Coverage Patterns-based Approach to Allocate Advertisement Slots for Display Advertising

Vaddadi Naga Sai Kavya and P. Krishna Reddy

Kohli Center on Intelligent Systems (KCIS)
IIIT Hyderabad, Gachibowli, Telangana State, India
{saikavya.vaddadi@research.iiit.ac.in and pkreddy@iiit.ac.in}

Abstract. Display advertising is one of the predominant modes of online advertising. A publisher makes efforts to allocate the available ad slots/page views to meet the demands of the maximum number of advertisers for maximizing the revenue. Investigating efficient approaches for ad slot allocation to advertisers is a research issue. In the literature, efforts are being made to propose approaches by extending optimization techniques. In this paper, we propose an improved approach for ad slot allocation by exploiting the notion of coverage patterns. In the literature, an approach is proposed to extract the knowledge of coverage patterns from the transactional databases. In the display advertising scenario, we propose an efficient ad slot allocation approach by exploiting the knowledge of coverage patterns extracted from the click stream transactions. The proposed allocation framework, in addition to the step of extraction of coverage patterns, contains mapping, ranking and allocation steps. The experimental results on both synthetic and real world click stream datasets show that the proposed approach could meet the demands of increased number of advertisers and reduces the boredom faced by user by reducing the repeated display of advertisements.

Keywords: Internet monetization, Computational advertising, Display advertising, Coverage patterns

1 Introduction

Banner advertising or display advertising is one of the predominant modes of online advertising along with contextual and sponsored search advertising [3]. The three major entities involved in display advertising scenario are advertiser, publisher (or ad server) and user (or visitor to the website). In the guaranteed contract of display advertising scenario, an advertiser demands certain number of views for his/her display ad. A publisher manages the available ad slots on web pages of the website through ad server and makes an effort to allocate appropriate sets of ad slots to meet the demands of advertisers. A user visits a set of web pages and the corresponding display ads which are placed in the ad slots. In this scenario, given the budget constraints, an advertiser aims to reach the maximum number of distinct potential users. That is, it may not be

beneficial for the advertiser if the same user sees the advertisement multiple number of times. A publisher aims to meet the demands of increased number of advertisers to maximize the revenue. Also, a user who visits multiple web pages and the corresponding ads wants to have a good browsing experience [2, 7, 9, 17]. The user is annoyed if the advertisement is displayed repeatedly. The research problem is to develop efficient allocation approaches to help the publisher in maximizing the revenue by satisfying the demands of increased number of advertisers and help the advertiser to reach the maximum number of distinct users without causing annoyance to users.

In the literature [13, 19, 26], the problem of display ad allocation is being studied by modeling it as a bipartite graph in which ad slots, advertisers and allocations between ad slots to advertisers are represented as supply nodes set, demand nodes set, and edges respectively. Efforts [7, 12, 15, 17] are being made to develop efficient ad allocation approaches using optimization techniques by mathematically formulating the ad serving scenario. These approaches have modeled the problem of display ad allocation as a stochastic optimization problem and attempted to develop optimal or near optimal allocation plans.

In this paper, we have made an effort to propose a different approach for efficient allocation of ad slots by extending the notion of coverage. In the literature, an effort has been made to extract the knowledge of coverage patterns from transactional databases [21, 22]. Each coverage pattern (or set) covers certain percentage of transactions. So, a large number of coverage patterns can be extracted from the transactional databases such that each pattern covers certain percentage of transactions. In the display advertising scenario, there is an opportunity to use the knowledge of coverage patterns which can be extracted from the click stream data of a website to efficiently identify the supply of user visits for various ad slots to meet the goals of both publisher and advertiser. Further, the boredom to the user can be reduced.

For display advertising, in this paper, we have explored how the knowledge of coverage patterns extracted from click stream transactions could be exploited in developing the efficient ad slot allocation approach. We have exploited the fact that a coverage pattern gives certain coverage/percentage of visitors and it contains a distinct set of visitors. We have proposed a framework to allocate ad slots to advertisers based on the knowledge of coverage patterns. The experimental results on both synthetic and real world click stream datasets show that the proposed allocation approach could allocate increased number of advertisers by reducing the repeated display of the same ad as compared to the baseline approach.

In this approach, it is assumed that the transactions formed from click stream data could be used to identify the set of ad slots that cover a given percentage of visitors. Such knowledge could be used to allocate ad slots to the advertisers by assuming similar access behaviour. The related issues will be investigated as a part of future work.

The rest of the paper is organized as follows. In Section 2, we present related work. In Section 3, we present an overview of coverage patterns and explain the

framework of display ad allocation. In Section 4, we explain proposed allocation framework. In Section 5, we present experiments. In Section 6, we present conclusion and future work.

2 Related Work

Regarding display advertising, research efforts are being made in the literature to propose efficient approaches for scheduling of display ads, allocation of display ads, preserving privacy, auction mechanisms and charging schemes [5, 11, 15, 20]. We discuss the related work concerning allocation of display ads. As we employ approaches related to click stream data and coverage patterns in the proposed approach, we also discuss the corresponding related work.

Efforts have been made to study the problem of display ad allocation by modeling it as a bipartite graph matching between the advertisers with demands and ad slots with supply of user visits with an objective of maximizing the revenue. Feldman et al. [13] have considered the online ad allocation problem by modeling it as an ad matching problem and generated upper bounds for the optimum allocation. Vee et al. [26] have formulated the optimization problem of ad allocation and described primal compact sample allocation plan which might generalize to a near optimal solution. Vahab et al. [19] have studied the problem of simultaneous approximations for the adversarial and stochastic online budgeted allocation problem and provided mathematical approximation bounds based on the arrival orders of nodes in the bipartite graph. Bharadwaj et al. [7] have considered the impression based mathematical formulation that minimizes the under delivery rate in a bipartite graph framework of ad allocation. Hojjat et al. [17] assign each user a predefined fixed stream of ads from a pool of simulated ad streams and use a column generation scheme to select a small set of ad streams that optimize the defined ad allocation problem.

