A Framework for the Optimizing of WWW Advertising

Charu C. Aggarwal, Joel L. Wolf and Philip S. Yu

IBM T.J. Watson Research Center, Yorktown Heights, New York

Abstract. This paper discusses a framework and provides an overview of general methods for optimizing the management of advertisements on web servers. We discuss the major issues which arise in web advertisement management, and describe basic mathematical techniques which can be employed to handle such problems. These include a number of statistical, optimization and scheduling models.

1 Introduction

With the recent explosion of web related sites and services, the internet has become a great opportunity for advertising [3]. For site administrators maintaining large numbers of web pages the process of optimizing the placement and time-based scheduling of advertisements on web pages has become more or less essential.

Advertisements on web pages typically occur as either inline or text links. Each web page on a server has a certain number of predefined standard size *slots* containing either the inline images themselves or text links to the actual advertisement pages. We say that an advertisement is *exposed* when a web page containing a slot with the advertisement is served to a client. Since a web page may contain more than one advertisement slot, more than one advertisement may be exposed at a single time. We say that an advertisement is *clicked* when a client chooses the link corresponding to an exposed advertisement. The number of clicks for an advertisement is thus a fraction of the number of exposures. Advertisers measure the number of times a person sees and/or clicks on an advertisement in order to measure the effectiveness of contracting with a particular server. In fact, the cost of an advertising contract may be depend directly or indirectly on either the number of exposures, the click rate, or both. In any case it is clearly desirable for the site administrator to maximize the total number of hits to advertisements on the server.

In this paper we discuss several general methods which attempt to optimize the assignment of advertisements to slots in web pages. To do so realistically we must take a number of important factors into consideration, and our primary goal here is to highlight these. The issues include web page access rate distribution, fairness, dynamic scheduling, and optimizations based on time-of-day, advertisement and web page content affinities, and client demographics. Since we cannot go into great detail in the space available here, we focus instead on providing a high level framework in which the optimized advertising decisions can be made. Thus this paper is intended primarily as a starting point for further discussion.

Since advertising agencies measure the effectiveness of an advertisement by the total number of exposures over all web pages and by the number of clicks that the advertisement generates, it is advantageous for the site administrator to maximize the overall click/exposure ratio. This will be a prime metric used in this paper in evaluating the assignment of advertisements to web pages.

The remainder of this paper is organized as follows: Section 2 describes some of the primary issues in more detail. In Section 3 we briefly discuss statistical data collection for the problem. The notion of a "possibility graph" is explained in Section 4. This graph makes the two optimizations described in Sections 5 and 6 more computationally easy. Section 7 contains conclusions.

2 Primary Issues

In this section, we discuss the primary issues involved in making the decisions about assignments and scheduling of advertisements on web pages. Doing so helps to demonstrate the overall complexity of the problem at hand.

(1) **Web Page Access Rate Distribution:** Some pages in a web site are likely to be accessed much more frequently than others. It is useful to exploit these hot pages by assigning them to advertisements which have a high click-exposure ratio.

(2) **Fairness:** The static placement of advertisements on web pages is likely to provide an overexposure to some advertisements and an underexposure or even starvation to others. Fairness criteria would dictate that this is not a good idea.

(3) **Dynamic Scheduling:** Dynamic advertisement scheduling avoids this by maintaining a minimum exposure percentage for each advertisement.

(4) **Time of Day:** Some advertisements are much more likely to be clicked at certain times of the day than others. Thus, an optimum matching of advertisement placement based on the time of day is important in order to maximize the effectiveness of advertisement placement.

(5) **Content classification:** It is important to choose advertisements which are appropriate to the corresponding web page content. For example, the sports page of a newspaper is probably unlikely to contain useful advertisements for a financial broker. This can be done by a two-stage classification scheme. The first stage is a keyword matching process, by which a rough initial classification is done, while the second stage is a refinement process in which a better matching is done using statistical forecasting information.

