Response Time Histograms for Composite Web Services

Johann Eder, Horst Pichler
University of Klagenfurt, Austria
Institute for Informatics-Systems
eder@isys.uni-klu.ac.at, horst.pichler@uni-klu.ac.at

Abstract

Probabilistic duration representations can be used to forecast the response time of (composite) web services, based on empirical data of past executions or calculated from control flow structures. The capability to choose the fastest among similar services or to optimize services based on probable process execution times are only two possible application areas.

1. Introduction

The next step in the evolution of web services (WSs) are composite web services (CWSs) to support business processes within organizations as well as business processes spanning several organizations like supply chains. Thus the most critical need in companies will be to provide services with a better quality than their competitors. To assess the quality of service (QoS) it is necessary to define measures which are significant indicators for certain quality aspects, where expected or guaranteed process duration ranks among the most important characteristics. Slow WSs, invoked by a CWS, can have an disastrous impact on the overall process response time. Thus techniques are needed to predict the process duration based on the anticipated response time of participating WSs, enabling us too avoid these services or to optimize them for faster execution. These are established problems in workflow management, a closely related application area. In this paper we explain the basic ideas of our probabilistic time management approach and then we outline the necessary steps to apply these ideas on CWS-environments.

2. Duration histograms

Activities as part of workflows are capable of hiding complex business processes with greatly varying durations (also called complex activities). Thus simple duration representations like statistical mean or median fall short, as well as interval representations (minimum and maximum) or distribution functions (normal distribution) because they are to imprecise. They all do not take into account that the existence of conditional structures in the control flow may result in a non-evenly distributed duration with multiple peaks. Thus the concept of duration histograms (DHs) was introduced in [1] as structure to represent a complex distribution function on the duration of an activity or workflow.

Definition 1 (Duration Histogram) A duration histogram H is a set of tuples (p, t), where $0 < p \in \mathcal{R}$ is the probability and $t \in \mathcal{N}$ is the according time information. A duration histogram H is valid, if $\sum_{i=1}^{n} p_i = 1$ for $(p_i, t_i) \in H$. A cumulated duration histogram C, based on a duration histogram H, is a relation of n rows (c_i, t_i), (cumulated probability c, and time-information t), with $\sum_{i=1}^{n} p_i = 1$, and $c_i = \sum_{t_j \leq t_i} p_j$ for $1 \leq i \leq n$.

DHs can be generated by extracting statistics from execution logs or by calculation, according to the structure of an underlying business process (see section 3). Figure 1 shows an example of the graphical representation for a valid DH and its cumulated version. According to the definition we write H = {(0.01,7), (0.09,8), (0.17,9), (0.11,10), (0.05,11), (0.1,12), (0.22,13), (0.08,14), (0.04,15), (0.01,16), (0.03,20), (0.07,21), (0.02,22)}. Note that cumulated histograms can be calculated from regular histograms and vice versa. Taking a look at these values we can state that the bigger part of future executions will last between 7 and 16 with peaks at 9 and 13 plus some additional outliers around 21. According to the associated cumulated representation we can for example state that there is a likelihood of 88% that the activity will last 16 or less, thus we are enabled to determine the probability for a given threshold duration or vice versa.

3. Duration histogram calculation

It may be necessary to calculate the DH for a workflow or complex activity based on its process definition (structure and activity-DHs). [1] show how this can be accomplished for workflow definitions using se-

Proceedings of the IEEE International Conference on Web Services (ICWS'04)
0-7695-2167-3/04 $ 20.00 IEEE