Research in click stream mining is mainly focused on the development of knowledge discovery techniques specifically designed for the analysis of web usage data. Web log databases provide rich information about web dynamics [24, 16]. In [18], several measures of interest to evaluate the association rules mined from web usage data were proposed and compared.

An approach to extract coverage patterns from transactional database has been proposed in [21]. Alternative approaches to extract coverage patterns have also been proposed in [22, 23]. In [25], a methodology to extract content-specific coverage patterns from transactional database has been proposed. It was shown that the knowledge of coverage patterns can be extracted from relatively large transactional databases and the potential application of coverage patterns to the display advertising and sponsored search advertising scenario was described. In [8], a framework has been proposed to improve the efficiency of adwords by considering a set of queries of a user as a transaction.

It can be noted that the preceding allocation approaches estimate the supply of user visits for ad slots. In addition, an objective function is modeled mathematically by simulating display ad serving scenario for maximizing the revenue.

The solution sets are generated based on the mathematical bounds for the specified constraints. The solutions produced require predicting the user visits in a concrete way to start with the ad allocation plans.

The ad allocation approach proposed in this paper is different from preceding approaches as we have extended the knowledge of coverage patterns extracted from click stream transactions to improve the performance of ad allocation in display advertising.

3 Background

In this section, we briefly explain the concept of coverage patters and explain the problem of display ad allocation.

3.1 Overview of coverage patterns

In the literature, the concept of coverage has been used to the solve set cover problem in set theory [10] and node cover problem in graph theory [14]. By extending the notion of coverage to transactional databases, a model to extract coverage patterns (CPs) was proposed to extract transactional coverage value of distinct sets of data items [21] or patterns.

Given a transactional database C, a pattern is a set of items. We attach the following notions to coverage patterns: coverage set, coverage support and overlap ratio. Given a pattern, the coverage set denotes the set of all distinct transactions n such that every transaction in n contains at least one web page of a pattern. The coverage support indicates the coverage i.e., the extent of the coverage of the pattern. It is the ratio of coverage set and total number of transactions. Another important parameter is overlap ratio. Consider a set of two items appearing in n transactions out of C transactions. If the two items are appearing in every transaction of n transactions, then the coverage support is n/C. But, this pattern is uninteresting with respect to coverage point of view. It would be more interesting, ideally, if each item in it appears from every different transaction of n transactions so that the coverage can be maximized. So, given the pattern, the degree of overlap of individual items coverage is captured by overlap ratio. So, a pattern is interesting if it has high coverage support and low overlap ratio.

Formally, given a transactional database, the CPs can be extracted based on the threshold values specified for relative frequency, coverage support and overlap ratio parameters. We briefly explain the three parameters as follows.

Let $I = \{i_1, i_2, \cdots, i_n\}$ be a set of items and C be a set of transactions such that each transaction T is a set of items where $T \subseteq C$. P is a pattern of items such that $P \subseteq I$ and $P = \{i_p, \cdots, i_q, i_r\}$ where $1 \leq p \leq q \leq r \leq n$ and T^{i_k} denotes a set of transactions containing the item i_k and its cardinality is denoted by $|T^{i_k}|$.

- **Relative frequency:** The fraction of transactions containing the item, i_k is called relative frequency of i_k. It is denoted by $RF(i_k)$ and measured as

follows.

$$RF(i_k) = \frac{|T^{i_k}|}{|C|} \tag{1}$$

– **Coverage support:** Coverage support of a pattern, P is the ratio of number of transactions which contains at least one item in P to the size of the transactional database, $|C|$ such that $|T^{i_p}| \geq \cdots \geq |T^{i_q}| \geq |T^{i_r}|$. It is denoted by $CS(P)$ and measured as follows.

$$CS(P) = \frac{|(T^{i_p} \cup \cdots \cup T^{i_q} \cup T^{i_r})|}{|C|} \tag{2}$$

– **Overlap ratio:** Overlap ratio of a pattern, P, where $|T^{i_p}| \geq \cdots \geq |T^{i_q}| \geq |T^{i_r}|$, is the ratio of the number of transactions common in $P - \{i_r\}$ and $\{i_r\}$ to the number of transactions in i_r (i.e., minimum number of transactions in either $P - \{i_r\}$ or $\{i_r\}$). It is denoted by $OR(P)$ and measured as follows.

$$OR(X) = \frac{|(T^{i_p} \cup T^{i_{p+1}} \cup \cdots \cup T^{i_q}) \cap (T^{i_r})|}{|T^{i_r}|} \tag{3}$$

Let $minRF$, $minCS$ and $maxOR$ be the threshold values specified for RF, CS and OR parameters respectively to extract CPs from a transactional database. A pattern P is said to be a CP, if $RF(i_k) \geq minRF$ where $i_k \in P$, $CS(P) \geq minCS$ and $OR(P) \leq maxOR$.