(6) **Client dependant decisions:** The IP-address of the person who accesses a web page can provide valuable information in the advertising optimization process. For example, an advertisement about job listings is likely to be very relevant to a person accessing the web page from the education

domain, though it is important not to starve advertisements from other domains. Similarly, for a car dealer in the Massachusetts area, it is critically important that the advertisement be accessed strictly by clients within that area. Using IP-addresses to make advertising more effective is more difficult because such decisions have to be made "on the fly", and there is a limit to the amount of computing which can be done without affecting the perceived web server performance. In sites which have a registration procedure (for example, *The New York Times*), explicit demographic information is available, and this may be used in order to perform advertisement placement. Typically one performs cluster analysis to partition either the individuals or their IP-addresses into groups with similar features. We cluster individuals into demographic groups based on age, salary and other relevant information. We also attempt to identify large segments of IP-addresses which have heavy correlation in terms of user behavior. For example, the IP-addresses corresponding to .edu represent a set of university students, professors and researchers, and are likely to have considerable correlation in web page access behavior. Due to space considerations, both the clustering and use of the same is beyond the scope of the current paper.

3 Statistical Data Collection

Since this entire decision process depends upon the method of statistical forecasting, it is important to be able to be able to assemble statistical data which is unbiased and truly representative of actual trends in user behavior. An attempt to collect data only for advertisements which are served using an optimization mechanism can lead to an inadvertent bias in data collection. More specifically, in order to determine which domains are most favorable for a particular advertisement, it is necessary to show the advertisements to all the possible domains for a statistically significant number of times.

Fortunately, the need to perform unbiased statistical data collection can be combined with the need to provide fairness to the system. Consequently, a certain fraction of advertisement assignments can be done by cycling them among all the web pages in strict round robin fashion, and the statistics calculated from these assignments is maintained separately from all the other statistics. For example, if it is decided that about 10% of the advertising assignments should be performed in order to satisfy the fairness criterion, then every tenth assignment can be treated separately.

4 The Possibility Graph

Our goal is to decide on appropriate potential assignments of advertisements to web pages, but we wish to make the job of the optimization routine as computationally easy as possible. Therefore, in this section we shall attempt to eliminate the irrelevant assignment possibilities to some extent. We can do this in several ways, either manually or in an automated fashion. The output of this process

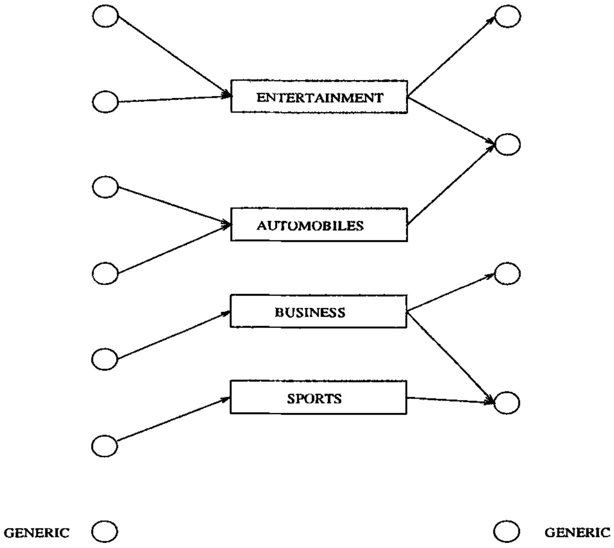

Fig. 1. Keyword Matching to Create The Possibility Graph

will be a so-called *possibility graph*. The possibility graph is bipartite, with one subset N_1 of nodes corresponding to the advertisements, and the other subset N_2 of nodes corresponding to the web pages. There will be a directed arc from an advertising node to a web page node if and only if it is "worthwhile" to assign the advertisement to the page.

Perhaps the simplest approach to this is manual, via a rough "keyword assignment". The site administrator can assign certain specific keywords to various subsets of both the advertisements and the web pages. These keywords correspond to content. We assume that a keyword "generic" is also available as a catchall for either the advertisements or the web pages. An advertisement categorized as catchall links to all web pages, and vice versa. See, for example, Figure 1, with four keywords plus one catchall advertisement and one catchall web page. Administrators can allow the process of assignment to be guided entirely by statistical information if desired. The idea of which keywords to use should be site specific. Now we generate the possibility graph by drawing a directed arc from an advertisement to a web page whenever there exists at least one keyword common to both. As noted, the outdegree of the generic advertising node is full, as is the indegree of the generic web page node. Figure 2 shows the resulting possibility graph for this example.

There are essentially two further types of data required as input to the assignment optimization model. Both of these can be forecasted by statistical techniques.

