Efficient Algorithms for Predicting Requests to Web Servers

Edith Cohen, Balachander Krishnamurthy, and Jennifer Rexford
AT&T Labs–Research, Florham Park, NJ
{edith,bala,jrex}@research.att.com

Abstract— **Internet traffic has grown significantly with the popularity of the Web. Consequently user perceived latency in retrieving web pages has increased. Caching and prefetching at the client side, aided by hints from the server, are attempts at solving this problem. We suggest techniques to group resources that are likely to be accessed together into volumes, which are used to generate hints tailored to individual applications, such as prefetching, cache replacement, and cache validation. We discuss theoretical aspects of optimal volume construction, and develop efficient heuristics. Tunable parameters allow our algorithms to predict as many accesses as possible while reducing false predictions and limiting the size of hints. We analyze a collection of large server logs, extracting access patterns to construct and evaluate volumes. We examine sampling techniques to process only portions of the server logs while constructing equally good volumes. We show that it is possible to predict requests at low cost with a high degree of precision.**

I. Introduction

The popularity of the World Wide Web is largely responsible for the dramatic increase in Internet traffic during the past few years. A growing number of users are accessing "Web pages," each consisting of one or more resources, including text, images, and applets. These resources must be retrieved from the Web server and transferred across the network, unless the contents are available in the browser cache or at a shared proxy cache. Retrieving a resource typically requires the browser to establish a TCP connection and generate a request with the appropriate URL. The server transmits a response message with the contents of the resource. The process of opening the TCP connection, submitting the request, and transmitting the response introduces perceivable latency and consumes bandwidth along the path from the server to the client. By storing popular resources closer to the client sites, proxy caches can reduce response latency, server load, and network traffic. Proxies can also perform other performance-enhancing functions, such as prefetching, cache validation, and TCP connection management, on behalf of clients.

The proxy's policies for managing cached resources and TCP connections rely on assumptions about client access patterns. For example, cache replacement strategies try to identify which resources are least likely to be accessed in the near future. Similarly, prefetching decisions are based on predictions of future requests, to avoid generating unnecessary traffic. The server, by receiving requests from a large number of proxies and clients, is in a unique position to gauge the popularity of its resources, and the likelihood that certain resources are accessed together. For example, the server can speculatively disseminate resources to the proxy [1], guide proxy prefetching decisions [2–5], guide cache-replacement decisions [6,7], and improve cache coherency by sending information about modified resources [8,9]. The overhead can be minimized by focusing on resources that are likely to be requested in the future.

In earlier work, we introduced an end-to-end framework for server response messages to include hints (in groups called *volumes*) about future requests, customized to the requesting proxy [10]. Here, we present a detailed design and evaluation of algorithms for grouping resources that are likely to be accessed together. In contrast to previous research on server hints, we discuss the theoretical aspects of optimal volume construction, and develop efficient heuristics that can be tuned to specific proxy applications. The algorithms estimate which resources will be requested within a time interval in the future. Tunable parameters allow the algorithms to strike a balance between the desire to predict as many accesses as possible with the need to control the number of false predictions, while limiting the number of hints that are sent to the proxy. We also consider simple techniques for reducing the computational overheads for constructing and evaluating the server volumes. The key contributions of the paper are:

- **Problem formulation:** We present a precise theoretical formulation of the volume-construction problem, based on the time interval for predicting future access and three important performance metrics. Based on this formulation, we describe several optimization problems that require balancing the trade-offs between the number of accesses that are predicted in advance, the number of false predictions, and the total number of hints sent to the proxy.
- **Greedy heuristic:** After proving that the optimization problems are NP-complete, we present a greedy algorithm for constructing volumes. The greedy volumes are constructed using a notion of marginal utility of adding a resource to a volume given the current set of volumes, until an overall threshold utility value is reached. Although the greedy volumes perform very well, volume construction incurs a significant computational overhead. Hence, we consider faster alternatives such as *pairwise volumes*.
- **Pairwise volumes:** Expanding on our earlier paper [10], we present a detailed treatment of heuristics for constructing volumes, based on the likelihood that pairs of resources are accessed together. Similar to previous work [1, 2, 5], we define a probability threshold for deciding which resources to include in a volume. We describe an efficient

algorithm for computing access probabilities, and elaborate on several techniques for identifying ineffective hints that do not usually generate new predictions.

• **Multifactor experiments:** To evaluate our algorithms, we characterize the interaction between the various parameters and performance metrics on a collection of server logs. One set of logs are from AT&T's EW3 Web hosting service and the rest were from past experiments [10]. The experiments provide guidelines for setting the algorithm parameters for different proxy applications. We also show that our approach of computing access probabilities and removing ineffective predictions is competitive with the more complex greedy heuristic. The pairwise volumes were also used as a seed for the greedy volume construction to reduce the overhead of the greedy algorithm.

• **Reduction of processing requirements:** Processing large server logs requires significant computational resources. We show that it is possible to substantially reduce the amount of data processing involved in generating volumes. The overheads of constructing and evaluating volumes can be reduced by operating on samples of the server logs. Generating volumes from a shorter duration of the log, or a fraction of the clients/proxies, is competitive with computing across the entire collection of accesses.

Section II describes the volume construction problem and its key parameters. Section III presents an efficient algorithm for computing the likelihood that pairs of resources are accessed together, including techniques for improving the effectiveness of the predictions and reducing computational complexity. The optimization problems being NP-complete, we present a greedy heuristic for computing volumes from server logs. The simulation experiments on a collection of server logs in Section IV systematically characterize the interaction between the various parameters, and investigate ways to reduce the amount of data that must be processed to construct and evaluate server volumes. Section V concludes the paper with a discussion of future research directions. The NP-completeness proof of optimal volume construction is presented in Appendix A, and the performance bounds for the greedy and pairwise heuristics are presented in Appendix B.

II. Volume Construction Problem

In this section we describe our prediction model of future requests and discuss the various optimization criteria of interest. The greedy algorithm developed to obtain an approximate solution is then presented.

A. Prediction Model

Studies of Web access patterns have shown that recent requests are useful for predicting which resources will be accessed in the near future. The exchange of request and response messages can be enhanced by the server predicting resources (and attributes, such as Last-Modified Time and size) the client will access in the future. Many of these accesses stem from navigating through the hypertext links of recently-retrieved Web pages. We consider techniques that generate hints based only on the current request rather than keeping detailed information about past accesses, since the large number of requesting proxies and contacted servers would introduce substantial overhead.

