USING DECISION TREE LEARNING TO PREDICT WORKFLOW ACTIVITY TIME CONSUMPTION

Liu Yingbo, Wang Jianmin
School of Software, Tsinghua University, Beijing, China

Sun Jiaguang
School of Information Science and Technology, Tsinghua University, Beijing, China

Keywords: Time analysis, Workflow management system, Machine learning.

Abstract: Activity time consumption knowledge is essential to successful scheduling in workflow applications. However, the uncertainty of activity execution duration in workflow applications makes it a non-trivial task for schedulers to appropriately organize the ongoing processes. In this paper, we present a K-level prediction approach intended to help workflow schedulers to anticipate activities' time consumption. This approach first defines K levels as a global measure of time. Then, it applies a decision tree learning algorithm to the workflow event log to learn various kinds of activities' execution characteristics. When a new process is initiated, the classifier produced by the decision tree learning technique takes prior activities' execution information as input and suggests a level as the prediction of posterior activity's time consumption. In the experiment on three vehicle manufacturing enterprises, 896 activities were investigated, and we separately achieved and average prediction accuracy of 80.27%, 70.93% and 61.14% with K = 10. We also applied our approach on greater values of K, however the result is less positive. We describe our approach and report on the result of our experiment.

1 INTRODUCTION

Time is always precious. An accurate knowledge of time consumption is serviceable to an enterprise's workflow management system to schedule the ongoing processes. However, strong interactions between human and computer in workflow applications often make it difficult for schedulers to anticipate activity's time consumption, which is an important reason that prevents existing scheduling techniques from being used in workflow (Greg et al., 2004).

Consider, an example of enterprises we investigated, within a period of 31 months (from Oct-31-2003 to Jun-06-2006), there are 922 activities that have been executed at least once and 147 performers left 99765 event entries in the workflow event log. Statistics of this event log shows a great variety of activity execution duration ranging from a low of only 1 second to a maximum of 252 days. Even if we exclude those outliers by neglecting top and bottom 5% of observed execution duration, the range is still greater than 16 hours. Thus, an essential first step in achieving good scheduling in workflow management system is to look for ways of predicting activity time consumption.

As a means of anticipating workflow activities' time consumption, we present a K-level prediction approach. This approach uses a machine learning technique to recommend to a workflow scheduler a level as the prediction of possible time consumption. This information can benefit a workflow application in at least two aspects: it may help activity performers to pick up suitable work items from their work lists. And, it may help a workflow scheduler to figure out feasible priority of ongoing processes.

Our approach requires an enterprise's workflow system to have had an event log for some period of time and the workflow models from which the patterns of activities' time consumption can be learned. We believe this information is generally available for most of current workflow management systems.

Yingbo L., Jianmin W. and Jiaguang S. (2007).
USING DECISION TREE LEARNING TO PREDICT WORKFLOW ACTIVITY TIME CONSUMPTION.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - AIDSS*, pages 69-75
DOI: 10.5220/0002404900690075

Table 1: General overview of three enterprises' workflow event log.

Enterprise	A	B	C
Operation Time	117 days	421 days	949 days
Event Entries	10808	42099	99765
Number of Actors	179	244	147
Workflow Models	21	24	49
Max Duration	22 days	122 days	252 days
95 Percentile Duration	11 hours	20 hours	17 hours
5 Percentile Duration	7 seconds	5 seconds	7 seconds
Min Duration	2 seconds	1 second	1 second

In the experiment on three vehicle manufacturing enterprises, a total number of 896 activities were investigated, and we have been able to correctly suggest the level of time consumption with an average prediction accuracy of 80.27%, 70.93% and 61.14% respectively with K = 10. We have also applied our approach on greater values of K, however, the prediction accuracy decreases monotonically. When K reaches 100, the average prediction accuracy decreases to only 46.91%, 36.06% and 30.91%. In addition, we also found that the operation time of workflow has a positive influence on the prediction accuracy.

This paper makes two contributions: It presents an approach for helping workflow schedulers to anticipate activity time consumption and it evaluates the approach on the data sets from three real world enterprises.

The remainder of this paper is organized as follow: we begin by presenting background information about workflow, and we provide an overview on workflow application in three enterprises (Section 2). Given this background, we describe our K-level approach to predict activity time consumption (Section 3) and evaluate the results of applying our approach on real data sets (Section 4). In the followed section, related efforts of time management and scheduling in workflow systems are presented (Section 5). Finally, we summarize the paper (Section 6).