Example 1: We explain the notion of CPs on an example transactional database C where $|C| = 10$ as shown in Table 1. Let us consider $minRF = 0.4$, $minCS = 0.7$ and $maxOR = 0.5$. From C, the set of items, $I = \{a,\ b,\ c,\ d,\ e,\ f\}$, $T^a = \{1,\ 2,\ 3,\ 4\}$ and $T^b = \{1,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10\}$. Thus, $RF(a) = \frac{4}{10} = 0.4$ and $RF(b) = \frac{7}{10} = 0.7$. Similarly, RF will be calculated for all other items in the set I. The items whose $RF \geq minRF$ will be considered and the rest will be removed. Here in this example, $RF(f) = 0.3 \leq minRF$, hence it will be removed. As a,b satisfy the $minRF$ constraint, {b, a} can be an item set. We need to measure CS and OR parameters of {b, a} to check whether it can be a CP. $CS(\{b, a\}) = \frac{|(T^b \cup T^a)|}{|C|} = \frac{|\{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10\}|}{10} = \frac{10}{10} = 1$. $OR(\{b, a\}) = \frac{|(T^b) \cap (T^a)|}{|T^a|} = \frac{|\{1\}|}{4} = \frac{1}{4} = 0.25$. As $CS(\{b,\ a\}) \geq minCS$ and $OR(\{b,\ a\}) \leq maxOR$, {b, a} is a CP. Similarly, all other valid CPs are extracted from C.

Table 1: Example transactional database

TID	items	TID	items
1	$a,\ b,\ c$	6	$b,\ d$
2	$a,\ c,\ e$	7	$b,\ d$
3	$a,\ c,\ e,\ f$	8	$b,\ e,\ f$
4	$a,\ c,\ d$	9	$b,\ e,\ f$
5	$b,\ d$	10	$c,\ b,\ d$

In the literature, an effort has been made to propose apriori-like approach [21] to extract complete sets of CPs from transactional database. A pattern growth-like approach [22, 23] is also proposed to extract the complete set of CPs in an efficient manner.

3.2 Framework of display ad allocation

In the guaranteed contract of display advertising scenario, an advertiser aims to reach the potential users and demands certain number of views (or impressions) for his/her advertisement on the publisher's website. The publisher who manages the ad space through the ad server, guarantees the supply of demanded impressions to the display ad, and generates revenue under appropriate revenue model. Cost per impression (CPI) is the most commonly used revenue model [17]. In this scenario, major components include ad slots on web pages of the website which supply certain number of impressions, advertisers who demand impressions and the publisher who allocates the ad slots that matches demands of advertisers through the ad server and generates revenue under CPI revenue model.

The entire scenario of display ad allocation naturally fits into a bipartite graph in which one set of nodes represent ad slots, another set of nodes represent advertisers and edges represent allocations [7, 17, 26, 27]. Figure 1(a) shows a bipartite graph $G(S, D, E)$ in which S contains five supply nodes (ad slots) on the left and D contains three demand nodes (advertisers) on the right. Here, each ad slot is modeled as a supply node and each advertiser (advertising contract) is modeled as a demand mode. Each supply node $S_i \in S$ supplies s_i impressions (weight of the node S_i) and each demand node $D_j \in D$ demands d_j impressions (weight of the node D_i). Th edge set E represents the eligible allocations between the supply and demand nodes. The problem is to achieve fair and optimal allocation between the supply nodes and demand nodes so as to maximize the publisher revenue, minimize the under-delivery of impressions to advertisers, and achieve desired user-level diversity.

Figure 1(b) shows the existing ad allocation framework. It contains the following steps.

1. **Identification of user visits:** The input to this module is the available ad slots. In this step, the probability of user visits to different ad slots is computed. The arrival of users to the different pages of the website is unpredictable. It would require identifying the arrival of specific user visits to the website to implement the ad allocation solutions. The output of this module is the supply nodes set (probability of user visits to ad slots). In the literature [7, 17, 26], several efforts have been made to predict or estimate the user visits for available ad slots using sampling procedures.
2. **Allocation approach:** The input to this step is the supply nodes and demands from advertisers. Based on the objective function, appropriate allocation algorithm is employed to allocate supply nodes to advertisers. Approaches to generate optimum ad allocation plans to achieve the objectives

of the defined problem is an important research issue. Research efforts [7, 17, 26] have also been made to develop theoretical solutions by modeling the ad allocation problem as a stochastic optimization problem.

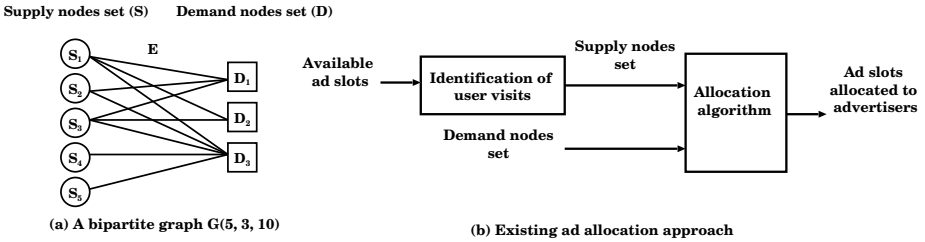


Fig. 1: Existing ad allocation framework

4 Proposed allocation framework

In this section, given a website, we first present the analysis of potential revenue from the impressions. Next, we explain the basic idea. Subsequently, we present the proposed ad allocation approach.

4.1 Analysis of potential revenue

Given a website which contains W web pages, the maximum revenue that can be generated from W is bound by the total number of page views ensured in a unit duration. The total number of pages views is equal to the summation of all individual page views of the website. It can be observed that, given a website, the revenue that can be extracted under the Cost Per Impressions (CPI) revenue model will be proportional to the total impressions served by all the ad slots associated with the web pages of the website in a unit duration. The ideal case would be when all the advertisers are allocated ad slots such that no two web pages in their allocated sets of ad slots are visited by the same user, and demanded impressions are exactly satisfied. This scenario is mathematically represented as follows.

If an advertiser D_i demands d_i impressions and a pattern of ad slots $S_i = \{s_j, s_{j+1}, \cdots, s_k\}$ is allocated, then the maximum revenue is obtained when the following conditions are satisfied.

$$Conditions = \begin{cases} T^{s_p} \cap T^{s_q} = \phi, \forall s_p, s_q \in S_i, p \neq q & \text{(4a)} \\ \sum\limits_{s_i \in S_i} |T^{s_i}| = d_i & \text{(4b)} \end{cases}$$

Here, given an ad slot x, T^x represents the set of transactions in which x has occurred.