The server has the history of past accesses including the triple (s, a, t), where client s requests resource a at time t. This information can be extracted from the server logs and post-processed to compute associations between resources. To reduce the processing overhead, these computations can be performed on a subset of clients and/or a smaller time interval. Volumes can be computed periodically to ensure that the server generates hints based on recent information from the logs. Each resource a has a *volume* v_a, which is a list of resources that are likely to be requested in the near future after an access to a. Effective and efficient algorithms for computing the volumes $\{v_a\}$ based on the sequence of triples (s, a, t) are needed. This formulation has guided recent studies of Web prefetching [1, 2, 5].

A prediction of an access to a resource b is most useful if it occurs some minimum time τ and maximum time T before the request. If the prediction is generated too close to the predicted access (e.g., less than half a second in advance), the proxy does not have sufficient time to act on the server hint. For example, a proxy is likely to begin downloading embedded images immediately upon receiving the enclosing page, lowering the value of predictions for these resources. On the other hand, predictions that arrive too far in advance (e.g., one hour before the predicted access) may also be of limited value. In practice, it may be difficult to predict accesses very far into the future, since client requests often depend on recent history. Hence, we consider that an access to resource b at time t was *predicted* if the proxy receives at least one hint for b during the time interval $[t-T, t-\tau)$. Similarly, a hint generated at time t for resource b is a *true prediction* if and only if the proxy receives at least one request for b during $[t+\tau, t+T)$.

B. Optimization Criteria

The hints returned to the proxy should maximize the proportion of requests that are predicted in advance, while minimizing the proportion of hints that are incorrect. A resource that appears in multiple hints in the same time interval is treated as a single prediction. To reduce the overhead of transmitting hints and the receiving proxy processing them, the server should minimize the total number of hints. We consider three key performance metrics:

• *Recall:* Likelihood that a resource requested by the client has appeared in at least one hint in the last $[\tau, T)$ seconds.

• *Precision:* Likelihood that a predicted resource is accessed by the client in the next $[\tau, T)$ seconds.

• *Average Hintsize:* Average number of hints in a message from the server to the client.

These metrics suggest several problems such as maximizing recall subject to an upper bound on hintsize, minimizing hintsize subject to a lower bound on recall, maximizing recall subject to a lower bound on precision, and maximizing precision subject to a lower bound on recall.

The choice of an optimization criterion depends on how the proxy uses the server hints [10]. For example, a prefetching application may require high precision to avoid the high cost of incorrect predictions. In contrast, a cache coherency policy may demand high recall, to receive the Last-Modified Time information for a large portion of the cached resources, even if the hints include excess information. Optimal volumes according to the first two criteria generate tradeoffs where recall increases with hintsize. Optimal volumes constructed according to the last two criteria are such that precision increases as recall drops. These desirable properties, consistent with user's expectations, are guaranteed for greedy volumes and as we shall see in Section IV, they generally hold for other heuristics.

These optimization problems are NP-complete, as proven in Appendix A. For efficient construction of server volumes, we offer heuristics based on computing the pairwise probability that an access to a resource a is followed by an access to resource b. The basic pairwise heuristic includes b in v_a whenever the implication probability exceeds a threshold (see Section III). The more refined and theoretically-motivated greedy algorithm iteratively recomputes these probabilities, based on past decisions (Section II-C). The greedy algorithm always performs within factor of $O(\log n)$ from optimal where n is the number of requests, whereas the pairwise heuristic has performance ratio of $\Theta(n)$ on worst case sequences, as discussion in Appendix B. In practice, the greedy algorithm outperforms the pairwise heuristic, but both techniques can achieve high precision and high recall with a reasonable average hintsize, as shown in Section IV. Greedy, however, incurs significant computational overhead; hence the need for the pairwise heuristic.

C. Greedy Algorithm

The greedy heuristic iteratively augments the sets $\{v_a\}$, based on the utility of the remaining resources. Each iteration performs a single pass on the server log to evaluate the *marginal utility* $u(b,a)$ of adding resource b to volume v_a, given the current set of volumes. The definition of $u(b,a)$ depends on the optimization criteria. For example, for the recall vs. hintsize trade-offs:

$$u(b,a) = \frac{\text{number of additional predicted requests for } b}{\text{total number of requests for } a}$$

captures the ratio of "benefit" (true predictions) and "cost" (total predictions), and has provable performance properties. Figure 1 provides a description of the greedy algorithm. Volumes are constructed subject to a threshold U on the marginal utilities. At the end of each iteration, for each resource b, we find the largest value of $u(b,a)$. If $u(b,a) \geq U$, we add b to v_a. The algorithm terminates when $u(b,a) < U$ for all pairs (b,a).

The algorithm starts with a set of empty volumes. In each iteration, each resource b is added to at most one volume v_a. Each item in v_a is labeled with the respective marginal utility. For each resource b, we maintain a set $\texttt{candidates}(b)$ of *candidate predictors*, consisting of all resources a such that we consider placing b in v_a. The running time of the greedy heuristic hinges on the number of iterations, which is bounded by the size of the largest candidates set (which is at most the number of resources).

Input:
  A threshold U
  For each resource b, a set $\texttt{candidates}(b)$

Iterate:
  Compute $u(b,a)$ for all b and $a \in \texttt{candidates}(b)$
  For every resource b:
    Let $a \in \texttt{candidates}(b)$ be the resource
      that maximizes $u(b,a)$.
    If $u(b,a) \geq U$, $v_a \leftarrow v_a \cup \{b\}$
    Remove from $\texttt{candidates}(b)$ all a
      with $u(b,a) < U$.
Until: $\forall b$, $\texttt{candidates}(b) = \emptyset$

Fig. 1. Greedy heuristic

The efficiency of the algorithm can be improved by limiting the size of the initial candidate set. A straightforward initialization places *all* resources in each of the sets $\texttt{candidates}(b)$, resulting in a quadratic number of candidates. We carefully initialize by including only pairs with initial $u(b,a) \geq U$; such pairs are found efficiently using our techniques for constructing pairwise volumes (see Section III). It is also possible to remove resources from the candidate set during iterations of the greedy algorithm. The marginal utility $u(b,a)$ is non-increasing with respect to the first optimization criteria. Hence, resource a can be removed from $\texttt{candidates}(b)$ as soon as $u(b,a) < U$.

III. Constructing Pairwise Probabilities

The greedy heuristic constructs high quality volumes at potentially high computation cost. We now discuss the cheaper pairwise approach. After describing how these access probabilities are computed, we introduce heuristics to improve accuracy and reduce volume size by determining which predictions are most effective.

A. Pairwise Counters

To avoid the complexity of the greedy algorithm, we construct volumes based on the probability that resources a and b are accessed together, independent of the predictions generated by other resources. We define $p_{b|a}$ as the proportion of requests for resource a that are followed by a request for resource b by the same client within $[\tau, T)$ seconds. Resource b is included in volume v_a if $p_{b|a}$ is greater than or equal to a threshold probability p_t. The value of the threshold probability depends on the desired trade-off between precision, recall, and hintsize. The server accumulates a counter of the number of accesses to resource a (c_a), and the number of times an access to a is followed by an access to b within $[\tau, T)$ seconds ($c_{b|a}$), where $p_{b|a} = c_{b|a}/c_a$.

