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 helps in generating informative RSS feeds out of legacy HTML Web sites. To produce timely and informative RSS feeds, and to be scalable, Dynamo needs a careful tuning and customization of its polling policies which have been evaluated against frequently-updated news portals.

Dynamo [1] is an experimental middleware for automated data collection and RSS delivery of data available from traditional HTML Web sites.

To locate the relevant data in the plain HTML pages, the architecture requires the insertion of some meta tags in the commented text. Such annotated HTML documents are then routinely pulled by our Web service, which then aggregates the data and serves them over several channels, e.g. RSS 1.0 or 2.0.

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 polices that we have developed in the experimental setting of two popular 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 by which the sites get updated. Next, our 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:

$$\tau_{n+1} = \alpha\tau_n + (1-\alpha)t_n \qquad (1)$$

where τ_{n+1} is the estimate at the $(n+1)$-th iteration, τ_n the estimate at the n-th iteration and t_n the *real* frequency at the n-th iteration. The parameter α, 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.

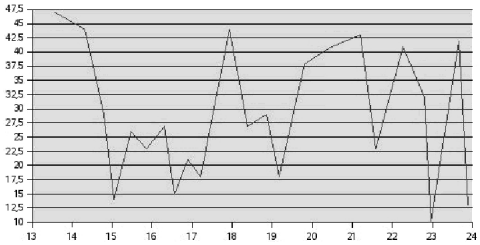

Figure 1. A time series from ANSA news site

The *cnn.com* and *ansa.it* (in Italian) 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.

The time series collected from the two news portals were used to compute the Mean Square Error (MSE), which measures the discrepancy between estimated and actually observed values. To find the parameter values that minimize MSE we applied the *golden ratio* method [3]. It consists of comparing the values of the function to be minimized over three points for each considered interval. The value of α that minimizes MSE over the available time series was the best estimate of the time elapsing between two updates of the page. Today, Dynamo is being deployed in the challenging scenario of on-line communities, i.e., to add RSS capabilities to existing community portals [2].

References

[1] S. Bossa, G. Fiumara, and A. Provetti. An architecture for policy-based rss polling. In *Proc. of Workshop from Objects to Agents (WOA06)*, 2006.

[2] F. DeCindio, G. Fiumara, M. Marchi, A. Provetti, L. Ripamonti, and L. Sonnante. Aggregating information and enforcing awareness across communities with the dynamo rss feeds creation engine: preliminary report. In *Proc. of COMINF06, an OTM 2006 Workshop*, volume 1, pages 227–236. Springer LNCS 4277, 2006.

[3] W. Press, S. Teukolsky, W. T. Vetterling, and B. Flannery. *Numerical Recipes in C, The Art of Scientific Computing, second edition*. Cambridge University Press, Cambridge, 1999.

Eighth IEEE International Workshop on
Policies for Distributed Systems and Networks (POLICY'07)
0-7695-2767-1/07 $20.00 © 2007 IEEE