The revenue of the publisher would be the maximum over the considered period of time when the preceding conditions are satisfied for all the advertisers and no web page is left unallocated. So, the maximum revenue is proportional to the total number of page views which is represented as follows.

$$\text{Maximum revenue} \propto \text{total number of page views} \tag{5}$$

$$\propto \sum_{i=1}^{|C|} |T^i| \tag{6}$$

$$= \beta \times \sum_{i=1}^{|C|} |T^i| \tag{7}$$

The proportionality constant β (> 0) depends on various factors. Some predominant factors include number of ad slots on each web page, ad geometry, ad frequency, ad pacing and bid of the advertisers [6]. Web pages with more advertising space yield more revenue. Bigger the size of the advertisement, it occupies more space and hence yields more revenue. Since the space is finite, each ad is displayed multiple times with pacing in line with the time spent on the web pages which helps in multiplying the revenue.

4.2 Basic idea

Website click stream data provides rich information about various dynamics of users who visit the website. It is possible to extract coverage patterns from a transactional database extracted from the website log files. A web page contains a set of ad slots. The notion of coverage support of a CP indicates distinct sets of ad slots, which ensures a certain percentage of views. Also, the overlap ratio of a CP indicates the degree of co-occurrence of ad slots in the same session of the user which can be exploited to achieve desired user-level diversity of ads by reducing the repeated display of the same ad. Most importantly, it can be noted that several low traffic ad slots can be part of a CP which can be used by the publisher to meet the demands of advertisers.

For example, consider an e-commerce website in which popular pages such as home page and frequently answered questions (FAQ) pages draw more users when compared to other product pages of the website. A typical user may visit more than one page, say p_1, p_2,..., p_n in a given session. If the publisher allocates p_1 and p_2 to the same advertiser and place the corresponding ad, the user encounters the same ad several times which leads to boredom. The coverage patterns extracted from click stream data allow the allocation of the ad slots of home page and FAQ page which are visited by the same group of users to different advertisers. The coverage patterns also allow combining the ad slots of frequently accessed web pages with ad slots of less popular web pages such that

the overlap is minimal, would result in displaying advertisement to a wide spectrum of website users and at the same time due to lesser overlap, repeatability can be minimized. As a result, the publisher would get an opportunity to satisfy demands of increased number of advertisers.

Since the coverage pattern contains ad slots with minimum overlap of click stream transactions, the impressions of a single user can be assigned to different advertisers. As a result, maximum number of impressions can be exploited for banner advertising. By carrying out allocations using coverage patterns having maximum coverage support and minimum overlap ratio, it is possible to maximize the revenue as given in Equation 4.

It can be noted that, in display advertising scenario, estimation of user visits to the ad slots is crucial for effective allocation. An improvement in the estimation of user visits to the website could improve the performance of allocation which in turn helps in realizing the objectives of both publisher and advertisers. Based on the knowledge of coverage patterns extracted from click stream transactions, it is possible to improve the performance of ad allocation. This is under the assumption that the knowledge of coverage patterns will be helpful to estimate the user visits. It can be observed that the patterns from click stream transactions are widely employed to recommend products in e-commerce environments. However, the investigation about predicting user visit behaviour using coverage patterns is investigated as a part of future work.

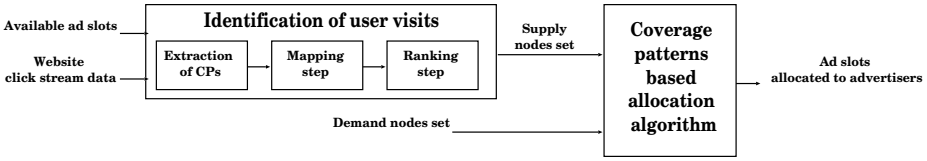


Fig. 2: Proposed ad allocation framework

4.3 Proposed ad allocation approach

The following assumptions are defined in the proposed allocation framework. We consider one ad slot per each web page and an advertiser's bid is a constant value predefined per impression. Advertisers demand the ad impressions to be shown and the publisher allocates a set of web pages (or ad slots) by considering the expected impressions in a unit duration. The duration can be fixed in terms of hours or days based on the user and advertiser dynamics [17].

In the proposed framework, the bipartite graph of display ad allocation is similar to the existing one shown in Figure 1(a). However, in the proposed framework, the supply nodes set S is formed and allocation will be carried out by exploiting the knowledge of coverage patterns from click stream data of a website.

Several issues need to be handled to form the supply nodes set using CPs in advertising scenario. Under the considered revenue model CPI, advertisers demand in terms of impressions. Hence, we propose a mapping methodology to convert coverage pattern with coverage support and overlap ratio into impressions. Also, a huge set of CPs are extracted from the click stream data of a website. Hence, we propose a ranking step to rank the CPs. Finally, an allocation algorithm is proposed to allocate appropriate set of ad slots to each advertiser.

The proposed approach is shown in Figure 2. It contains the following steps.

