Quality-Aware Retrieval of Data Objects from Autonomous Sources for Web-Based Repositories

Houtan Shirani-Mehr [1], Chen Li [2], Gang Liang [3], Michal Shmueli-Scheuer [4]

Department of Information and Computer Science, University of California, Irvine
Irvine, CA 92697, USA
{ [1]hshirani, [2]chenli, [3]liang, [4]mshmueli}@uci.edu

I. Introduction

As the Web becomes the dominant medium for information delivery and commerce, an increasing number of applications have their services heavily rely on data repositories in utilizing information collected from various Web sources. For example, comparison shopping service providers such as MySimon.com search and index products from a large number of merchant Web sites, and allow users to compare prices from competing stores by collecting a substantial amount of information from many autonomous sources on the Web.

Building a high-quality repository is challenging especially when remote Web sources are dynamic: new data objects are added, and existing data objects are deleted or updated frequently. In addition, these Web sources might not be collaborative in the sense that applications will not be notified for any changes. Therefore, applications have to query remote sources periodically to retrieve the latest information. As a consequence, the local repository cannot mirror the remote sources completely. Hence, we need *quality metrics* to quantify the goodness of the local repository with respective to the remote sources.

The goal of this paper is to develop a framework for designing good data repositories for Web applications. The central theme of our approach is to employ statistical methods to predict quality metrics. These prediction quantities can be used to answer important questions such as: How soon should the local repository be synchronized to have a quality of at least 90% precision with certain confidence level? Suppose the local repository was synchronized three days ago, how many objects could have been deleted at the remote source since then?

We make three main contributions. First, we conducted an empirical analysis on datasets collected from 6 Web sites in 4 different domains. Our study provides valuable insights into the dynamics of Web sources (Section II). Second, we propose a survival-arrival approach for analyzing behaviors of individual objects in remote Web sources (Section III). The changes in the remote sources are decomposed into two independent processes, arrivals and removals, and the estimation of various quality metrics are derived. Third, we develop an adaptive framework to track the dynamism of remote sources (Section IV). We use a time-series analysis to model decay properties of quality metrics. The adaptivity of the framework allows us to detect and react to changes at remote sources automatically.

The complete results of our work can be found at [1].

II. Objects and Quality Metrics

A Web source is viewed as a collection of objects. Each object (e.g., a car) can have multiple attributes, such as vehicle ID, make, model, price, etc. We have a local repository that stores the object information retrieved by either following links of Web pages at various Web sources, or posing queries using their search forms. There are three possible events during the life span of an object at a Web source: *insertion*, *deletion*, and *update*. The remote Web source has to be re-crawled periodically due to these events.

In the literature, several quality measures have been used for individual objects, such as age and freshness [2]. In this paper we consider two important quality metrics to describe the overall quality of the data at the local repository: *precision* and *recall*. Formally, let $L(t)$ and $R(t)$ be the set of objects at the local repository and remote Web source at time t, respectively. We define

$$\text{precision}(t) = \frac{|R(t) \cap L(t)|}{|L(t)|}; \qquad \text{recall}(t) = \frac{|R(t) \cap L(t)|}{|R(t)|}. \tag{1}$$

The quality metrics above are defined with respect to a reference time t_0, i.e., the time of the last synchronization of the local repository. We assume that the local repository is fully synchronized each time, i.e., $L(t_0) = R(t_0)$. Both precision and recall can also be defined over a subset of objects, such as "BMW cars made after 2002."

An Empirical Study: We have conducted a thorough empirical study on real data sets collected from 6 web sites in 4 domains on a daily basis: cars, job postings, books, and forums. An overview of all the data sources is summarized in Table I. For all the sources, we relied on the URL of each page to extract a unique identifier of the corresponding object.

Our empirical study shows that quality metrics decay over time, and the decay rate varies from site to site, from category to category, and from time to time. An application needs to take all these factors into consideration when modeling the Web dynamics in order to determine a good synchronization schedule.

	Website (category #)	Period (# of months)	object #
Car	Car Web source 1 (10)	10 (2006/01 - 2006/12)	43,597
Car	Car Web source 2 (10)	13 (2005/10 - 2006/11)	37,173
Book	Book Web source 1 (3)	3 (2005/08 - 2005/12)	3,349
Job	Job Web source 1 (2)	3 (2006/12 - 2007/03)	2,959
Job	Job Web source 2 (7)	4 (2006/06 - 2006/10)	31,223
Forum	Forums Web source 1(4)	2 (2006/12 - 2007/02)	25,370

TABLE I
CRAWLED WEB SOURCES.