Nuances arise in handling repeated accesses to resources a and b within a small interval of time. For example, when resource b is requested multiple times after an access to resource a (e.g., $abbb$), we count this as a single occurrence of the pair of resources. Although the counter c_b increases with each access to b, the counter $c_{b|a}$ is incremented just once. Otherwise, the ratio $c_{b|a}/c_a$ could be larger than 1. Counting predictions more than once is somewhat misleading, as it artificially inflates the probability $p_{b|a}$. For example, if one request for a is followed by three requests for b, and two other requests for a are not followed by requests for b, we assign $p_{b|a}$ to $1/3$, not $3/3$. This is crucial for the precision metric, since it accounts for false predictions generated on requests for a not followed by b. Repeat accesses to resource a (e.g., $aaab$) are handled differently. We credit each occurrence of a with generating a prediction, as long as b occurs within the subsequent prediction interval[1]. For example, if three accesses to a are followed by one access to b, we assign $p_{b|a}$ to $3/3$, not $1/3$.

The counters $c_{b|a}$ and c_a can be computed by sequencing through the requests in the server log. The probabilities are computed by sequencing through the triples (s, a, t), where client s requests resource a at time t. The algorithm processes the requests one client at a time. For a given client, we maintain a FIFO list of requests from the preceding T seconds. In processing a request for b, we first remove requests that occurred more than T seconds ago from the head of the FIFO. Then, we sequence from the tail of the list and increment $c_{b|a}$ for each resource a in the list, after skipping any requests that occurred less than τ seconds ago. Sequencing continues till the end of the list, unless we encounter a previous request for b. Else, the algorithm would credit some resource a with multiple predictions of resource b in a single time interval. Suppose a client has a sequence of requests for $axbyb$ in a small interval of time. The first request for b increments $c_{b|a}$ and $c_{b|x}$; the second access to b increments $c_{b|y}$, but not $c_{b|a}$ and $c_{b|x}$.

Computing the pairwise probabilities $p_{b|a}$ could potentially require a quadratic number of counters $c_{b|a}$, though many pairs of resources rarely occur together. Storage complexity can be reduced using heuristics that create counters only for pairs of resources that are likely to be accessed together [11]. For example, a counter $c_{b|a}$ could be created only if b is reachable from a through a sequence of up to $h \geq 1$ hypertext links. Alternatively, the counter could be created if the URLs for a and b match in the first $k \geq 0$ directories. Finally, randomization could be used to generate counters. While processing the log, we create counter $c_{b|a}$ with probability $\frac{C}{p_t c_a}$ upon encountering the pair, where C is some constant that governs tradeoffs between the number of counters and the likelihood of missing important pairs.

B. Volume Thinning

The implication probability $p_{b|a}$ estimates the value of including b in volume v_a, independent of all of the other resources that may predict b. But local decisions can reduce precision. Suppose that nine out of every ten accesses to a are followed by an access to b, resulting in $p_{b|a} = 9/10$. But, suppose these nine accesses to a occur with other resources x and y that also have high implication probabilities for b (e.g., sequences $xyab$ where $p_{b|x}, p_{b|y} \geq p_t$). If the volumes v_x, v_y, and v_a all include resource b, then a never actually generates a *new* prediction for b (every occurrence of a is preceded by another resource that generates a prediction for b). In fact, including b in volume v_a actually *decreases* precision, since a *mis*predicts b one-tenth of the time. In contrast, the greedy algorithm avoids placing b in volume v_a by reevaluating the utility of a as predictor after adding b to volume v_x or v_y.

To increase precision and reduce hintsize, we consider heuristics for removing ineffective predictions, without incurring the high computational complexity of the greedy algorithm. After computing the probabilities $p_{b|a}$ and applying p_t to construct an initial set of volumes, we perform a second pass through the log to measure the effectiveness of each predictor. The various heuristics differ in how they distribute "credit" for each prediction across the multiple resources that precede b. One possible heuristic favors the *first* resource that generates a prediction for b, within the $[\tau, T)$ interval [10] (in $xyab$, credit x for prediction); an alternate approach credits the *last* resource in the $[\tau, T)$ interval (i.e., a), under the assumption that the current access to b is more correlated with the most recent request from the same client than with the earlier accesses.

These heuristics count the number of times $e_{b|a}$ that resource a (with $p_{b|a} \geq p_t$) is credited with generating a prediction for b. Then, resource b is included in volume v_a if the proportion of predictions exceeds a threshold e_t (i.e., if $e_{b|a}/c_{b|a} \geq e_t$). Hence, a resource b is included in volume a only if $p_{b|a}$ and $e_{b|a}$ are high enough, relative to p_t and e_t, respectively. Although these heuristics attempt to locate the most effective predictors, they thin the volumes without reevaluating the dependencies between resources. For example, suppose resources x, y, and z occur in two types of sequences xyz and yxz. Then, both heuristics would credit both x and y with half of the predictions of z. If the threshold e_t exceeds 0.5, neither v_x nor v_y would include resource z, even though including z in either volume would have successfully predicted all accesses to z. Or, if $e_t \leq 0.5$, the heuristics would place z in both volumes, even though one of the two predictions is always redundant. In comparison, the greedy algorithm would select one of the two resources (x or y) to include z in its volume, and then reevaluate the utility of the other resource.

IV. Performance Evaluation

After discussing the server logs used in our performance study, we present trade-offs between the performance metrics for "regular" pairwise volumes (without effectiveness thresholding). We begin with the basic properties and then study the effect of varying different parameters such as time interval and effectiveness threshold. Then, we compare the performance of all our volume-construction methods. Lastly, we investigate two trade-offs important for improv-

[1] Related work has been similar: [1] computes $p_{b|a}$ such that multiple occurrences of b after a are credited just once. [2] gives multiple credit for such a pattern but conditional probabilities are not exploited. In [5], the counters are incremented only if there is a hyperlink from resource a to b and only when b is accessed immediately after a.

Server Log (days)	Number Requests	Number Clients	Unique Resources
Sun (9)	13,037,895	218,518	29,436
EW3_4 (94)	6,273,792	79,623	2638
EW3_6 (94)	6,281,303	102,454	784
EW3_18 (94)	3,902,302	58,839	279
EW3_60 (94)	45,903,646	481,628	2,031

TABLE I
CHARACTERISTICS OF SERVER LOGS REPORTED IN THIS PAPER

ing the computational complexity of volume construction: how often volumes need to be regenerated, and the impact of operating on a subset of the log. A more detailed performance evaluation is presented in [11].