1. **Identification of user visits:** Available ad slots and the website click stream data are inputs to this step and output is supply nodes set. It performs the following sub steps to identify the supply nodes set from the click stream data of a website. We explain the steps in detail as follows.
   (a) **Extraction of CPs:** The input to this step is website click stream data considered over a period of time. From the website log data, click stream transactional database is generated in such a way that each transaction represents a user and the items in the transaction represents the pages visited by that user in a session. Complete set of CPs can be extracted from the click stream transactional data by employing the *CPPG* algorithm [23] with specified thresholds for RF, CS and OR parameters. The output of this step is a set of CPs extracted from the click stream transactional data. Each CP is a set of web pages. As we assume one ad slot is available on each web page, each CP can be referred as a set of ad slots along with the corresponding CS and OR values.
   (b) **Mapping step:** The CS of a CP provides the proportional value of number of transactions (unique visitors). The advertisers demand in terms of impressions. So, to facilitate the allocation, we need to estimate the number of impressions ensured by a CP to match with the demands of the advertisers for facilitating the allocation. We associate the value of impressions with each CP in this step. The number of impressions (I) ensured by a CP is equal to the summation of individual frequencies of the web pages in the CP multiplied with the respective number of ad slots available on each web page. The formula is as follows.

$$\text{Impressions}(S_i)=\text{I}(S_i)=\sum |T^{w_i}| * n_i, \forall w_i \in S_i \tag{8}$$

   where $S_i = \{s_j, s_{j+1}, \cdots, s_k\}$ is a supply node which is a CP of web pages, $|T^{w_i}|$ is the number of transactions in which the web page w_i occurred in the click stream transaction, n_i is the number of ad slots on the web page w_i and $I(S_i)$ is the number of impressions mapped to S_i. The output of this step is CPs with the corresponding value of impressions.
   (c) **Ranking step:** A large number of CPs are extracted from the click stream transactional data. It is possible that several hundreds of CPs match an advertiser demand. In such a case, it is important to allocate the most interesting CP first to the advertisers. A CP is interesting if

Table 2: Extracted CPs with CS, OR, $Impression(I)$, $Uniqueness(U)$

ID	CPs	CS	OR	I	U	ID	CPs	CS	OR	I	U
1	{b}	0.7	0.0	7	7	11	{d}	0.5	0.0	5	5
2	{b, a}	1.0	0.25	11	7.5	12	{d, a}	0.8	0.25	9	5.5
3	{b, c}	1.0	0.4	12	6	13	{d, a, e}	1.0	0.5	13	5
4	{b, e}	0.9	0.5	11	4	14	{d, a, f}	1.0	0.3	12	7
5	{c}	0.5	0.0	5	5	15	{d, e}	0.9	0.0	9	9
6	{c, e}	0.7	0.5	9	2	16	{d, f}	0.8	0.0	8	8
7	{c, d}	0.8	0.4	10	4	17	{a}	0.4	0.0	4	4
8	{c, d, e}	1.0	0.5	14	5	18	{a, e}	0.6	0.5	8	1
9	{c, d, f}	1.0	0.3	13	7	19	{a, f}	0.6	0.3	7	3
10	{c, f}	0.7	0.3	8	4	20	{e}	0.4	0.0	4	4

it has high coverage support and low overlap ratio. This is due to the fact that high CS value reflects more unique users and low OR value reflects less repetition or together occurrence of the web pages in the same session. We define the interestingness measure *Uniqueness (U)*, to capture both the aspects of high CS and low OR of a CP which is equivalent to the difference between CS and OR parameters. The difference gives a proportional measure which is equal to the number of unique visitors visiting the corresponding ad slots of a CP. The formula is as follows.

$$\text{Uniqueness}(S_i) = \text{U}(S_i) = |CS(S_i) - OR(S_i)| * |\text{C}| \tag{9}$$

where S_i is a CP, $|C|$ is the transactional database size, CS and OR is the overlap ratio and.

2. **Allocation step:** Supply nodes set generated from the preceding step and demand nodes set is the input to this step. The approach first sorts the S set in the increasing order of impressions. Next, the patterns with equal number of impressions are sorted based on uniqueness measure. The algorithm considers each D_i in D set to identify eligible appropriate supply nodes. Out of eligible supply nodes, the supply node with weight close to the demand of D_i is allocated and an edge is formed between the demand node D_i and eligible supply node S_k. The procedure is repeated by considering the remaining supply nodes till no demand node is left or remaining supply nodes are unable to satisfy any demand further. The allocation algorithm (CPs-based allocation) is given in Algorithm 1.

Example 2: We explain the proposed approach by considering the click stream transactional data of the Table 1. We consider a set of three advertisers $D_1,\ D_2,\ D_3$ coming with impressions demands of $d_1 = 8,\ d_2 = 11,\ d_3 = 9$ respectively. Here $d_1,\ d_2,\ d_3$ are the weights associated with $D_1,\ D_2,\ D_3$ in the set D.

The algorithm of $CPPG$ [23] is employed with the parameters RF=0.3, CS=0.4 and OR=0.5 on the Table 1 to extract CPs. The CPs which satisfy

Algorithm 1 CPs based allocation algorithm

Input: D: Demand nodes set, W: Website click stream data
Output: O: Ad slots allocated to advertisers.

procedure ALLOCATION(D, W)
  $\mathcal{S} \leftarrow CoveragePatterns(W)$ ▷ $\mathcal{S}$ is a supply nodes set
  for $i \leftarrow 1, |D|$ **do**
    $C \leftarrow arg\min_{C \in \mathcal{S}} |Supply(C) - D_i|$ ▷ C is a supply node
    $O[i] \leftarrow O[i] \cup \{C\}$
    $\mathcal{R} \leftarrow \{r \in \mathcal{S} : \exists w \in W\{w \in C\}\}$ ▷ R is an allocated supply nodes set
    $\mathcal{S} \leftarrow \mathcal{S} \backslash \mathcal{R}$
  end for
  return O
end procedure

the specified threshold requirements are given in Table 2. The algorithm has extracted 20 CPs from Table 1 with the length of the patterns varying from 1 to 3 for the specified parameters.

As the advertisers demand in terms of impressions in the considered model, we need to estimate and map the impressions ensured by each extracted CP. We associate the value of impressions to each extracted CP using the formula given in the mapping step. For example, consider the CP {d, a, e} from the extracted CPs. The number of impressions ensured by {d, a, e} using the above mentioned formula is the summation of individual frequencies of web pages appearing in the CP. Hence impressions ensured by {d, a, e} is $|T^d + T^a + T^e| = 13$. Similarly, the number of impressions are mapped to each extracted CP which forms the supply set S with respective mapped impressions as weights.

