Improving Performance of Internet Services Through Reward-Driven Request Prioritization

Presentation of a paper by A. Totok and V. Karamcheti from the IEEE International Workshop on Quality of Service, June 2006

Ken Salem

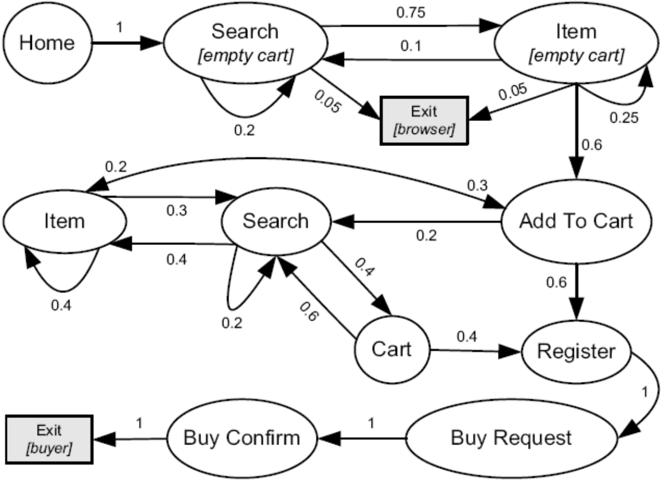
David R. Cheriton School of Computer Science
University of Waterloo

November 15, 2006

Problem Setting and Objectives

- web services
- differential session QoS targets
- argument: in some cases, target QoS should be determined dynamically, during the session
  - in on-line shopping, give buyers better QoS than browsers
  - given better QoS to sessions that visit revenue-generating advertising links

Customer Behaviour Model Graphs



Determining the Value of a Session

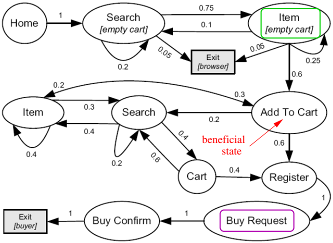
- associate a benefit (“reward”) with each type of state in the customer behaviour models
  - example: define benefit of “Add To Cart” state to be 1, benefit of all other states to be zero
  - each session either succeeds (exits normally) or fails because one of its requests is not served quickly enough.
  - define the benefit of a successful session to be the sum of the benefits of the states that are actually visited during the session
  - define the benefit of a failed session to be zero.
- associate a cost with each type of state, depending on the execution cost of that state’s request

Reward-Driven Request Prioritization

- the following are given in advance:
  - a set of customer behaviour models M_i, each of which describes a type of session
  - a prior probability p_i for each type of session
- each arriving HTTP request is associated with a particular active session
- when a request arrives, the RDRP mechanism estimates the expected benefit and cost from the request's session
- the expected session benefit and cost are used to prioritize request's access to resources. Higher benefit and lower cost give improved priority.

Estimating Future Session Cost and Benefit

- if we know that a request's session is of type M_i, we can estimate its future benefit (and cost):



- suppose we have a request R and session history H_R

$$\text{benefit}(R) = \sum_i \text{benefit}(R|M_i)\text{Prob}(M_i|H_R)$$

- future cost can be estimated the same way

Guessing a Request's Session Type

- Bayesian estimate:

$$\mathrm{Prob}(M_i|H_R) = \frac{\mathrm{Prob}(H_R|M_i)p_i}{\sum_j \mathrm{Prob}(H_R|M_j)P_j}$$

- $\mathrm{Prob}(H_R|M_i)$ is easy to determine in CBMGs and other Markov models

Prioritizing Requests

- assign a priority to each request

$$\text{priority}(R) = \frac{\text{attained plus predicted session benefit}}{\text{incurred plus predicted session cost}}$$

or

$$\text{priority}(R) = \frac{\text{attained plus predicted session benefit}}{\text{predicted session cost}}$$

- use priorities in the application server to regulate access to two resources:
  - execution threads
  - database connections

Comments

- this is a dynamic optimization mechanism
- no feedback is involved - it is assumed that accurate customer models are known in advance
- simple alternative (not considered) is to prioritize requests based only on attained benefits and incurred costs - how much benefit does prediction really bring?