2 BACKGROUND

Understanding our approach requires a basic knowledge of workflow. These concepts will be covered in this section. In addition, we provide an overview of workflow application in three enterprises.

2.1 Workflow Structure and Event Log

We first present a set of definitions that will be used throughout this paper.

A *workflow* or *workflow model* is a description of a business process in sufficient detail that it is able to be directly executed by a workflow management system. A workflow is composed of a number of *activities* or *tasks*, which are connected in the form of a directed graph. An executing instance of a workflow is called *workflow instance* or *case*. There may be multiple instance of a particular workflow running simultaneously, however each of these instances is assumed to have an independent existence and they typically execute without reference to each other(Russell et al., 2005).

In the discussion of this paper, we treat activities in a workflow as a single unit of work, which will be undertaken by some *actors* or *performers*. Each invocation of an activity that executes is termed a *work item*. In general, a work item is directed to an *actor* for execution. An *activity's time consumption* or *execution duration* is the interval calculated from the time when the work item is accepted by an actor to the time that work item is committed by him. Once the actor commits a work item, corresponding activity will be marked as completed and other activities will be invoked, mean while, an *event entry* is created to log the actor's operation, including work item's time stamp, actor's identity and workflow instance id etc. These event entries form a workflow system's *event log*.

2.2 Overview of Workflow Event Log in Enterprises

In previous section, we have outlined basic concepts of workflow management system. As a further introduction to the background, we provide information about three enterprises. All these three enterprises are vehicle manufacturing enterprises. We investigate them because workflow is successfully used in many aspects of their business,

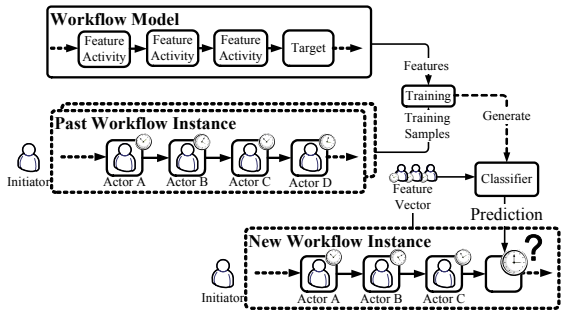

Figure 1: Machine learning based activity time consumption prediction.

like: configuration change, order processing, design review, technical notification, standard release, and new material classification etc.

Table 1 is a general overview of workflow event log in these enterprises. In order to maintain confidential, we use A, B and C to represent them. As illustrated in the table, the workflow system in these enterprises has a different length of operation time. Besides, there are many actors who have left lots of event entries, which clearly reveals the fact that workflow has been heavily used. Nevertheless, in all these enterprises, the activity time consumption varies greatly, which leads to the introduction of our K-level prediction approach.

3 K-LEVEL ACTIVITY TIME CONSUMPTION PREDICTION

Our approach of activity time consumption prediction is based on machine learning, its rationale is illustrated in figure 1. First, we define K levels so as to make different observation of time consumption uniformly distributed into the ranges of levels. Then, for a given activity, each event entry of this activity can be viewed as a training sample (or instance) and the event entries of those prior activities in the same workflow instance can be viewed as this training sample's features. The training sample may have a label that indicates its time consumption level. A supervised machine learning algorithm takes as input a set of training samples with known labels and generates a classifier. The generated classifier can then be used to assign a label to an unknown sample, which, in the context of workflow, is the time consumption level of unexecuted activity. The process of creating a classifier from a set of instances is known as training the classifier.

As is typical in machine learning, we evaluate the performance of each classifier using 10-fold cross-validation(Jiawei and Kamber, 2001).

In order to train a classifier, we take following steps:

- Selecting appropriate levels and target activities
- Determining features activities from workflow model
- Constructing training set from event log
- Applying machine learning to obtain a classifier

3.1 Selecting Appropriate Levels and Target Activities

The first step of our approach is to discretize observed time consumption into K levels so as to assign appropriate label to a given event entry. However, it is unwise to simply divide the maximum duration by K, and equally segment the time into K levels, because, in real situation, the frequency distribution of time consumption skews greatly. In our experiment, we use *a-quantile (a=1/K, 2/K,…1)* of observed time consumption as levels, this selection makes the interval between consecutive levels changes according to the density of time consumption distribution. Figure 2 is an example of 10-level selection in three enterprises.