A. Server Logs

Our evaluation was based on several server logs gathered from outside and inside AT&T. One set of server logs were studied in earlier work dealing with volumes [10]–the logs from Amnesty International USA (AIUSA), Apache, Marimba, and Sun Microsystems. A new set of logs were obtained from AT&T's Easy World Wide Web (EW3) [12], a web hosting part of the AT&T Worldnet business unit. Several thousands of companies and organizations have their web pages (with product information and other company related material) hosted on EW3. Roughly three months of server logs were obtained from EW3 and we examined server logs with significant traffic. We chose four out of the top ten logs based on the number of requests.

The data from the logs were cleaned to generate the triples and the pairwise volumes were generated. As is typical of server logs, several records were malformed due to very long lines (often due to inclusion of certain cgi resources that had an abnormal number of <name, value> pairs). The entries needed for our study were the IP address of the client (proxy), the accessed resource, and the time of the access. We converted the first two into hashed integers, the ASCII time representation into an integer form (number of seconds since Dec 31 1969). To generate volumes and compute counters, we used stable sorting on the hashed forms of source IP addresses and processed the logs efficiently. The details of our process for cleaning and processing the server logs is discussed in [11,13].

B. Maximum Time Interval (T)

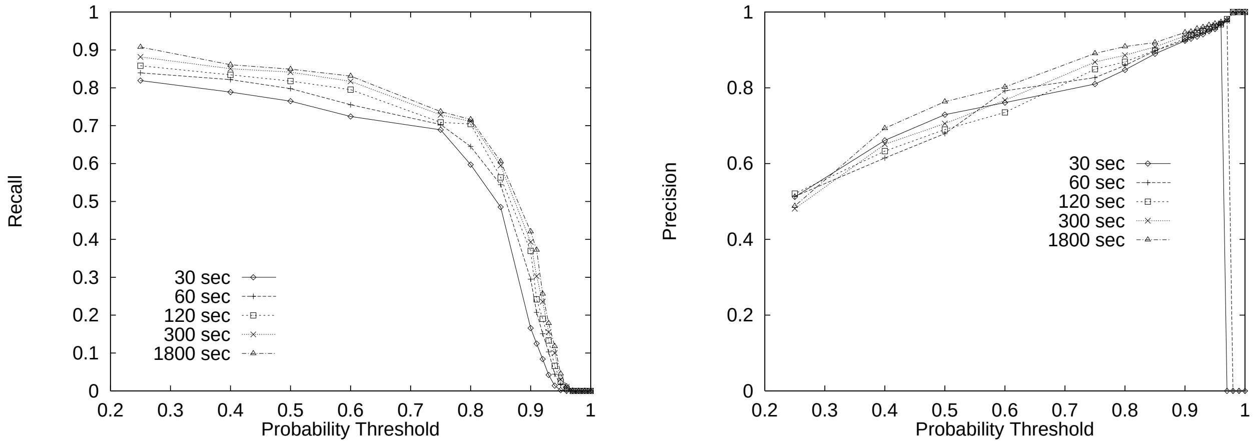
Figures 2–3 show the behavior of regular pairwise volumes for EW3_6, across several values of the time interval (T). Though we show only one log here, other logs show similar patterns of results. As can be seen, a similar general behavior is depicted across all time intervals: recall and hintsize decrease with increasing probability threshold, whereas precision increases. The decrease of recall and hintsize with probability threshold, and the increase of recall with hintsize, are due to the fact that pairwise volumes with higher probability threshold are contained in the respective volumes with lower probability threshold. The increase in precision with decrease in recall and hintsize, is a property of optimal volumes and a desirable property from any heuristic. Pairwise volumes mostly have this property (but not always — for an exception on the Sun log, see [10]).

Figure 2(a) shows that recall initially increases sharply as the probability threshold drops to about $p_t = 0.8$, For threshold values below 0.8, the recall steadily and slowly increases with decreased p_t. This indicates that very few requests are "covered" by high threshold values, but beyond that, "coverage" steadily increases, indicating that lower p_t values provide useful information. Figure 3(b) shows that the dependence of recall on average hintsize is concave (has a non-increasing first derivative). This corresponds to the fact that hints with higher pairwise probabilities are typically more valuable than hints with lower pairwise probabilities. Again, this is a property of optimal volumes which is consistent with our expectations, and is desirable in a heuristic.

Different applications may warrant the use of different lookaheads into future requests. Cache replacement benefits from longer time windows into the future (larger values of T), while cache validation, desiring minimal staleness, benefits from the more recent updates (small values of T). Most of the "achievable" recall is captured in the smallest (30 seconds) time interval, reflecting the fact locality of reference is most present in smaller time scales. Still, there is a measurable increase in recall with larger time intervals (T), while maintaining the same or better performance in terms of trade-offs with hintsize and precision, as the smaller time intervals. This indicates that there is added value in considering longer time intervals, when appropriate for the application. The added value, however, tapers off as larger and larger time windows are needed to achieve similar increases in recall. For example, the increase in recall from 300 seconds (5 minutes) to 30 minutes is comparable to the increase from 120 to 300 seconds.

C. Effectiveness Threshold (e_t)

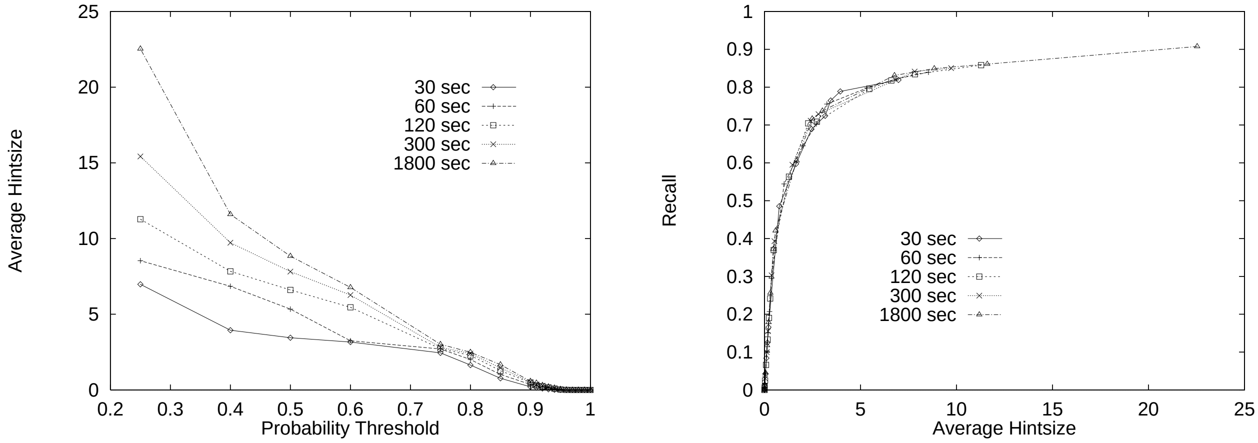
As discussed earlier, pairwise volumes can be thinned using effective thresholding which is controlled by a parameter e_t ($e_t = 0$ corresponds to regular pairwise volumes). We used the heuristic of distributing higher credit to the last resource in the $[\tau, T)$ interval (in contrast to the first-resource heuristic used in [10]). Figure 4 varies the effectiveness threshold for various logs. Higher e_t values produce volumes that are subsets of respective volumes with lower e_t values. Consistently across logs, small e_t values exhibit significant drop in hintsize and a significant jump in precision, with only a moderate drop in recall. The improvements in hintsize and precision taper off with larger e_t values, while recall continues to decrease steadily. The increase in precision levels off and then even drops, while hintsize levels off. Precision decreases because the effectiveness thresholding does not evaluate the impact of removing a volume element on the effectiveness of other predictors. Eventually, the thresholding becomes an overly aggressive