The uniqueness measure for each CP is calculated using the formula given in the ranking step of the algorithm. For example, consider the CP {c, d, e} from the extracted CPs. The uniqueness measure of {c, d, e} using the above mentioned formula is the difference between the CS and OR values associated with it and multiplied by $|C|$. From the Table 2, the CS and OR of {c, d, e} are 1.0 and 0.5 respectively and $|C|$ is the total number of transactions in Table 1. Hence uniqueness measure of {c, d, e} is $|1.0 - 0.5| * 10 = 5.0$. The values of impressions (I) and uniqueness measure (U) for the rest of the extracted CPs are calculated similarly and given in the Table 2.

The proposed algorithm allocates the demands in first-come-first-serve basis. In the first iteration, the algorithm identifies an eligible subset of S with weights able to satisfy the demand $d_1 = 8$ of the first advertiser D_1. The patterns or the supply nodes {c, f}, {d, f} and {a, e}, all ensure 8 impressions, are eligible to satisfy the demand and hence picked by the algorithm. From Table 2, the uniqueness values of {c, f}, {d, f} and {a, e} are 4, 8 and 1 respectively. Hence, the CP {d, f} is allocated to the advertiser D_1. In the second iteration, the algorithm identifies a possible subset of S with weights eligible to satisfy the demand $d_2 = 11$ of the second advertiser D_2. The patterns or the supply nodes {b, a}, {b, e} are eligible to satisfy the demand and hence picked by the algorithm. The

CP {b, a} has high uniqueness measure and hence allocated to the advertiser D_2. In the third iteration, the algorithms picks and allocates the CP {c, e} close to the demand $d_3 = 9$ to the third advertiser D_3 similarly. The algorithm terminates as no advertiser is left to allocate in this example.

5 Experiments

In this section, we explain datasets, approaches implemented, performance metrics and results.

5.1 Description of datasets

To evaluate the proposed approach, three types of click stream datasets and simulated advertiser dataset are being used.

T40I10D100K dataset (TIK) [1]: It is a synthetic dataset. The format of the dataset is {Item ID1, Item ID2,..., Item IDk}. Each record corresponds to one distinct transaction. We consider each record as distinct session and the items comprise each distinct session as the web pages visited by the visitor during that session. We consider 100000 transactions.

Kosarak dataset (KSK)[1]: It is an anonymized click stream transactional dataset of a Hungarian online news portal. The format of the dataset is {Item ID1, Item ID2,..., Item IDk}. Each record corresponds to one distinct session where the items are the web pages visited by the visitor during that session. We consider 20000 sessions of the dataset for the experiment.

Yoochoose dataset (YCT)[4]: It is a real world click stream dataset of an e-commerce website provided by a retailer. The dataset comprises of click-through transactions in the form of web sessions. The format of the dataset is $\langle$ *Session ID*, *Timestamp*, *Item ID*, *Category* $\rangle$. We consider 10000 sessions of the dataset for the experiment.

The statistics of the datasets are given in the Table 3. It should be noted that the three datasets used in the experiments capture different aspects of click-through behaviour.

Simulated advertisers dataset: The datasets available are either click-stream datasets of websites or individual advertiser datasets for sponsored search advertising. So we have created an advertiser dataset to run experiments on both the datasets. For each dataset, five advertiser datasets were simulated containing 20, 40, 60, 80 and 100 advertisers. For creation of the dataset, impressions required by each advertiser were generated randomly satisfying the condition that every advertiser in the corresponding dataset can be satisfied within the available impressions of the respective click stream data. For example, in the 4th column of Table 3, the average impressions in the dataset 20 is 1325.55. The impressions demanded by the 20 advertisers are generated around that average between the maximum and the minimum impressions available in the respective click stream datasets. Similarly, the impressions are generated for the rest of the datasets. The statistics of the average impressions demands are provided in

Table 3 for TIK, KSK and YCT datasets respectively. The upper bound on the advertisers demands data size is set by analyzing the website click stream data of respective datasets such that we can exploit all available page views to achieve a comparison between both the approaches. For each advertisers dataset, the impressions demands are uniformly generated around the average impressions ensured in the respective click stream transactional dataset. This indicates that the advertising load on the system is increasing with a uniform distribution of advertisers demands.

Table 3: Statistics of click stream transactions and Advertisers demands data

Dataset	Page views	Average transaction length	Advertisers demands dataset				
			20	40	60	80	100
TIK	131530	13.15	1325.55	1328.35	1344.02	1294.30	1310.30
KSK	162513	8.12	1746.95	1675.10	1710.50	1728.20	1607.27
YCT	62861	6.28	1045.65	1003.75	1013.86	725.42	626.66

5.2 Approaches implemented

In the literature [7, 17, 27], the work on display advertising is focused upon mathematical formulations of display ad scheduling by defining parameters and carried out experiments by randomly generated numerical values that simulate appropriate scaled versions of a real world data instances of a website. CPs based allocation is the first allocation approach proposed using click stream transactional database of a website. Consequently, there are no concrete existing benchmarks for comparing the performance. As the earlier approaches dealt the problem by considering simulated user visit frequencies, we compare the proposed approach with the same underlying notion as our baseline approach and hereafter referred as visit frequency based allocation approach (VF-based allocation).