III. MODELING OBJECT BEHAVIORS

In this section we propose a general framework in which we use an arrival-survival approach to modeling the behaviors of individual objects at a Web source. Based on the behaviors, the application can derive quality metrics and use them to decide the crawling schedule and predict the quality metrics.

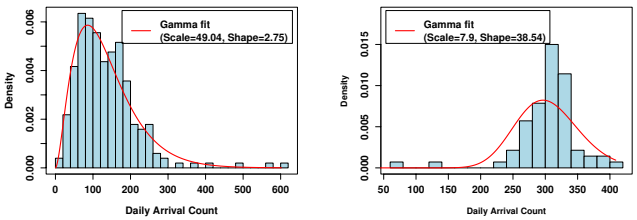
There are several advantages associated with this approach. First, understanding behaviors of individual objects can give us insights into the dynamics at the source. Second, censored information can be easily incorporated into the analysis. Third, it is natural to incorporate attributes of objects (such as car make) in the survival model to improve the estimation accuracy.

A. Arrival Analysis

The arrival analysis is to analyze the behavior of arrival counts at the remote source. Figure 1 shows the histograms of the daily-arrival count for two crawled Web sites. Given the shape of the histograms, Gamma distributions are used to model the arrival counts. Its density function is

$$f(x;\lambda,\theta) = \frac{x^{\lambda-1}}{\Gamma(\lambda)}\theta^{-\lambda}e^{-x/\theta}, x > 0,$$

where $\lambda > 0$ is the *shape parameter*, $\theta > 0$ is the *scale parameter*, and $\Gamma(\cdot)$ is the Gamma function. The parameters can be estimated by the maximum likelihood method based on collected data [3].



(a) Car Web source 1. (b) Job Web source 2.

Fig. 1. Arrival analysis.

The expected value of a Gamma distribution, i.e., the expected arrival counts, is $\lambda\theta$. Under the assumption that the daily arrival counts are independent, the summation of Gamma distributions is still a Gamma distribution. Hence, we can compute the distribution of arrival counts during some period by looking up a Gamma table.

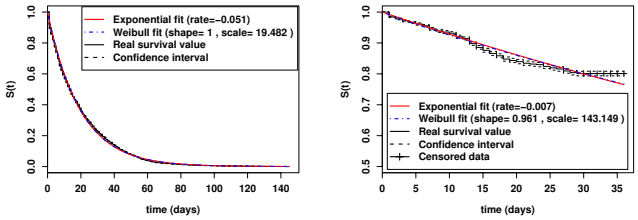
B. Survival Analysis

Survival analysis [4] is a statistical approach on studies of the failure events or deaths of patients. Its idea is to analyze objects over time, and study the pattern of failure events of interests. In our study, an object can be viewed as a "patient," and the removal event is a "death" event.

Let T be the life span of an object. The *survival function*, $S(t) = P(T > t)$, is the probability that the life span of an object is greater than t. Exponential and Weibull distributions are the two widely used parametric survival functions. The survival function of Weibull is

$$S(t) = e^{-(at)^k},$$

with exponential being a special case ($k = 1$), i.e., Weibull has two parameters while exponential has only one. Both survival functions are fitted using our datasets, and the curves are shown in Figures 2. Due to the negligible difference between distributions, we use an exponential distribution as the underlying survival function. Given the observed data, we can see that the maximum likelihood estimate of the parameter $\hat{a}$ is simply the inverse of the sample mean. The estimated parameter can be then used to predict the number of objects in a future time.



(a) Car Web source 1. (b) Job Web source 2.

Fig. 2. Survival analysis.

C. Deriving Quality Metrics

Using the arrival and survival models, we can estimate the decay of quality metrics over time. We can derive formula for both precision and recall under the assumption that the arrival and survival processes are independent [1].

IV. MODELING SYSTEM ADAPTIVITY

In some cases the object changing frequencies at the source can also change over time. We now propose an adaptive framework, in which we can automatically adjust to such changes. We use time-series methods [5] due to the natural time component in our problem. In this framework, we first collect enough data to build a time series model for the Web source to decide the next synchronization time. After each

synchronization, we adjust the parameters in the model based on the newly collected data.

The static framework in Section III is an indirect model in the sense that it is just a by-product of the modeling of individual objects in the arrival-survival models. In contrast, the adaptive model is a direct model, since the quality metric is directly modeled by a time series model. A direct approach usually offers better predictions but lacks the insights of the process itself, while an indirect one generally has better insights but worse predictions due to the extra complication.