In practice, K indicates the resolution of prediction. Higher value of K means finer granularity and stronger comparability of prediction result. Although, a higher resolution tends to make workflow schedulers to be more sensible, it, as we will see in the experiment results, usually leads to lower prediction accuracy.

After levels have been defined, each event entry can be assigned a label. The following step of selecting target activity is quite simple. In order to

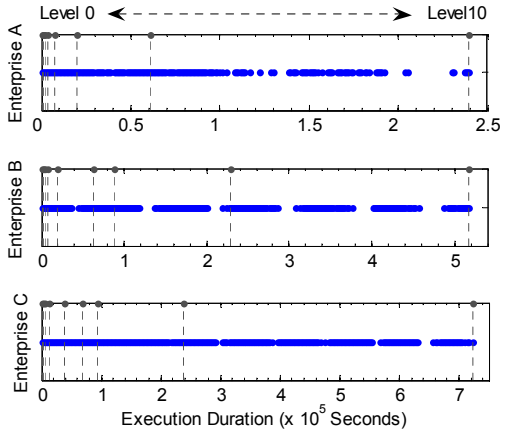

Figure 2: 10-Level selection of three enterprises.

cover as many activities as possible, we just excluded those activities whose event entries is not sufficient for 10-fold cross-validation and finally 896 activities are included in our investigation. The numbers of activities in three enterprises are listed in Table-2.

Table 2: Investigated activities in three enterprises.

Enterprise	A	B	C
Total Activities	256	399	922
Investigated Activities	104	243	522

3.2 Determining Feature Activates from Workflow Model

In order to train a classifier for a given activity, we need to find out which event entries are similar to each other, so that typical time consumption patterns can be derived by learning algorithm. The similarity is based on characteristics of prior activities' event entries in the same workflow instance. We find out these precedence activities by first excluding edges in a workflow model that might cause loop execution, and we make the relation on activities to be a directed acyclic graph, thus for any activity its precedence activities set can be obtained.

3.3 Constructing Training Set from Event Log

After feature activities have been selected, each event entry of the target activity can be characterized by a feature vector, and the associated label for this event entry is represented by the time consumption level.

However, there is a substantial amount of information in event entries that can be used as feature. Which part of information is selected fundamentally determines the performance of classifiers. In our experiment, we use three parts of information as features:

- The first part is actor's identity for those prior activities. We make this selection because it is commonly believed that, staff assignment has a strong influence on activity time consumption;
- The second part is prior activity's time consumption, this selection is based on the assumption that prior activity's time consumption may reveal posterior activities' characteristics in a workflow instance;
- The last part is the start time of prior activities calculated from the time corresponding workflow instance started, we choose this part of information because actors are not always interacting with workflow systems, a pending work item means there are some external reasons that prevent the instance from being completed, hence, it might has some influence on posterior activity's time consumption.

Finally, we construct the training set by collecting all the features of target activities' event entries from the workflow event log.

3.4 Applying Decision Tree Learning to Obtain a Classifier

In the final step, we use C4.5 decision tree(Quinlan, 1993) to obtain a classifier, we choose this algorithm because it is proposed by previous research(Ly et al., 2006). For the purpose of this paper is to testify the applicability of machine learning approach in activity time consumption prediction, we use existing tool WEKA (Witten and Frank, 2005) to train our classifiers and to perform the test.

4 EXPERIMENT RESULT AND EVALUATION

To demonstrate how well our approach can be used in real world applications and to see the relationship between prediction accuracy and resolution. We applied our approach on three enterprises' data sets with K = 10, 20, 40, 60 and 100.

Because, considerable number of classifiers is going to be trained in the data sets of each enterprise, we use average prediction accuracy of all classifiers as a global measure to represent the main feature of the performance of our approach.

4.1 Experiment Results

The exact numbers of average prediction accuracy are listed in Table-3 and the trend of prediction accuracies with regard to different values of K are depicted in Figure 3.

Table 3: The exact number of average prediction accuracy in three enterprises.