For the experiment, the VF-based allocation approach is implemented in the following manner. Each advertiser comes with a demand of certain impressions. The visit frequency of a web page is the number of times the web page has appeared in the click stream transactional data considered over a period of time. Thus, visit frequency of each web page gives the number of impressions ensured by that web page. VF-based allocation allocates web pages by considering the visit frequency of individual web pages. Given an advertisers impressions demand d_i, the approach selects a set of web pages such that the total number of expected visits (or impressions) is greater than or equal to the request made by the advertiser. This approach is repeated for every advertiser till there are no advertisers left or it is not possible to identify a set of web pages that meet the requests of rest of the advertisers.

For the experiment, to extract CPs from the click stream transactional database, we have set $minRF = 0$ to engage every web page. Also, we set $minOR = 0.5$

and $minCS$ = average value of $minRF$ in the respective datasets. The CPPG [23] algorithm is employed to extract CPs with the respective thresholds.

5.3 Performance metrics

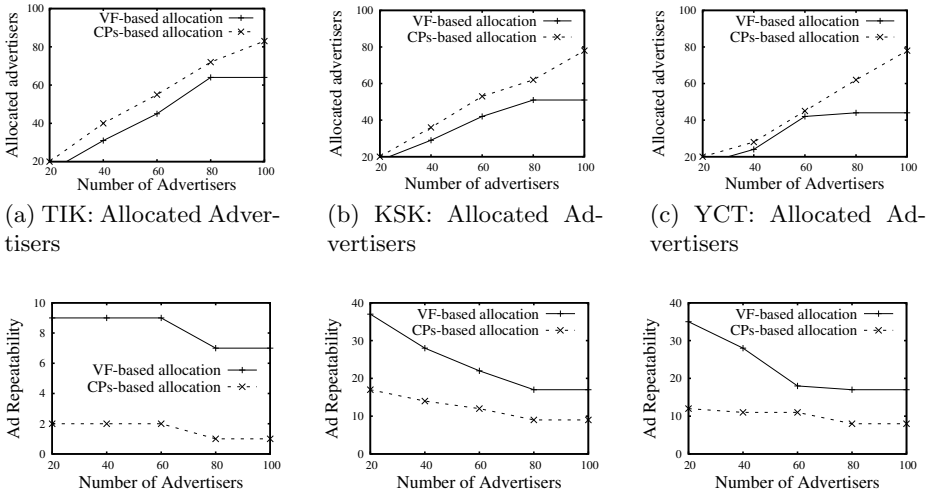
We employ two measures to compare the performance.

- **Allocated Advertisers:** It is equal to the number of allocated advertisers per unit time duration. At most, the number of allocated advertisers is equal to the number of advertisers who have requested the ad slots. The approach which gives the high value of allocated advertisers is better as it meets the demands of more number of advertisers which helps in maximizing the revenue.
- **Ad Repeatability (AR):** Ad repeatability measure indicates how many times the same advertisement appeared in the same session of a user. It is measured by considering a set of ad slots in which at least two occur in the same session of a user. It captures the maximum co-occurrence of ad-slots on which the same ad is being displayed to the same user. For the advertiser, less ad repeatability is preferred to maximize the reach. In addition, the boredom of the visitor is directly proportional to the value of the metric i.e., higher the value of ad repeatability more the boredom of the visitor. The lower value of the metric signifies the efficient utilization of ad slots which indicates improved user-level diversity (displaying diverse set of ads to users). For a pattern $P_i = \{a_i, a_{i+1}, \cdots, a_j\}$, where $a_i's$ are ad slots, the metric is defined as follows:

$$AR(P_i) = \begin{cases} 0 & \text{if } |P_i| = 1 \\ |T^{a_i} \cap T^{a_j}| & \text{if } |P_i| = 2 \\ \max\limits_{a_i, a_j \in P_i, a_i \neq a_j} AR(\{a_i, a_j\}) & \text{if } |P_i| > 2 \end{cases}$$

5.4 Results

Figure 3a, Figure 3b and Figure 3c show the performance results of allocated advertisers for the experiments conducted on TIK, KSK and YCT datasets respectively. It can be observed that initially both the approaches could meet the demands of advertisers. However, as the number of advertisers is increasing, it can be observed that the VF-based allocation could not allocate more number of advertisers whereas CPs-based allocation is meeting the demands of increased number of advertisers. This is because the VF-based allocation carries out allocation based on user visit frequencies and by arbitrarily combining the web pages. As a result, there is a high possibility that the same user will see the advertisement multiple times which leads to under utilization of user views. The CPs-based allocation carries out allocation based on the interesting coverage patterns which has maximum coverage support and minimum overlap ratio.



(a) TIK: Allocated Advertisers (b) KSK: Allocated Advertisers (c) YCT: Allocated Advertisers

(d) TIK: Ad Repeatablity (e) KSK: Ad Repeatablity (f) YCT: Ad Repeatablity

Fig. 3: Performance results on TIK, KSK and YCT datasets

As a result, multiple advertisements are shown to the same visitor which leads to an efficient utilization of user views.

Figure 3d, Figure 3e and Figure 3f show the performance results of ad repeatability for the experiments conducted on TIK, KSK and YCT datasets respectively. It can be observed that, the average ad repeatability of the both approaches is decreasing with the number of advertisers. This is due to the fact that as the number of advertisers increases, the user may encounter the same advertisement multiple times. However, it can be observed that the ad repeatability reduced significantly in the proposed approach as compared to the VF-based allocation. The proposed allocation approach ensures that there is less user overlap among the web pages of a CP by carrying out the allocation based on the coverage patterns with high value of uniqueness measure thereby leading to more unique visits to a web page. As a result, ad repeatability is significantly reduced in the proposed allocation approach.

