Adaptive Web Data Extraction Policies

Giacomo Fiumara
Dip. di Fisica, Univ. di Messina
gfiumara@unime.it

Massimo Marchi
DSI, Univ. di Milano
marchi@dsi.unimi.it

Alessandro Provetti
Dip. di Fisica, Univ. di Messina
ale@unime.it

Abstract

Dynamo is a middleware that assists in generating informative RSS feeds out of legacy HTML Web sites. To make RSS feeds be timely and informative and to be scalable, Dynamo needs a careful tuning and customization of its polling policies which have been evaluated against news portals such as cnn.com.

Dynamo [1] is an experimental middleware for automated data collection and RSS delivery of data available from traditional HTML Web sites. To be effective (i.e., deliver information timely) and scalable Dynamo needs an accurate fine-tuning of its polling policy, i.e., the frequency at which it downloads and examines a given Web page. This poster describes the adaptive, ad-hoc polling policies that we have developed in the experimental setting of two popular (and busy) News portals: CNN Most Recent, (http://www.cnn.com) and ANSA Top News, (http://www.ansa.it).

To compute the frequency of the requests of updated Web documents we first make an estimate of the frequency. Then this estimate is compared to the real frequency with which Web documents are updated or newly generated. Both the estimate and the real times are used to compute a new estimate. That is:

\[ \tau_{n+1} = \alpha \tau_n + (1 - \alpha) t_n \]  

where \( \tau_{n+1} \) is the estimate at the \((n+1)\)-th iteration, \( \tau_n \) the estimate at the \(n\)-th iteration, \( t_n \) the real frequency at the \(n\)-th iteration. The parameter \( \alpha \), whose value stands in the interval between 0 and 1, represents the relative weight of the previous estimate w.r.t. the real frequency. In order to gain some insight, we have considered historical data (time series) collected as follows. Two popular news portals, maintained by respected and well-known press agencies, have been tracked continuously over several days to determine the rate of update of their front page. Those Web sites were www.cnn.com and www.ansa.it. The latter is in Italian only and obviously generates less traffic. However, the dynamics of the news would in principle be the same.

Given the time series collected from the two news portals described above, we computed the Mean Square Error (MSE), which measures the discrepancy between estimated and actually observed values. To find the parameter values that minimize MSE we proceeded applying the golden ratio method [3], which consists in comparing the values of the function to be minimized over three points for each considered interval. The value of \( \alpha \) that minimizes MSE over the available time series was the best estimate of the time elapsing between two updates of the page.

Today the intended application of Dynamo is in assisting on-line communities to add RSS channels to their existing community portals [2].

References