Levels	Enterprise A	Enterprise B	Enterprise C
10	61.14	70.93	80.27
20	51.88	59.94	72.77
40	43.56	48.52	62.55
60	38.03	43.15	54.95
80	34.52	38.10	50.85
100	30.91	36.06	46.91

In Figure 3, the trends of prediction accuracy in three enterprises are rather alike, but the absolute

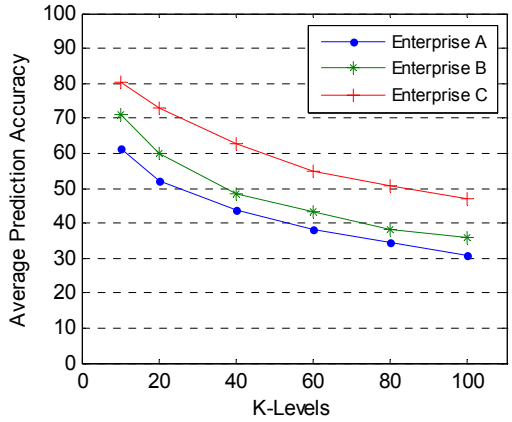

Figure 3: Average prediction accuracy of three enterprises with different values of K.

value is different. Our approach always performs the best in Enterprise C (with the best accuracy of 80.27%) while the worst in Enterprise A (with the best accuracy of just 61.14%).

We believe this is mainly because of the quality of training set. As shown in the discussion of previous sections, the performance of our approach depends on activity time consumption pattern and on how clearly the pattern displays itself in event log. However, it is quite rare for the situation to be so obvious, especially, in the initial phase of workflow application. Comparing with the data listed in Table-1, one may find that the workflow operation time in Enterprise A is rather short(less than 4 months), which means, typical time consumption patterns are not so clear to be generalized by C4.5 learning algorithm. Hence, the performance of classifiers is not likely to be good. Whereas, the workflow operation time of enterprise C is much longer than that of the other two. This long operation time has lead to a larger number of event entries and more importantly, bigger sample space for C4.5 algorithm to learn. Therefore, we believe, as time goes by, the prediction accuracy in Enterprise A and B will steadily increase.

4.2 Evaluation on Applicability

In our opinion, whether or not our approach is acceptable for scheduling will depend on requirement. According to experiment results, there appears to be a contradiction between accuracy and resolution, but the overall performance of our approach will gradually increase as time goes by. Therefore, a workflow scheduler needs to tradeoff between accuracy and resolution according to workflow operation time.

For example, if a scheduler requires some fixed prediction accuracy, then, at the beginning, this prediction have to be based on indefinite resolution and few activities can be well predicted, so, decisions have to be made on a vague knowledge of time consumption, and, these decisions tends to be rough. While, after a period of time, with resolution level becomes higher and higher and well predicted activities becomes more and more, the schedule can be more specific.

However, to the best of our knowledge, most of scheduling approaches presented in the literature of workflow (Combi and Pozzi, 2006) (Greg et al., 2004) (Johann et al., 2003) haven't consider too much about adaptively adjusting the scheduling strategies according to given condition. While, we believe the results reported in this paper is sufficient to warrant the development of such an adaptive scheduling approach.

5 RELATED WORKS

Our work is related to workflow time management and workflow scheduling.

5.1 Workflow Time Management

Analysis and Management of temporal information in workflow is by no means straightforward, its difficulty mainly comes form two aspects: the first is undetermined execution sequence of tasks and the second is variety of activities' time consumption.

In (Johann et al., 1999), Johann Eder et al investigated various time constrains in workflow. And, they presented a framework for computing activity deadlines so that the overall process deadline is met and all external time constraints are satisfied. Later on, he and Euthimios Panagos presented a method for incorporating detailed time information into workflow management systems(Eder and Panagos, 2000), their method is based on extend PERT(Pozewaunig et al., 1997). By adding elements like duration, deadline, earliest possible start time, earliest possible end time etc., their method can express different possibility of process execution time. In their paper, they also discussed issues in runtime handling of workflow time information.