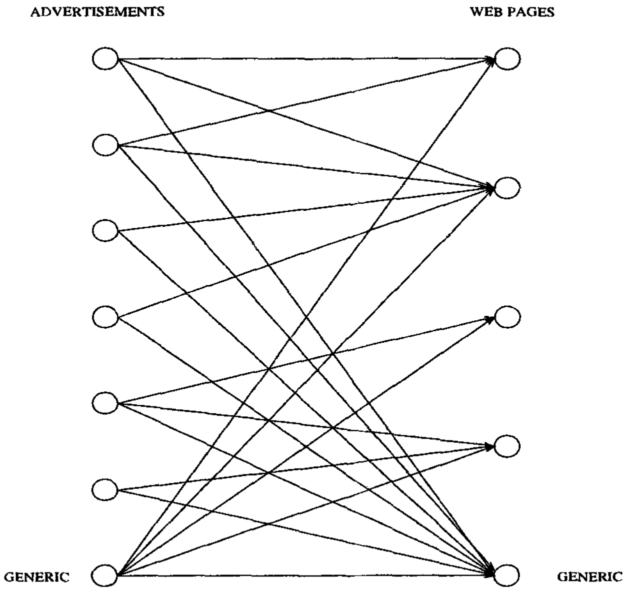

Fig. 2. Possibility Graph

(1) **Click/Exposure Ratio:** The *click/exposure* ratio is the ratio of the number of times an advertisement is accessed to the number of times a user views a web page containing the advertisement. The click/exposure ratio thus depends on both the page which is being accessed and the particular advertisement occurring on that page. As we have already indicated, we expect that the click/exposure ratios is also dependent on the time of day. This means, in turn, that the assignment model should be solved in each time interval. Dividing the day into hours would seem a reasonable alternative. At the beginning of each one-hour slot, an optimization model using forecasted data as input is solved in order to decide the final advertisement assignments to web pages. Once these assignments are generated, a pure round-robin scheme among these pages could be used for that particular hour. The click/exposure ratios may be forecasted by using a simple linear prediction model as discussed in [2].

(2) **Recurrence Factor:** The *recurrence factor* is the average number of times that a particular client will access the same web page in a single time-slot, given that the client accessed it once. For most pages, the recurrence factor is expected to be very close to 0. However, for pages containing certain types of real-time results (such as sports, stocks and weather) the recurrence factor for some web pages is likely to be higher.

5 A Simple Assignment Model

The possibility graph describes the worthwhile advertisements which may be assigned to each web page. Our goal here is to a pick an optimal subset of these advertisements for each web page and then *schedule* them equally in round robin fashion. For example, suppose that a web page A has a single advertisement slot and the possibility graph indicates that it may be assigned to advertisements 1, 2, 3, 4 and 5. Suppose further that our optimization model picks the subset 2, 3, and 5. Then the scheduling algorithm will keep assigning them round robin to the clients of page A. We can assume that the number of advertisements which are assigned to a web page must be at least equal to the number of slots on the page. On the other hand, each advertisement should be assigned to at least one web page. Finally, pages with large recurrence factors should be assigned a greater number of different advertisements than other pages, because a greater rotation of advertisements provides better overall exposure.

We are ready now to discuss how to construct the optimization model in order to find a solution to the advertisement assignment problem. This optimization model is a *minimum cost flow* problem [1]. The model which we shall discuss in this section assumes a flat rate which a site administrator charges for advertising services. Furthermore it does not consider the possibility of guaranteeing minimum numbers of exposures to the advertisers. On the positive side, the model is relatively simple. In the next section we shall discuss a more flexible network flow model.

