analogy-based estimation is to determine the cost of the target project as a function of the known costs from similar historical projects [5]. Compared with the other two categories of estimation methods, It can be not only applied in the very early phase of a software project when detailed information about the project is not yet available, and can be but also later improved when more detailed information is accessible. The accuracy of cost estimation depends on similar projects selected, so it isn't suitable for projects to which they aren't similar. Collaborative Filtering is an important estimation technique in the information retrieval research domain. It is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). It is suitable for cost estimation of software projects. The similar projects can affect the accuracy of estimation; therefore, the problem is researched in the paper.

Collaborative Filtering

The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue x than to have the opinion on x of a person chosen randomly. The common methods in computing similarity are Euclidean Distance, Cosine, Modify cosine and Pearson correlation.

Euclidean Distance: if uses rating look as the points in Euclidean space, then the distance in the points is similarity for them. If the common item set is Ij which include the items rated by user i and user j, $R_{i,c}$ and $R_{j,c}$ are the rate which are rated separately by user I and j, then the similarity between user I and user j is computed used Formula 1.

$$sim(i, j) = \frac{1}{1+\sqrt{(\sum_{c\in I_j}(R_{i,c}-R_{jc})^2)}} \quad (1)$$

Where $R_{i,c}$ is the rate of item c by user I; $R_{j,c}$ is the rate of item c by user j.

Cosine：If uses rate look as the vectors in n space, then the similarity between one user and the other user is defined as cosine between one vector and the other vector. If $\vec{i}$ and $\vec{j}$ are rating vectors by user i and user j, then the similarity between user I and user j is computed used Formula 2.

$$sim(i, j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\| \|\vec{j}\|} \quad (2)$$

Modify cosine: As the different user's rating scale does not considered in the cosine similarity, the modified cosine similarity is used to improve the defect by minus the average score of user rating for the project. If I_{ij} is the common item set that are rated by user I and user j, I_i and I_j are separately the rate which is rate by user I and j, then the similarity between user I and user j is computed used Formula 3.

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

Where $R_{i,c}$ is the rate of item c by user I; $\overline{R}_i$ and $\overline{R}_j$ are respectively the average rate for the whole items by user I and user j.

Pearson correlation：If the common item set is Iij which include the items rated by user i and user j,, then $sim(i, j)$ of the Pearson correlation similarity of two users I and j is defined as Formula 4.

$$sim(i, j) = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \overline{R}_i)(R_{j,c} - \overline{R}_j)}{\sqrt{\sum_{c \in I_{ij}} (R_{i,c} - \overline{R}_i)^2} \sqrt{\sum_{c \in I_{ij}} (R_{j,c} - \overline{R}_j)^2}} \quad (4)$$

Where $R_{i,c}$ is the rate of item c by user I; $R_{j,c}$ is the rate of item c by user j. $\overline{R}_i$ and $\overline{R}_j$ are average value of rate for the whole items by both user x and user y.

Cost Estimation Based on Collaborative Filtering

Collaborative Filtering is an important estimation technique in the information retrieval research domain [9-15]. It has been successfully applied in both information filtering and E-commerce applications. Collaborative Filtering is applied in the cost estimation of software project, and the processes are illustrated in figure 1.

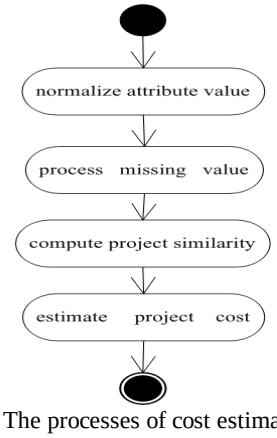

Fig. 1 The processes of cost estimation

A Definition is given for simplicity of description.

Definition1. The historical data set DB where there are m projects with n attributes is defined for a m×n matrix, where rows in the matrix show projects and series show attributes. is project i, shows attribute j of project; and the values in the matrix show the value of attribute j for project i. if the attribute value is missing, then the value is empty.