A. Time Series Analysis

Suppose the local repository is synchronized at time t_0. Let $q(t_0, t)$ be the quality metric (recall or precision) at time t, assuming there is no synchronization between t_0 and t. Define $d(t)$ as the decay rate of the metric over period $[t-1, t)$:

$$d(t) = \log(q(t-1, t)).$$

Our empirical results show that the quality metric is of an exponential form; hence, the quality metric at time t, $q(t_0, t)$, is related to the accumulated decay through

$$q(t_0, t) = \exp\left(\sum_{i=t_0+1}^{t} d(i)\right). \quad (2)$$

Notice the decay rate $d(t)$ is independent of the reference date t_0, so we can impose a time series model on $d(t)$, then the above equation links the decay rate with the quality metric of interest. It also implies that the future quality metric can be predicted as long as estimates of $d(t)$ are available.

AR Model. Even though any model can be used within the general adaptive framework, in practice it is important to choose a good time series model. In this paper, we use an autoregressive (AR) model [6] for the decay rate sequence $d(t)$. Note that the choice of the model is tightly coupled with the application, and should be selected on a case-by-case basis.

Formally, given an *order parameter* k, an AR(k) model can be written as

$$d(t) = \alpha + \sum_{i=1}^{k} \beta_i d(t-i) + \varepsilon_t, \quad (3)$$

where α and β_i's are coefficient parameters, and ε_t's are white noises with variance σ^2. In other words, the quality decay rate at t is a linear combination of the decay rates in the last k periods plus a random error term under an AR(k) model. The AR model is one of the most widely used time series methods due to its simplicity, flexibility, and power. It can be tailored to reflect various properties of the remote source. For instance, if weekly variation is observed, we can add a 7-th order term, $\beta_7 d(t-7)$, to capture such a weekly trend.

In AR analysis, the order parameter k can be determined based on the observed data. It is called *model identification* [5]. For example, AR(3) is appropriate to model the decay rate of recall of Car Web source 1.

Given k and the collected data, a method using maximum likelihood can be applied to obtain parameter estimate $(\hat{\alpha}, \hat{\beta}_i, \hat{\sigma}^2)$. We can obtain the estimated future decay rates. Using Equation 2 we can obtain an estimate of the quality metric $q(t_0, t)$. The confidence interval of $d(t)$ then $q(t_0, t)$ can be obtained by the delta method [3].

B. Adaptive Autoregressive Model

Here we present a novel adaptive approach to modifying an AR model based on newly observed information. We expect the AR model will slowly adapt to the current dynamics of the remote source. The principle of the parameter-updating scheme is to update AR parameters such that the estimated decay is pushed towards the newly observed one.

In order to keep a balance between the historic and the newly observed information, a learning rate parameter τ ($0 < \tau < 1$) is introduced to control the speed of adaptation. In practice, the learning rate parameter was set to be a small number $\tau = 0.1$ or 0.2 for a slow adaption.

An example of our experiments is shown in Figure 3. The remote Web source was crawled once every 12 days, and we predicted the recall metric at the end of each cycle. The AR model is trained on the data of 40 days starting from April 25, 2006, and the result was tested for a 60-day period from September 25, 2006 to November 24, 2006. The learning rate is set to $\tau = 0.2$.

In the plot, the recall curve of the adaptive AR model is against the true curve and that of an AR model with static parameters. There is a 78% reduction (from 0.136 to 0.030) in terms of sum of square errors (SSE) for the adaptive approach over the static one. We can see that the adaptive model yielded more accurate estimates than its static counterpart.

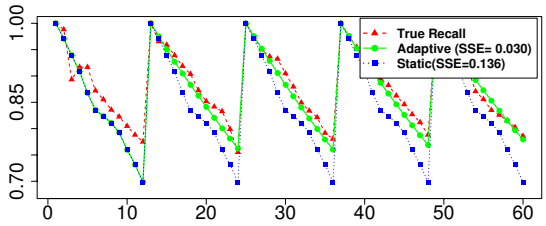


Fig. 3. Adaptive versus static AR decay models for recall of the car Web source 1.

REFERENCES

[1] H. Shirani-Mehr, C. Li, G. Liang, and M. Shmueli-Scheuer, "Quality-aware retrieval of data objects from autonomous sources for web-based repositories (full version)," Department of Computer Science, UC Irvine, Tech. Rep., 2007.

[2] J. Cho and H. Garcia-Molina, "Synchronizing a database to improve freshness," *SIGMOD*, 2000.

[3] J. Rice, *Mathematical Statistics and Data Analysis*, 2nd ed. Duxbury Press, 1994.

[4] E. T. Lee and J. W. Wang, *Statistical Methods for Survival Data Analysis*. Wiley, 2003.

[5] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. Springer-Verlag, 1991.

[6] G. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.