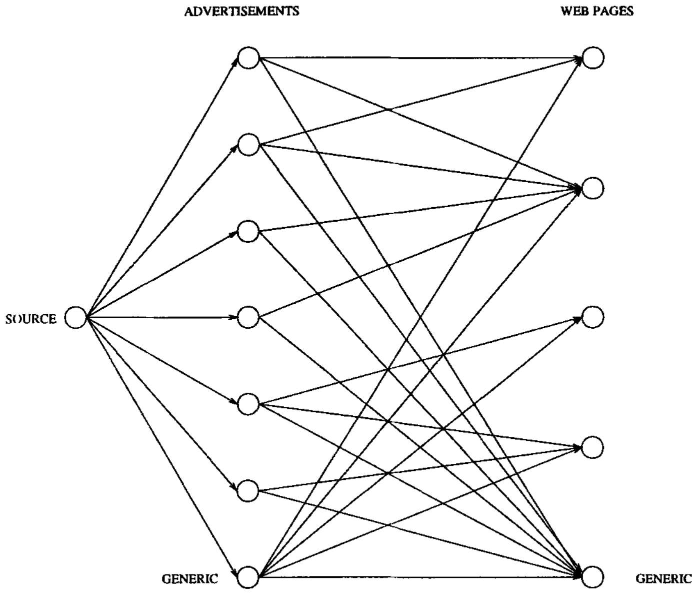

Fig. 3. Directed Graph for Network Flow Problems

We construct the network $G = (N_1 \cup N_2 \cup \{s\}, A)$ in which each node $i \in N_1$ corresponds to an advertisement, and each node $j \in N_2$ corresponds to a web page. The node s is the source. See Figure 3 for an illustration of the underlying network. A directed arc (i, j) from $i \in N_1$ to $j \in N_2$ exists if and only if the advertisement i can be assigned to web page j according to the possibility graph. The capacity of each of these directed arcs is one unit, and the decision variable x_{ij}, which represents the flow, is either 1 or 0 depending on whether or not the advertisement i is assigned to web page j. The cost $-R$ of the directed arc (i, j) is the negative of the click/exposure ratio R for that advertisement / web page pair, and thus measures the desirability of the potential assignment. The flow on each of the directed arcs (s, i) represents the number of pages to which a particular advertisement is assigned. Consequently the lower bound on each of these directed arcs is one unit. The cost on each of these directed arcs is zero. Note that the network G is obtained from the possibility graph by adding a source node s and the directed arcs from s to each node $i \in N_1$. The demand on each web page node $j \in N_2$ denotes the number of advertisements that will be assigned to that page. The demand on node j should at least be equal to the number of advertisement slots available on that page and should at most be equal to the indegree of the node j in the possibility graph. Another factor which needs to be taken into account while deciding the demand of node j is the recurrence factor corresponding to the page j. If a page has high recurrence, then the demand of the node needs to be increased so as to provide viewers with a greater variety of advertisements.

If a minimum cost flow is obtained in this network, and the flows on the directed arcs from the advertisements to the pages are used in order to determine which advertisements are assigned, then this mechanism maximizes the average click/exposure ratio of the advertisements. Once the set of advertisements for a particular page are obtained, the number of such advertisements typically being larger than the number of slots available on that page, a round robin scheduling scheme is used in order to rotate those eligible set of advertisements among the slots. In the next section we shall discuss a more generic model which takes actual contractual constraints into account.

6 A More Flexible Assignment Model

Site	Cost
Netscape	$ 30000 per month per million impressions
Yahoo	$ 20000 per month per million impressions
Pathfinder	$ 38000 per quarter per 1.3 million impressions
Playboy	$ 30000 per quarter per 10 million impressions

Table 1. Advertising Costs for Some Popular Sites

Most current web sites offer a very wide range of contracts and rates. Some information may be found in [4]. Many sites currently charge advertisers on the basis of *cost per month* per one thousand impressions, also referred to as *CPM.* Some typical values of costs which are currently being charged are illustrated in Table 1, which is taken from [4]. (Note that the units in the table are not entirely identical, though they can easily be converted to CPM.) It should be noted that the typical rate on a per thousand impression basis is very dependent upon the degree of the popularity of a site. For the purpose of advertisement management, the number of accesses to a site have to be viewed by the site administrators as a resource. For example, Playboy can afford to offer such low rates on a CPM basis because of the large number of hits to its site. In the context of a model which charges on the basis of the number of hits, the concept of advertisement space available may be redefined to a quantitative notion of the number of times that an advertisement may be shown to viewers. Thus, popular sites have more available advertisement space, and can afford to give much more competitive rates.

We shall now proceed to discuss a model which provides a graded pricing scheme, which we consider a suitable cost model in terms of its flexibility in incorporating different situations. A site administrator has the flexibility of offering a number of different types of contracts which are based on "graded costs". Each of the different categories of contracts, is based on the number of impressions for an advertisement being in a certain range. The per impression cost of an advertisement goes down as the number of impressions increases. A possible example of such a graded cost scheme is illustrated in Table 2. As we can see, the cost structure is concave with the number of impressions, and this provides the advertiser an incentive for choosing a contract with a larger number of impressions.