Normalize Attribute Value. Since each metric has different value range, this first step normalizes values of metrics so that the value range becomes [0, 1]. There are quantitative values and non-quantitative values in the set of attributes of projects, so the different method is used to normalize these attribute value. The method in [18] is used in the paper.

Process Missing Value. One of the practical problems in using the estimation methods is that the historical project data usually contain substantial numbers of missing values. Especially, process metrics contain larger numbers of them since they are collected by hand. MDTs can give bad influences to the accuracy of estimation, so some complementary techniques have been developed for dealing with missing values. The techniques were: listwise deletion, mean imputation and some types of hot-deck imputation [14]. Listwise deletion is the simplest technique to ignore data sets that have missing values. Mean imputation is a technique to fill the missing values on a variable with the mean of data sets that are not missing. Hot-deck imputation is alternative forms of imputation that are based on estimates of the missing values using other variables from the subset of the data that have no missing values. Mean imputation is applied in the paper to compute simply.

Compute Project Similarity. In this step, similarity $sim(p_a, p_i)$ is computed between the target project p_a and other projects p_i using 4 methods which are described in the foregoing in order to compare the accuracy of cost estimation [9-11].

Estimate Project Cost. The cost is calculated for the target project p_a using $sim(p_a, p_i)$ calculated in previous step. The steps are as following: Firstly, the k-nearest Projects are chosen based on similarity. Then the weighted sum is employed to compute estimation whose value is computed as the sum of the metrics values given by the other projects similar to p_a. Each value is weighted by the corresponding the $sim(p_a, p_i)$ between p_a and p_i. Formally, the value is defined using formula (5).

$$w_{ab} = \frac{\sum_{i \in k-nearest} (w_{ib} \times sim(p_a, p_i))}{\sum_{i \in k-nearest} sim(p_a, p_i)} \quad (5)$$

Where k-nearest Projects denotes set of k projects chosen (called neighborhoods) that have highest similarity with pa.

Example

Several projects are chosen from USP05 as an example to show the cost estimation, in that project 101 is known as the target project and the remaining projects are known as historical project set. As there are missing values and non- quantitative values in project set, the whole attribute values are normalized by the method in [7], and the mean imputation are using for missing value in the experiment, the result normalized is shown in table 1.

Table 1 The Result Normalized

ID	Func	IC	DF	DE	DO	UFP	Lang	Tools	Texpr	Aexpr	TS	DBMS	Mtd	SAT
101	0.333	1	1	0.392	0.26	0.429	0.33	0.143	1	1	0	1	1	1
102	0.25	1	0.471	1	0	0	0.33	0.143	1	1	0	1	1	1
208	0.167	0.75	0.059	0.013	0.02	0	1	0.429	0	0.667	0.2	0	1	1
210	0.167	0	0.059	0.004	0.02	0	1	0.429	0	0.333	0.2	0	1	1
509	0.2	0.25	0.647	0	0	0.286	0.78	0.31	0.35	0	0.4	0.5	1	0
510	0.143	0.75	0	0.083	0.4	0.571	0.78	0.31	0.35	0	0.4	0.5	1	0
511	0.167	0.5	0.235	0.083	0.1	0.714	1	0.286	0.06	0.333	0.2	0.5	1	0
512	0.167	0.75	0.588	0.25	1	1	1	0.429	0.06	0	1	0.5	1	0

If project 101 is known as the target project to estimate cost, the similarity between project 101 and the others is computed using four methods, then the result of similarity is shown in figure 2.