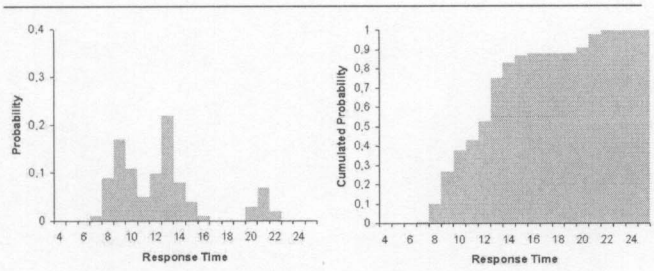

Figure 1. Response time histogram

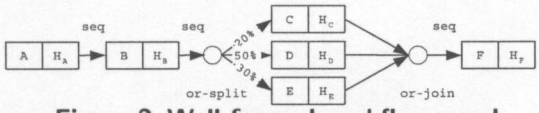

Figure 2. Well-formed workflow graph

quences, conditional routing (and-splits/joins), parallel routing (or-splits/joins) or iterations (loops). We use a (directed acyclic) workflow graph holding activities and the control flow dependencies between these activities. Additionally probabilistic information about the branching behavior of a process and DHs for activities must be provided. We only allowed well-formed workflow graphs, which have to adhere to certain constraints: a) every split-node has exactly one join-node of the same type, b) every path that originates from a split-node must merge in the associated join-node, and c) dependencies originating from activities between split/join-pairs must not leave their paths or connect to activities outside of these pairs. Figure 2) shows a simple graph consisting of activities (A,B,C,...), probabilistic duration information (DHs named H_A,H_B,H_C,...) and control flow structures. To calculate the DH of the workflow it is necessary to traverse the graph from the first to last activity, summing up and weighting the activity-DHs according to control structures and branching probabilities. The result of the calculation is a DH for the workflow holding tuples for every possible overall duration with its according probability.

4. Mapping the concepts

Due to the resemblance between the concepts workflow and CWSs, an adoption of the above presented seems reasonable. The following outlines our future research topics for a solution using executable processes in BPEL4WS.

Response time histograms: The basic idea remains the same: to forecast the duration of a process, based on probabilistic data. Workflow activities are mapped to invoked WSs and the duration is mapped to response times of these services, thus we call them response time histograms (RTH). The histogram concept will prove valuable as infrequent outliers will be located in the upper regions of the cumulated histogram. This enables us to cut them of by defining a threshold less than 100%. E.g. the response time of a service with a likelihood of 98% will not include very infrequent delays but still yield acceptable forecasts.

Adaption for BPEL4WS definitions: Our approach was designed for graphs representing well-formed workflow definitions. This is too strict for BPEL-WSCs defined with different types of primitive activities (invoke, receive, reply, etc.) and structured activities (sequence, link, flow, etc.). Therefore an extended graph-representation which allows non-well-formed structures must be found and the calculation algorithms must be adapted. Future research will build upon the pattern based analysis of BPEL4WS [2]. Another research topic will be the implementation of a parser to extract a non-well-formed graph from a BPEL-definition.

Gathering data for RTHs: In BPEL-environments empirical data can be extracted from the execution log (either existent or implemented as part of the process itself). However, to generate RTHs and branching probabilities it is necessary to log the system times of each request and each response as well as the control flow of all instances. Alternatively WS statistics could be stored and offered by trusted third parties.

Possible application scenario: WSCs are published as WSs to the outside world. These WSs have an additional interface (port type) *getResponseTime* which yields the expected service response time as RTH. The RTH can either be generated by accessing statistical data or calculated by applying the above mentioned algorithm on the structure of the composition. WSs invoked by the WSC are themselves equipped with *getResponseTime*. Dependent on the resulting RTH requestor and provider are enabled to make decisions (using this service or selecting an alternative one or optimizing the process for faster execution).

5. Conclusion

The prediction of overall process response times, based on the anticipated response time of participating WSs, be a valuable support and optimization criterion for service customers as well as for service providers.

References

[1] J. Eder and H. Pichler. Duration histograms for workflow systems. In *Proceedings of the Working Conference on Engineering Information Systems in the Internet Context*, Kanazawa, Japan, 2002.

[2] P. Wohed, W. van der Aalst, M. Dumas, and A. ter Hofstede. Pattern-based analysis of BPEL4WS. Technical report, Queensland Univ. of Technology, Brisbane, 2002.