(a) Recall vs. probability threshold

(b) Precision vs. probability threshold

Fig. 2. Recall and precision against probability threshold for different maximum time intervals T (EW3_6 log with $\tau = 0$ seconds)



(a) Average hintsize vs. probability threshold

(b) Recall vs. average hintsize

Fig. 3. Average hintsize vs. threshold and recall for different maximum time intervals T (EW3_6 log with $\tau = 0$ seconds)

mechanism that removes good predictors. Precision starts to increase again on e_t values exceeding 0.4, accompanied by a sharp decrease in recall.

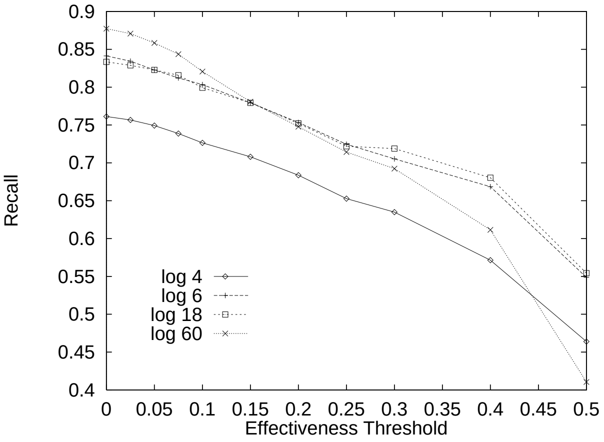
D. Comparison of Algorithms

Figure 5 compares the performance of our volume construction heuristics on log EW3_18 (other logs showed similar results). We used greedy volumes, constructed for the optimization criteria of hintsize vs. recall, for all our experiments. As proposed in Section II-C, the greedy algorithm was seeded with pairwise volumes, in order to achieve reasonable computation time. The greedy algorithm, as expected, outperforms all other techniques. The plots show how effectiveness thresholding helps in dramatically reducing the average hintsize and improving precision (by lowering false predictions), across the range of recall values. The plots verify the behavior argued in Section IV-C: minimal effectiveness thresholding ($e_t \leq 0.1$) is the most valuable, while higher threshold values ($e_t \geq 0.2$) are both worse performers and exhibit undesirable dependencies (where precision decreases for lower recall).

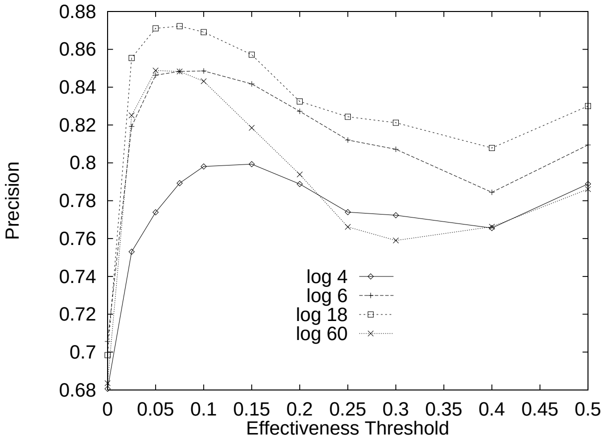
Experiments on the Sun log show that pairwise volumes exhibit initial "flat areas" of the recall vs. average hintsize curves, where recall almost did not increase as the average hint size was increasing. We observed that Greedy not only obtained significantly better trade-offs, but also eliminated the flat areas. The flat areas are due to resources b with correlated predictors, that is, resources a with high $p_{b|a}$, that tend to predict the same set of requests for b. Once the highest (in terms of $p_{b|a}$ predictor is "in" (b is placed in v_a), placement of b in volumes of additional correlated predictors adds to the average hintsize without predicting many additional requests for b. Greedy avoids this by iteratively reevaluating the utility of each predictor. We also observed cases where pairwise volumes had a non-monotonic dependence of precision and recall. With greedy volumes, precision always increases as recall drops.

E. Aging of Volumes

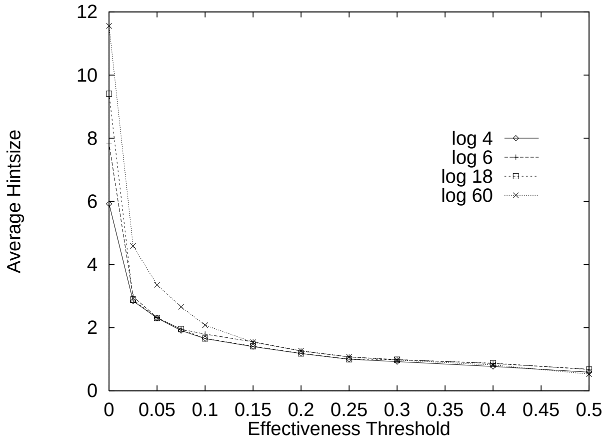
We study how often volumes should be recomputed by the server. This often depends on how transient user access patterns are and the underlying URL structure. Daily to weekly updates, or updates only upon major changes to the site, sufficed for the examined logs. We partitioned the server logs into separate days, built volumes for particular days, and applied these volumes to logs of different days.