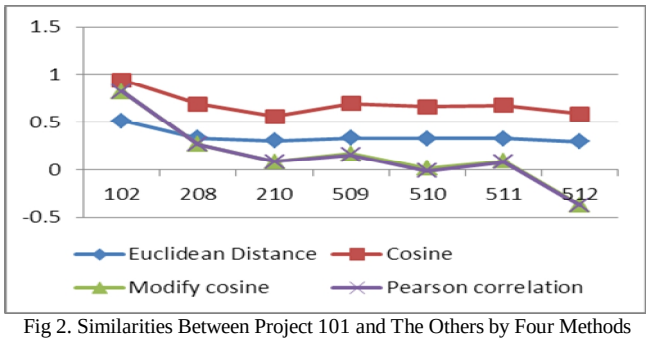

Fig 2. Similarities Between Project 101 and The Others by Four Methods

As the amount of historical projects is too small, 2 projects are chosen as nearest neighbors by similarity, the cost is estimated using formula (6), the result of estimation are 3.880135019, 3.989211486, 3.114245525 and 3.116081903, and cost of project 101 is 2.5 in fact, so modify cosine is the most accuracy, The accuracy is over 75%, the amount of historical projects and nearest neighbors can decrease the accuracy.

Summary

Cost estimation is an important content in software project management, the method based on analogy is an effect algorithm to estimate cost whose accuracy has direct relationship with the choice of similar projects. Collaborative Filtering is an important estimation technique in the information retrieval research domain. It has been successfully applied in both information filtering and E-commerce applications. CF is applied in cost estimation in the paper, which chooses similarity projects with the target project in historical projects set by similarity which is computed by Euclidean distance, cosine similarity, modified cosine similarity and Pearson correlation similarity, and then the cost of the target project is predicted using weight sum. The method to estimate cost introduced in the paper is used to project set of USP05-FT, The results of experiment show that modified cosine similarity and correlation similarity are better than the others, and the accuracy of estimation may achieve to 75%.

References

[1]. Zhang haipan. Introduction to software engineer. Tsinghua University press.2015:50

[2]. Boehm BW, Valerdi R. Achievements and challenges in software resource estimation. Technical Report, 2005.

[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]. Boehm BW, Royce W. Ada COCOMO and the Ada process model. In: Proc. of the 5th Int'l COCOMO User's Group Meeting. Pittsburgh, 1989.

[5]. B.W. Boehm et al "The COCOMO 2.0 Software Cost Estimation Model", American Programmer, 1996:5-7.

[6]. M. Shepperd and C. Schofield, "Estimating software project effort using analogy", IEEE Trans. Soft. Eng. 1997: 736-738..

[7]. LI Ming-Shu1, HE Mei. Software Cost Estimation Method and Application. Journal of Software.2007:777-783

[8]. Ralph Bergmann. Introduction to case-based reasoning.
http://www.dfki.uni- kl.de/~aabecker/Mosbach/Bergmann-CBR-Survey.pdf,2014.12

[9]. Watson. Case-based reasoning is a methodology not a technology. knowledge-based system. elsevier,1999:304-307.

[10]. Comparison Study of Internet Recommendation System, http://d.wanfangdata.com.cn/Periodical/rjxb200902013

[11]. Collaborative filtering, https://en.wikipedia.org/wiki/Collaborative_filtering,2015.5

[12]. Benjamin Marlin.Collaborative Filtering:A Machine Learning Perspective,21-23

[13]. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl.Item-Based Collaborative Filtering Recommendation Algorithms.ACM,2001:286-289.

[14]. Euclidean distance, https://en.wikipedia.org/wiki/Euclidean_distance, 2015.8

[15]. Pearson Correlation Coefficients, http://baike.baidu.com/link?url=RfN3xBwLjuYgg6BFT2cQEKWHADgy_KUVjObXC0Kd6CcdOxTVF2hKT-scdPwjtK1UXYIYEcATIehBYsT3MP6hJa,2015.3

[16]. Missing value[EB/OL]. http://baike.baidu.com/view/1578358.htm,2014.10

[17]. Shichao Zhang, Yongsong Qin.. Optimized parameters for missing data imputation. Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2006.

[18]. Xueli Ren,Yubiao Dai. Application in Effort Estimation of Collaborative Filtering,iscid,2013:331