Impressions/month	Incremental cost per month per 10000 impressions
0-10000	$ 140
10000-20000	$ 100
20000-30000	$ 70
30000-60000	$ 50
60000-100000	$ 30

Table 2. Graded Cost Scheme

An interesting feature of cost contracting in advertisements is the issue of the pricing of "click-throughs". For example, for a Rolls-Royce manufacturer it might be worthwhile to pay a few dollars for each click-through, given the high price of a single unit. Of course, they might not be interested in paying for exposures. On the other hand, a company which manufactures soaps might rather pay a higher price on a per exposure basis than pay for click-throughs. We introduce another term here, namely the *cost per click-through*, or *CPC*. Thus a site might offer a higher number of possible contracts, which are generally structured in such a way that a higher CPM means a lower CPC, and vice versa. Typically advertisers whose products have a very high profit margin for each unit sold but a lower number of units actually sold are expected to be targeted by contracts which have click-through costs. An example of such a pricing scheme is illustrated in Table 3. We will now discuss an optimization model which maximizes the revenue for a flexible pricing scheme.

The model which we shall discuss again uses the possibility graph, but now is employed to determine the actual number of desired assignments for each of the slots in a page. In many ways the model is quite similar to the model that we discussed in the previous section. In fact, the underlying network is the same, though the costs, flows and such are different. See again Figure 3. Specifically, we construct a network $G = (N_1 \cup N_2 \cup \{s\}, A)$ in which each node $i \in N_1$ corresponds to an advertisement, while each node $j \in N_2$ corresponds to a page. As before, s is the source. The directed arc (i, j) exists in the network G if and only if advertisement i and web page j are joined in the possibility graph. Each of these directed arcs is uncapacitated. The flow x_{ij} on this directed arc will determine the desired number of assignments of advertisement i to web page j. The cost of this directed arc is the negative of the expected profit associated with each assignment of page i to advertisement j. The expected profit has two components. The first component is the profit associated with the exposure itself, while the second component is the expected profit associated with a click-through, where R is, as before, the forecasted click-exposure ratio for an advertisement using historical information. Thus, the cost of the directed arc (i, j) is $-(CPM/10000 + CPC * R)$. Note that the CPM is normalized by a factor of 10000, because the CPM values are expressed in terms of groups of 10000 exposures. The source node s has directed arcs to each node $i \in N_1$. The flow on the directed arc (s, i) represents the number of impressions corresponding to the advertisement i. Thus, if a contract corresponding to an advertisement i has a constraint corresponding to a minimum number of desired exposures, this is handled by putting a lower bound on the directed arc (s, i). The demand for each node $j \in N_2$ is indicative of the forecasted popularity (number of accesses) of web page j and is obtained using a time-series analysis, based on the historical access data for that page. The supply for the source node s is the sum of the demands for the sink nodes. (It should be noted in passing that typical contracts require a minimum number of exposures on a monthly basis, while this model might typically be used for time slot assignments whose lengths are on the order of an hour. To handle this, the lower bound on the directed arc (s, i) can be obtained by dividing the residual commitment for the month by the number of slots for which the advertisement may be scheduled in the current month.)

Contract	Impressions	CPM	CPC	Advertiser
1	10000-20000	$ 100	None	Soaps
2	20000-40000	$ 80		
3	40000-80000	$ 60		
4	10000-20000	$ 20	$ 0.25	Personal Computers
5	20000-40000	$ 16		
6	40000-80000	$ 12		
7	10000-20000	None	$ 2	Luxury Cars
8	20000-40000			
9	40000-80000			

Table 3. Flexible Pricing Scheme

7 Conclusion

This paper provides a structure for making optimal decisions about assigning and scheduling advertisements on web pages. We have indicated some of the complexities involved, and highlighted some realistic solutions. The optimized placement of advertisements on web pages is likely to become a hot area of research as the commercial aspects of the world wide web grow.

References

1. Ahuja, R., T. Magnanti, and J. Orlin, "Network Flows", *Prentice Hall*, 1993.
2. Press W., S. Teukolsky, W. Vetterling, and B. Flannery, "Numerical Recipes in C", *Cambridge University Press,* 1992.
3. Welz G., "New Deals", *Internet World*, April 1995, pp. 36-41.
4. Welz G., "The Ad Game", *http://www.iw.com/1996/07/adgame.html.*