(a) Recall vs. e_t

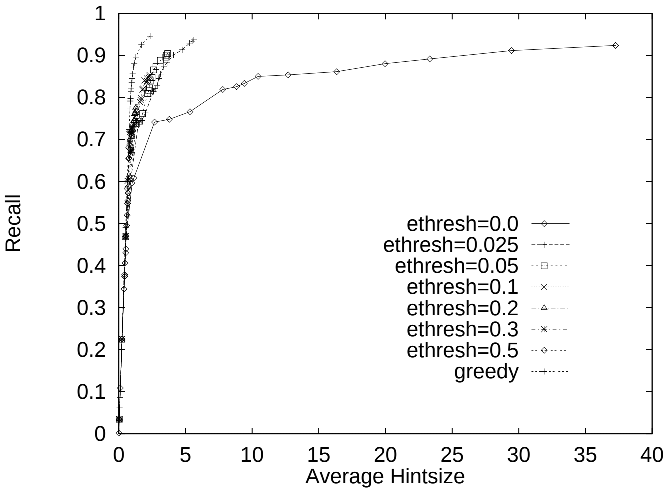


(b) Precision vs. e_t

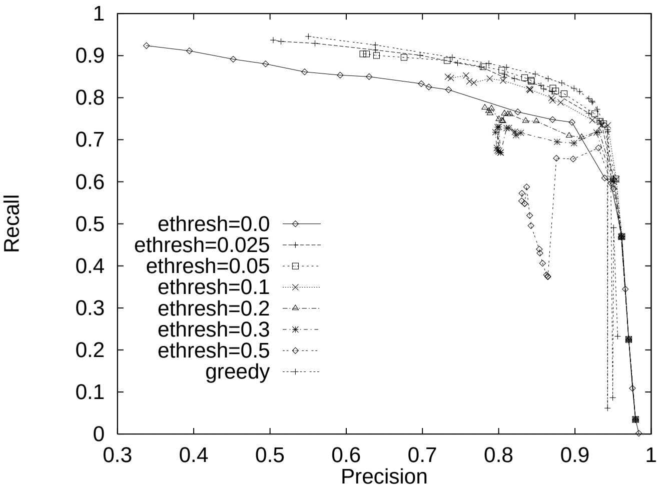


(c) Average hintsize vs. e_t

Fig. 4. Recall, precision, and average hintsize vs. effectiveness threshold for EW3 logs ($T = 300$ seconds, $\tau = 0$ seconds, $p_t = 0.25$)



(a) Recall vs. average hintsize



(b) Recall vs. precision

Fig. 5. Recall and precision vs. average hintsize for various techniques (EW3_18 log with $T = 300$ seconds and $\tau = 0$ seconds)

The Sun log was partitioned into nine days (labeled 0–8). Figure 6(a) shows the trade-off between recall and hintsize when greedy volumes for day 0 (with $T = 300$ and $\tau = 0$) are applied to all other days. The results show that access patterns did not change significantly during the time frame of a week. We partitioned the EW3_18 log into separate days (0–93) and generated greedy volumes (with $T = 300$ and $\tau = 0$) for day 10. Figure 6(b) plots the trade-off when volume of day 10 is applied on days in the same week and same day-of-the-week in the previous and next several weeks. Figure 6(b) shows three clusters of days for which the day 10 volumes had similar performance. The clusters are for days 3–14, day 15, and days 16-66. Within each cluster there is only a slight deterioration with time. This indicates that a change in access patterns occurred during day 15, but otherwise, access patterns remained steady for long periods of time. The change in access patterns was due to reconstruction on the site: the number of unique resources accessed increased by about 60% on day 15, and was relatively steady during previous and subsequent days.

F. Constructing Volumes with Sampled Logs

The size of the logs often dictates the efficiency of our volume construction methods. We examine trade-offs of computation time and performance obtained by processing only a *sample* of the log ("sublog") to produce volumes. These sublogs were generated by including requests made by a fraction (random sample) of the clients. Sampling requests alone is likely to distort access patterns statistics whereas client sampling provides an unbiased sample. Representative behavior for our logs is depicted in Figure 7. On the Sun log, we generated pairwise volumes using 1/2, 1/4, 1/8,1/16, and 1/32 of the clients (with $T = 300$ and $\tau = 0$). These volumes were applied to the full log. The largest client, responsible for 5% of all accesses, exhibited significantly different access patterns and its inclusion during volume generation significantly affected the resulting volumes and their performance, and was excluded from all sublogs. This client was a spider that contacted the site daily. The recall vs. hintsize tradeoffs, using various size sublogs, are shown in Figure 7(a). Figure 7(b) plots results for a similar experiment for greedy volumes on EW3_18 log, using 1/2-1/128th of the clients. The figures show a very small decrease in performance for significant reduction in the size of the log that needs to be processed.

G. Summary of Evaluation

Greedy volumes yielded the best performance in terms of trade-offs of hintsize, recall, and precision. Though we were able to speed up greedy by seeding it with pairwise volumes, it requires scores of iterations, with each iteration requiring a pass through the log. An alternative two-pass construction is obtained by using pairwise volume with effectiveness thresholding. Low e_t values (around 0.1) consistently achieve large improvements in performance over pairwise volumes. All volume construction methods can benefit from sampling the log. Sampling clients results in very slight degradation of performance for large reductions in the size of the log processed. Access patterns, typically, are fairly stable; it thus suffices to update volumes daily or weekly, or upon reconstruction of the site.

V. Conclusions

We have examined a variety of techniques for grouping resources at a Web server site for predicting future requests. Predictions can be used in applications such as prefetching of resources to reduce user-perceived latency, reduction of network traffic by obviating cache validation requests, and better cache replacement at a proxy site. We have provided a strong theoretical basis for the optimal volume construction problem and a variety of heuristics have been examined on a collection of large real-world server logs. We looked at three primary metrics (precision, recall, and hintsize), analyzed server logs, and experimented with thinned volumes. The results show that it often suffices to look at subsets of the logs. Since the problem naturally decomposes to an independent task of predicting requests for a single resource, we plan to look at incremental techniques in future.