6 Conclusion and Future Work

In display advertising scenario, the objective of publisher is to maximize the revenue by meeting the advertising demands of increased number of advertisers. In this paper, we have proposed an improved allocation framework for display advertising scenario by exploiting the knowledge of coverage patterns extracted from the click stream transactions of the website. We have discussed how the nature of coverage patterns could improve the efficiency of ad slots allocation to the advertisers. We have developed the allocation framework based on coverage

patterns by adding mapping, ranking steps and allocation algorithm. The experimental results on both synthetic and real world click stream datasets show that the proposed allocation approach meets the demands of increased number of advertisers as compared to the baseline approach by improving the user-level diversity and reducing the repeated display of ads.

As a part of future work, a part from carrying out intensive experiments, we are planning to explore how the knowledge of content-specific coverage patterns improves the efficiency of ad allocation. We will investigate how the knowledge of coverage patterns captures the dynamics of user visit behavior. We will also explore how the knowledge of coverage patterns will be useful in improving the efficiency of coverage patterns in both guaranteed and non-guaranteed contract scenario.

References

1. Frequent itemset mining implementations repository, http://fimi.cs.helsinki.fi/
2. Double click (2015), http://www.doubleclick.net
3. Interactive advertising bureau (2015), http://www.iab.net
4. Recsys challenge 2015 (2015), http://2015.recsyschallenge.com/challenge.html/
5. Right media (2015), https://en.wikipedia.org/wiki/Right_Media
6. Adler, M., Gibbons, P.B., Matias, Y.: Scheduling space-sharing for internet advertising. Journal of Scheduling 5(2), 103–119 (2002)
7. Bharadwaj, V., Chen, P., Ma, W., Nagarajan, C., Tomlin, J., Vassilvitskii, S., Vee, E., Yang, J.: Shale: an efficient algorithm for allocation of guaranteed display advertising. In: The 18th International Conference on Knowledge Discovery and Data mining. pp. 1195–1203. ACM (2012)
8. Budhiraja, A., Reddy, P.K.: An approach to cover more advertisers in adwords. In: The 2nd International Conference on Data Science and Advanced Analytics. pp. 1–10. IEEE (2015)
9. Caruso, F., Giuffrida, G., Zarba, C.: Heuristic bayesian targeting of banner advertising. Journal of Optimization and Engineering 16(1), 247–257 (2015)
10. Chvatal, V.: A greedy heuristic for the set-covering problem. Mathematics of Operations Research 4(3), 233–235 (1979)
11. Feige, U., Immorlica, N., Mirrokni, V., Nazerzadeh, H.: A combinatorial allocation mechanism with penalties for banner advertising. In: The 17th International Conference on World Wide Web. pp. 169–178. ACM (2008)
12. Feldman, J., Henzinger, M., Korula, N., Mirrokni, V.S., Stein, C.: Online stochastic packing applied to display ad allocation. In: The 18th Annual European Conference on Algorithms: Part I. pp. 182–194. Springer (2010)
13. Feldman, J., Mehta, A., Mirrokni, V., Muthukrishnan, S.: Online stochastic matching: Beating 1-1/e. In: The 50th Annual Symposium on Foundations of Computer Science. pp. 117–126. IEEE (2009)
14. Garey, M.R., Johnson, D.S., Stockmeyer, L.: Some simplified np-complete problems. In: The 6th Annual ACM Symposium on Theory of Computing. pp. 47–63. ACM (1974)
15. Ghosh, A., McAfee, P., Papineni, K., Vassilvitskii, S.: Bidding for representative allocations for display advertising. In: The 5th Workshop on Internet & Network Economics. pp. 208–219. Springer (2009)

16. Han, J., Chang, C.C.: Data mining for web intelligence. Computer 35(11), 64–70 (2002)
17. Hojjat, A., Turner, J., Cetintas, S., Yang, J.: Delivering guaranteed display ads under reach and frequency requirements. In: The 28th AAAI Conference on Artificial Intelligence. pp. 2278–2284. AAAI Press (2014)
18. Huang, e., Cercone, N., An, A.: Comparison of interestingness functions for learning web usage patterns. In: The 11th International Conference on Information and Knowledge Management. pp. 617–620. ACM (2002)
19. Mirrokni, V.S., Gharan, S.O., Zadimoghaddam, M.: Simultaneous approximations for adversarial and stochastic online budgeted allocation. In: The 23rd Annual Symposium on Discrete Algorithms. pp. 1690–1701. ACM-SIAM (2012)
20. Nakamura, A., Abe, N.: Improvements to the linear programming based scheduling of web advertisements. Journal of Electronic Commerce Research 5(1), 75–98 (2005)
21. Srinivas, P.G., Reddy, P.K., Sripada, B., Kiran, R.U., Kumar, D.S.: Discovering coverage patterns for banner advertisement placement. In: The 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 133–144. LNCS (2012)
22. Srinivas, P.G., Reddy, P.K., Trinath, A.V., Sripada, B., Kiran, R.U.: Mining coverage patterns from transactional databases. Journal of Intelligent Information Systems 45(3), 423–439 (2015)
23. Srinivas, P.G., Reddy, P.K., Trinath, A.: Cppg: Efficient mining of coverage patterns using projected pattern growth technique. In: International Workshops - Trends and Applications in Knowledge Discovery and Data Mining - PAKDD. pp. 319–329. LNCS (2013)
24. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations News Letters 1(2), 12–23 (2000)
25. Trinath, A., Gowtham Srinivas, P., Krishna Reddy, P.: Content specific coverage patterns for banner advertisement placement. In: The 1st International Conference on Data Science and Advanced Analytics. pp. 263–269. IEEE (2014)
26. Vee, E., Vassilvitskii, S., Shanmugasundaram, J.: Optimal online assignment with forecasts. In: The 11th Conference on Electronic Commerce. pp. 109–118. ACM (2010)
27. Yang, J., Vee, E., Vassilvitskii, S., Tomlin, J., Shanmugasundaram, J., Anastasakos, T., Kennedy, O.: Inventory allocation for online graphical display advertising. CoRR: Computing Research Repository (2010)