In complex workflow models, the existence of conditional structures in the control flow may result in many execution paths, which makes it difficult to analyze task duration. Therefore, in (Johann et al.,

2003) (Eder and Pichler, 2002), Johann Eder and Horst Picler et al introduced the concept of time histogram. Their approach requires a well-formed workflow and probabilistic information about branching behavior of a process, then for each activity, possible execution time can be calculated. They also discussed ways to apply their approach to automatic process scenario like composite web-service process(Eder and Pichler, 2004). The probabilistic time management approach is also used by Martin Bierbaumer et al to analysis the phenomenon of unnecessary delay caused by fixed date constraints (Bierbaumer et al., 2005a) (Bierbaumer et al., 2005b). In order to assist participants of workflow appropriately select their work items, they use time histogram to calculate the delay time of ongoing process, and remind participants about possible delay according to the calculated result.

In addition to Johann Edier's works, there are some other researches that are related to workflow time management, In (Aalst and Reijers, 2003), Aalst et al use stochastic petri-nets to analysis workflow performance. and Carlo Combi et al also developed a set of models to address time constrains in organizational point of view(Combi and Pozzi, 2003b) (Combi and Pozzi, 2003a).

Previous work of workflow time management and time analysis concerns the variety of execution time caused by complex workflow model and branching probability. However, the variety of activity execution duration caused by interactions between human and workflow management system are not discussed. Our work focuses on this kind of variety.

5.2 Workflow Scheduling

Scheduling, however, despite its successful application in manufacturing fields, is not widely accepted in workflow.

Grego'rio Baggio Tramontina et al discussed some of the problems that prevent existing scheduling techniques from being used in workflow (Greg et al., 2004), in addition, they proposed a "Gauss and Solve" scheduling approach. Their approach consists of two steps, first, making a guess on the execution times and routes the case will follow, and second, solving the corresponding deterministic scheduling problem using a suitable technique. In the simulation, they used genetic algorithms as a means to schedule artificially generated cases. According to their result, if the error in guessing is bound by 30%, their approach is better than the commonly used FIFO rules regarding the number of late jobs. Besides, they envisioned the approach of using machine learning or statistical techniques to predict activity time consumption, however, in their paper, they didn't provide much detail. Our work can be viewed as a complementary effort to their work.

In (Combi and Pozzi, 2006), Carlo Combi and Giuseppe Pozzi focuses on temporalities in the conceptual organizational model and task assignment policies. They proposed a temporal organizational model, which extends traditional organizational models, to describe different temporal constrains of resources(Combi and Pozzi, 2003b) (Combi and Pozzi, 2003a), like availability constrains, and deadline constrains etc. Based on the description of these constrains, they designed a scheduling algorithm, which evaluates the priority of tasks according to the expected deadline for completion and expected duration. As a proof-of-concept, a running prototype implements the algorithms of the temporal scheduler for a WfMS.

Despite works that mainly concerns macro-level scheduling from workflow system's point of view, the work of Johann Eder et al(Johann et al., 2003) provides us another view on workflow scheduling: the personal scheduling. By admitting a commonly overlooked fact that people are actually the driving force of workflow(Moore, 2002), they changed their objective of scheduling from ordering cases in workflow system to assisting individual workflow participants. To meet this end, they provide workflow participants information about upcoming tasks so that they can proactively take measures to prepare for those tasks. Their approach is based on a probabilistic time management system(Eder and Pichler, 2002) which uses duration histograms to express the uncertainty of workflow time consumption.

Other work about workflow scheduling concerns scheduling in a single workflow instance, In (Senkul et al., 2002) (Senkul and Toroslu, 2005), Pinar Senkul and Ismail H. Toroslu proposed a architecture which provides a specification language that can model resource information and resource allocation constraints, and a scheduler model that incorporates a constraint solver in order to find proper resource assignments. Particularly, they use constraint programming to schedule workflows with resource allocation constraints.

6 SUMMARY

In this paper, we have discussed a K-level approach to anticipate activity time consumption in workflow management system. Our approach uses a supervised machine learning algorithm that is

applied to workflow event log. In the experiment on three enterprises, a total number of 869 activities were investigated and our approach separately achieved an average prediction accuracy of 80.27%, 70.93% and 61.14% with K = 10. In addition to presenting these results, we have analyzed the performance trend of different values of K, however the results is less positive. In addition, we also found that the operation time of workflow system has a positive influence on the performance of our approach.

We believe that our approach shows some promise for improving the current state of workflow scheduling. Our future plans include an investigation of additional sources of information, further development of adaptive scheduling approaches, and simulation using real data sets to test the applicability of workflow scheduling.

