A Comparison of Decision-Making Models for Determining File Importance

K. C. Wong (k-wong@govst.edu)
Division of Science
Governors State University
University Park, IL 60466 USA

Keywords: cognition; user modeling

Introduction

File replication is the most popular approach used to promote system reliability and file availability in a network-based environment (Purdin, et al. 1987; Son, 1987; Rodrigues, et al., 2002). However, all of the distributed file systems equipped with the functionality of file replication require their users to determine how important their files are in order to assist systems in making decisions regarding how many replicas should be made and distributed in the networks (Blair, et al., 1983). As such, system users are inevitably burdened with this potential responsibility. The problem can be partially alleviated if systems can take more responsibility for their users on determining file importance. To achieve this goal, however, we need to better understand how system users cognitively make decisions regarding determining file importance. In this paper, we quantitatively compare the performance of three decision-making models popularly used in juror decision-making (Pennington & Hastie, 1981) to examine how satisfactorily they model the process of determining file importance. The three models are the linear weighting model, the Bayesian model, and the Poisson model.

The Three Decision-Making Models

The Linear Weighting Model

The linear weighting model postulates that file importance can be determined by linearly combining those weighted pieces of information (referred to as predictors in this paper) during the session of determining file importance. The set of weights associated with the predictors identified can be determined in such a way that predicted file importance is optimally correlated with observed file importance using multiple regression analysis (Rawlings, 1988).

The Bayesian Model

The Bayesian model postulates that file importance can be determined by a series of simple inferences, in which importance is revised according to the direct impact of the predictors identified independently. In other words, the determination of file importance using the model is concerned with determining the posterior odds for importance (R_n), which is defined in terms of determining the ratio of the probability of importance given all the predictors identified, to the probability of unimportance given all the predictors identified. Once R_n is determined, it is compared with the decision criterion (dc) adopted by the system user to judge if the file under consideration is important (if $R_n \geq dc$) or not (if $R_n < dc$).

The Poisson Model

The Poisson model postulates that determining file importance is a Poisson process. In the process, it assumes that there exists an apparent weight of predictors (w) important to the file under consideration. The apparent weight accumulates constantly with time during the session of determining file importance until either a critical predictor is identified or the end of the session is encountered. The apparent weight accumulated (w_a) is then compared with the decision criterion (dc) adopted by the system user to judge if the file under consideration is important (if $w_a \geq dc$) or not (if $w_a < dc$).

Data Collection And The Experiment

Five predictors were systematically identified in this study for model comparison: the number of characters keyed, the computer cost spent, file length, file dependency, and the frequency of file access. Correlation coefficients between observed and predicted file importance were used to quantitatively evaluate the performance of the three models. A computer program, written in C++, was designed and implemented on a laptop to collect data for observed file importance and the five predictors. The data collected were classified into five importance ratings (from important to unimportant) and mapped proportionally to an importance rating scale (from 1 to 5, respectively). There were 41 subjects (randomly selected in an academic environment) participating in the experiment. These subjects accessed a total of 169 files. Since the subjects were asked to randomly pick up their files created by them, the sample may contain various types of file contents.

Model Comparison

Correlation Coefficients

The correlation coefficients computed for each of the models suggest that the linear weighting model and the Bayesian model with dc = 1 perform much more satisfactorily than the Poisson model using the empirical data collected in the study. The poor performance of the Poisson model may be resulted from the following three possible sources of errors: (1) the data collected may not be representative; (2) the assumptions made in this study may not hold for the model; (3) the model itself is inferior. More studies are needed to clarify the issue.

Nature of File Importance Determination

The linear weighting model is characterized by the nature of determining file importance slightly different from the Bayesian model and the Poisson model. The former model determines how important the file under consideration is (a rated outcome), while the latter models determine whether or not the file under consideration is important (a binary outcome). Moreover, the linear weighting model associates file importance ratings directly with predictor ratings in determining file importance. On the other hand, the Bayesian model and the Poisson model convert predictor ratings into predictor appearance probability, which may not be directly related to file importance ratings. As such, the linear weighting model provides more information about how each of the predictors is correlated with each other, and how each of the predictors is weighted by the subjects.

Implementation Efficiency

There is no noticeable performance difference in model implementation and file importance determination using the three models. All of the three models need an order of $O(n \times m)$ accesses to various data items for model implementation and an order of $O(m)$ accesses to determine predicted file importance, where n = the number of files created by a subject and m = the number of predictors each file has.

Decision-Making Processes

The three models studied have quite different decision-making processes, reflecting how system users cognitively make decisions regarding determining file importance. The linear weighting model postulates that determining file importance is a process consisting primarily of two phases: predictor collection and predictor evaluation. In the predictor collection phase, all possible predictors are collected. The predictors collected are then assigned weights in the predictor evaluation phase and combined linearly to determine file importance.

The Bayesian model postulates that in the process of determining file importance, once a predictor is identified, it will be evaluated to examine how likely the predictor is the one identified, given that the file under consideration is important and unimportant, respectively. The likelihood ratios thus computed constitute a series of inferences, in which posterior odds for importance is revised according to the direct impact of the predictors identified independently. At the end of the process, the revised posterior odds is compared with the decision criterion adopted by the subject to determine whether or not the file under consideration is important.

The Poisson model postulates that there exists an apparent weight of predictors important to the file under consideration. The apparent weight accumulates constantly with time in the process of determining file importance until either a critical predictor is identified or the end of the process is encountered. In the process, once a predictor is identified, it is judged by the subjects to examine if it is a critical predictor. The apparent weight accumulated up to the time when the critical predictor appears or the process ends is compared with the decision criterion adopted by the subject to determine whether or not the file under consideration is important.

References

Blair, G. S. Blair, Mariani, J. A., Nicol, J. R. and Shepherd, W. D. A Knowledge Based Operating System, *The Computer Journal*, vol. 30, no. 3, pp. 193-200, 1983.

Pennington, N. Pennington & Hastie, R. Juror Decision Making Models: The Generalization Gap, *Psychological Bulletin*, vol. 89, no. 2, pp. 246-287, 1981.

Prudin, Titus D. M., Schlichting, Richard D., & Andrews, Gregory R. A File Replication Facility for Berkeley Unix, *Software – Practice and Experience*, vol. 17(12), pp. 923-940, 1987.

Rawlings, J. O. Applied Regression Analysis: A Research Tool, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1988.

Rodrigues, Rodrigo, Liskov, Barbara, & Shrira, Liuba. The Design of Robust Peer-To-Peer System, *Tenth ACM SIGOPS European Workshop*, September, 2002.

Son, Sang Hyuk. Using Replication to Improve Reliability in Distributed Information Systems, *Information and Software Technology*, vol. 29, no. 8, pp. 440-449, 1987.