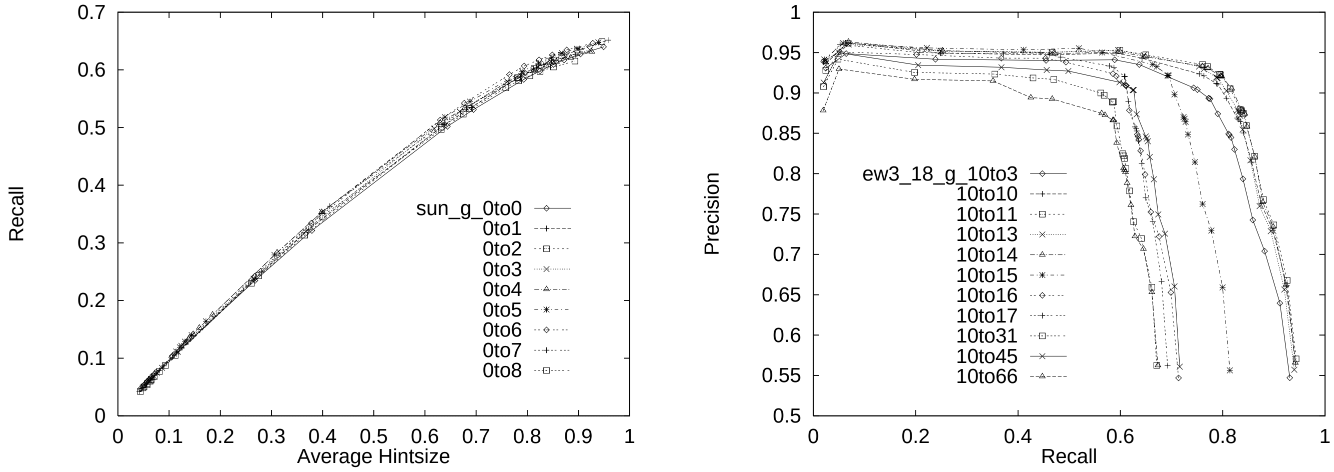
Appendix

I. NP Completeness Proof

In Section II we formulated several optimization criteria for volumes construction, based on obtaining optimal tradeoffs of hintsize and recall or of recall and precision. The NP-completeness of both problems can be established by considering decision problems [14]. For example, given a server log (sequence of triples) and two parameters, h and r, find a set of volumes (if one exists) with average hintsize at most h and recall at least r. Alternatively, given a server log and two parameters, p and r, find a set of volumes (if one exists) with precision at least p and recall at least r.
Lemma: The optimization problems defined above are NP-complete.

Proof: We provide a reduction from 3SAT [14] to the first decision problem. A similar reduction works for the second problem, and is omitted for brevity. Given an instance of 3-SAT, construct a log as follows: Create a resource for each variable and its negation. Create another "special" resource R. The log consists of many "independent subsequences" that capture the constraints of the 3SAT instance. Two requests are independent if made by different clients or are far enough in time (more than T seconds apart) so that predictions made using the volume for the first resource will expire before the other request arrives. Below we use $\tau = 0.1$ and $T = 2$.

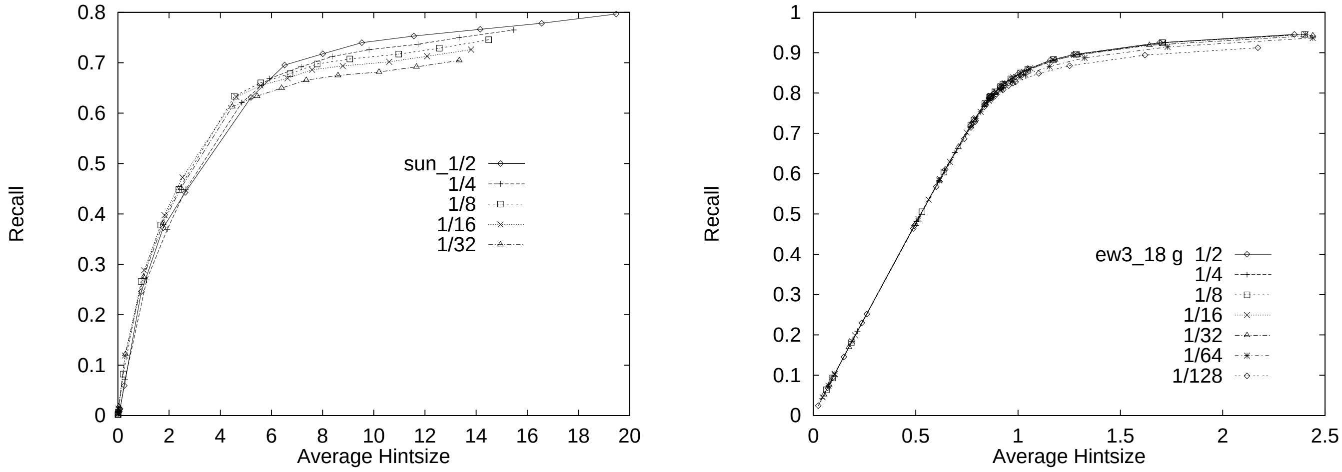
Let M be the number of clauses and N the number of variables in the 3SAT instance. We use the same symbol for each variable and its corresponding resource. For each variable x and its negation $\overline{x}$ we create $3M$ subsequences of the form: "$x, \overline{x}$ at time t and R at time $t+1$". For each clause of the form $(x \vee y \vee z)$ we create a subsequence of the form "x, y, z at time t, R at time $t+1$". We claim that there exist a volumes set with recall of $r \geq M(3N+1)/(9MN+4M)$ and average hintsize of $h \leq (3MN+3M)/(9MN+4M)$, if and only if if the 3SAT instance is satisfiable.



(a) Sun log

(b) EW3_18 log

Fig. 6. Recall vs. average hintsize (a) Greedy volume of day 0 in Sun log, applied to Days 0–8, $T = 300$, $\tau = 0$. (b) Greedy volumes of day 10 in EW3_18 log, applied to various days in days 3–66.



(a) Pairwise volumes with the Sun log

(b) Greedy volumes with the EW3_18 log

Fig. 7. Recall vs. hintsize trade-offs of volumes generated using sublogs for a fraction of the clients, and applied to the full log

Consider a set of volumes that solves the problem above. Observe that only requests for the resource R can be predicted, since other requests have no preceding requests during the previous $[\tau, T]$ seconds. There are $M(3N+1)$ requests for R, and $(9MN+4M)$ requests in total. Hence, recall of $M(3N+1)/(9MN+4M)$ means that the volumes predict all the requests for R. Now, for each variable x, there are requests for R that are preceded (and hence, can be predicted) only by x and $\overline{x}$. Hence, our volumes must be such that for each variable x, at least one of v_x or $v_{\overline{x}}$ contains R. Note, however, that each variable (or negation) with nonempty volume generates at least $3M$ hints. We had allowed for $(3MN+3M)$ hints in total, M of them must be used to predict the requests for R in the "clause" subsequences. Hence, at most N of the volumes can be non empty. By combining the above we obtain that exactly one of v_x and $v_{\overline{x}}$ equals $\{R\}$ and the other is empty.

Consider now the M subsequences that correspond to clauses. Each subsequence contains one request for R, and all these requests need to be predicted to achieve the desired recall. Consider a subsequence corresponding to the clause $(x \vee y \vee z)$. The request for R in the subsequence is predicted, if and only if at least one of v_x, v_y, and v_z equals $\{R\}$. It follows that for any set of volumes with the stated recall and precision, a satisfying assignment can be obtained as follows: For each variable x we set x to be true if and only if the volume of x is equal to $\{R\}$. The considerations above establish that this is well defined and that all clauses are satisfied. Conversely, given a satisfying assignment, the following set of volumes has the desired recall and hintsize: $v_R = \emptyset$, for each true variable x, $v_x = \{R\}$, and for each false variable y, $v_y = \emptyset$.