ACKNOWLEDGEMENTS

We are grateful to Tsinghua InfoTech Company for providing the workflow event-log data of their TiPLM system. This work is supported by the Project of National Natural Science Foundation of China (No. 60373011) and the 973 Project of China (No.2002CB312006).

REFERENCES

Aalst, W. M. P. V. D. & Reijers, H. A. (2003) Analysis of Discrete-time Stochastic Petrinets. *Journal of the Netherlands of Society for Statics and Operations Research,* 58.

Bierbaumer, M., Eder, J. & Pichler, H. (2005a) Accelerating Workflows with Fixed Date Constraints *24th International Conference on Conceptual Modeling.* Klagenfurt, Austria.

Bierbaumer, M., Eder, J. & Pichler, H. (2005b) Calculation of Delay Times for Workflows with Fixed-date Constraints. *Seventh IEEE International Conference on E-Commerce Technology, CEC 2005.*

Combi, C. & Pozzi, G. (2003a) Temporal Conceptual Modelling of Workflows *Conceptual Modeling - ER 2003.* Springer Berlin / Heidelberg.

Combi, C. & Pozzi, G. (2003b) Towards Temporal Information in Workflow Systems *Advanced Conceptual Modeling Techniques.* Springer Berlin / Heidelberg.

Combi, C. & Pozzi, G. (2006) Task Scheduling for a TemporalWorkflow Management System. *Thirteenth International Symposium on Temporal Representation and Reasoning, Time'06.*

Eder, J. & Panagos, E. (2000) Managing Time in Workflow Systems. IN FISCHER, L. (Ed.) *Workflow Handbook 2001.* Future Strategies Inc., USA.

Eder, J. & Pichler, H. (2002) Duration Histograms for Workflow Systems. *Working Conference on Engineering Information Systems in the Internet Context (IFIP TC8/WG8.1).* Kanazawa, Japan.

Eder, J. & Pichler, H. (2004) Response time histograms for composite Web services. *IEEE International Conference on Web Services, 2004*

Greg, Rio, B., Jacques, W. & Clarence, E. (2004) Applying Scheduling Techniques to Minimize The Number of Late Jobs in Workflow Systems. *Proceedings of the 2004 ACM symposium on Applied computing.* Nicosia, Cyprus, ACM Press.

Jiawei, H. & Kamber, M. (2001) Data *Mining : Concepts and Techniques* San Francisco, Morgan Kaufmann.

Johann, E., Euthimios, P. & Michael, R. (1999) Time Constraints in Workflow Systems. *11th International Conference on Advanced Information Systems Engineering: , CAiSE'99,.* Heidelberg, Germany, June 1999.

Johann, E., Horst, P., Wolfgang, G. & Michael, N. (2003) Personal Schedules for Workflow Systems. *Proceedings on Business Process Management: International Conference, BPM 2003, Eindhoven, The Netherlands, June 26-27, 2003.*

Ly, L., Rinderle, S., Dadam, P. & Reichert, M. (2006) Mining Staff Assignment Rules from Event-Based Data. *Lecture Notes in Computer Science Vol. 3812.*

Moore, C. (2002) Common Mistakes in Workflow Implementations. Giga Information Group, Cambridge MA(2002).

Pozewaunig, H., Eder, J. & Liebhart, W. (1997) ePERT: Extending PERT for Workflow Management Systems. *1 st East European Symposium on Advances in Database and Information Systems ADBIS ' 97.* St. Petersburg, Russia.

Quinlan, R. (1993) *C4.5: Programs for* Machine *Learning,* San Mateo, CA., Morgan Kaufmann Publishers.

Russell, N., Hofstede, A. H. M. T., Edmond, D. & Aalst, W. M. P. V. D. (2005) Workflow Resource Patterns. Eindhoven, Eindhoven University of Technology.

Senkul, P., Kifer, M. & Toroslu, I. H. (2002) A Logical Framework for Scheduling Workflows Under Resource Allocation Constraints. *Proceedings of the Twenty-eighth International Conference on Very Large Data Bases***,** 694-705.

Senkul, P. & Toroslu, I. H. (2005) An Architecture for Workflow Scheduling Under Resource Allocation Constraints. *Information Systems,* 30**,** 399-422.

Witten, I. H. & Frank, E. (2005) Data *Mining: Practical Machine Learning Tools and Techniques,* San Francisco, Morgan Kaufmann.