II. Performance Bounds

A. Greedy Heuristic (Recall vs. Hintsize)

Lemma: The greedy heuristic with respect to recall vs. hintsize trade-offs has performance ratio of $O(\log n)$, where n is the number of requests.

Proof: We write each instance of our problem as an instance of the *weighted set-cover problem* [15, 16]. The input to the weighted set cover problem is a set of elements, and a set of subsets over the elements. Each subset has an associated *weight*. The goal is to find a collection of subsets of minimum combined weight such that their union contains all elements. An important variant of the problem which is relevant to us is to find a collection of subsets of minimum weight which covers *at least a fraction of the elements*. Given an instance of our problem, we view each request as an "element." Each placement of a resource s in a volume v_r constitutes a "subset." The weight of the subset is the total number of request for r (the number of hints produced when s is placed in v_r). The subset contains all elements corresponding to requests for s occurring within $[\tau, T)$ time units following some request for r.

It is easy to establish that a solution for the weighted set cover instance with a certain coverage and weight corresponds to volumes with recall that equals the fraction of covered elements and weight that corresponds to the total hintsize. For example, a set cover that covers 50% of the elements and has weight (combined weights of all participating subsets) 1000, corresponds to a solution of our problem with recall 50% and (total) hintsize 1000 (average hintsize is $1000/n$). Therefore, there is correspondence between the optimal solution of the greedy instance and an optimal solution for our problem. Furthermore, a solution to the set cover instance with approximation ratio of k (k times worse than optimal) corresponds to a volume set with the same performance ratio.

The greedy algorithm has a logarithmic performance ratio [16] for the weighted set cover problem. The well-known greedy set cover algorithm reduces to our greedy heuristic on instances of our problem. To see that, observe that predictions made for different resources are independent (each "subset" covers only requests to a single resource and hence, subsets covering requests to different resources have empty intersection.) Thus, in constructing a set cover it suffices to consider covering requests of each resource r independently (that is, choosing a set of volumes to place r), as our greedy algorithm does. It follows that our greedy heuristic has a logarithmic performance ratio.

B. Pairwise Heuristic (Without Effectiveness Threshold)

Lemma: The pairwise heuristic has performance ratio of $\Omega(n)$, where n is the number of requests.

Proof: We construct a family of logs, parameterized with k. The log has a set of resources $p_1, \ldots, p_q, q, R$. The request sequence consists of many "independent" subsequences (subsequences are independent if more than T seconds separate the last request in one subsequence and the first request in another):

- 9 repetitions of the subsequence: simultaneous (within τ seconds) requests for $p_1, \ldots, p_k$ followed (within (τ, T) seconds) by a request for R.
- 8 repetitions of the subsequence: request for q, followed (within (τ, T) seconds) by a request for R.
- For $i = 1, \ldots, k$, a single-request subsequence "p_i."
- 2 single-request subsequences "q."

The pairwise probabilities are $p_{R|p_i} = 0.9$ and $p_{R|q} = 0.8$. Hence, a pairwise volume set includes R in the volumes $v_{p_1}, \ldots, v_{p_k}$ before including it in the volume v_q. Hence, a pairwise volume set that predicts all requests for R have total hintsize $10(k+1)$ (number of requests for $p_1, \ldots, p_k$ and q). This volume set predicts all $9+8=17$ requests for R, but makes $k+2$ "false" predictions. Hence, it has precision of $17/(19+k)$. The optimal set of volumes contains two non-empty volumes: $v_{p_1} = v_q = \{R\}$. It predicts all requests for R with total hintsize 20 and precision 17/20 (it generates only 3 "false predictions"). It follows that the performance ratio of pairwise is $\Omega(k)$. Since n is proportional to k, the performance ratio is $\Omega(n)$.

To conclude we note that repeated sequences of requests of the form $p_1, \ldots, p_k$ are common as web request sequences (due to embedded images or following hyperlinks), and pairwise volumes perform suboptimally on such sequences.

References

[1] A. Bestavros, "Using speculation to reduce server load and service time on the WWW," in *Proc. ACM Inter. Conf. on Information and Knowledge Management*, 1995.

[2] V. N. Padmanabhan and J. C. Mogul, "Using predictive prefetching to improve World Wide Web latency," *Computer Communication Review*, vol. 26, no. 3, pp. 22–36, 1996.

[3] T. M. Kroeger, D. E. Long, and J. C. Mogul, "Exploring the bounds of Web latency reduction from caching and prefetching," in *Proc. USENIX Symp. on Internet Technologies and Systems*, pp. 13–22, December 1997.

[4] M. Crovella and P. Barford, "The network effects of prefetching," in *Proc. IEEE INFOCOM*, April 1998.

[5] Z. Jiang and L. Kleinrock, "An adaptive prefetching scheme," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 358–368, April 1998.

[6] J. C. Mogul, "Hinted caching in the Web," in *Proc. ACM SIGOPS European Workshop*, 1996.

[7] E. Cohen, B. Krishnamurthy, and J. Rexford, "Evaluating server-assisted cache replacement in the web," in *Proc. European Symposium on Algorithms*, August 1998.

[8] C. Liu and P. Cao, "Maintaining strong cache consistency in the World Wide Web," in *Proc. IEEE Inter. Conf. on Distributed Computing Systems*, pp. 326–334, May 1997.

[9] B. Krishnamurthy and C. E. Wills, "Piggyback server invalidation for proxy cache coherency," in *Proc. World Wide Web Conference*, April 1998.

[10] E. Cohen, B. Krishnamurthy, and J. Rexford, "Improving end-to-end performance of the web using server volumes and proxy filters," in *Proc. ACM SIGCOMM*, September 1998.

[11] E. Cohen, B. Krishnamurthy, and J. Rexford, "Efficient algorithms for predicting requests to web servers," Tech. Rep. 981221-01, AT&T Labs – Research, December 1998.

[12] "AT&T Easy World Wide Web." `http://www.att.com/work-net/easywww`.

[13] B. Krishnamurthy and J. Rexford, "Software issues in characterizing web server logs," in *Proc. Web Characterisation Workshop*, November 1998. `http://www.w3.org/1998/11/05/WC-workshop`.

[14] M. R. Garey and D. S. Johnson, *Computers and Intractability*. San Francisco, CA: W. H. Freeman and Co., 1979.

[15] D. S. Hochbaum, "Approximating covering and packing problems," in *Approximation Algorithms for NP-Hard Problems* (D. S. Hochbaum, ed.), Boston, MA: PWS Publishing Company, 1995.

[16] V. Chvátal, "A greedy heuristic for the set-covering problem," *Mathematics of Operations Research*, vol. 4, pp. 233–235, 1979.