An improved approach for long tail advertising in Sponsored Search

by

amar.budhiraja , P Krishna Reddy

in

22nd International Conference on Database Systems for Advanced Applications
(*DASFAA-2017*)

Report No: IIIT/TR/2017/-1

Centre for Data Engineering
International Institute of Information Technology
Hyderabad - 500 032, INDIA
March 2017

An improved approach for long tail advertising in Sponsored Search

Amar Budhiraja and P. Krishna Reddy

FC Kohli Center in Intelligent Systems,
IIIT-Hyderabad, Hyderabad, India - 500032.
amar.budhiraja@research.iiit.ac.in, pkreddy@iiit.ac.in

Abstract. Search queries follow a long tail distribution which results in harder management of ad space for sponsored search. During keyword auctions, advertisers also tend to target head query keywords, thereby creating an imbalance in demand for head and tail keywords. This leads to under-utilization of ad space of tail query keywords. In this paper, we have explored a mechanism that allows the advertisers to bid on concepts rather than keywords. The tail query keywords are utilized by allocating a mix of head and tail keywords related to the concept. In the literature, an effort has been made to improve sponsored search by extracting the knowledge of coverage patterns among the keywords of transactional query logs. In this paper, we propose an improved approach to allow advertisers to bid on high level concepts instead of keywords in sponsored search. The proposed approach utilizes the knowledge of level-wise coverage patterns to allocate incoming search queries to advertisers in an efficient manner by utilizing the long tail. Experimental results on AOL search query data set show improvement in ad space utilization and reach of advertisers.

Keywords: Data Mining, Computational Advertising, Coverage Patterns, Pattern Mining, Sponsored Search.

1 Introduction

Sponsored search is one of the most dominant mediums to advertise on the web. In sponsored search, advertisers create ad campaigns and bid on keywords that they deem relevant to their product. For an incoming search query, advertisements from the ad campaigns containing the query keywords are shown along with the search results. If multiple advertisers demand to be shown on the same query's results page, they are ranked for the allocation of ad space. The ranking is based on multiple factors including the bid amount of the advertiser on the query keywords, relevance of ad content to the search query, Click-Through-Rate (CTR) and budget of the advertiser.

It has been established that search queries follow a long tail distribution of a small but fat head of frequent queries and a long-thin tail of infrequent queries [5, 8]. Advertising on tail queries is challenging as tail queries are encountered rarely which makes them harder to interpret for sponsored search. Also, it has been observed that during keyword auctions, advertisers tend to bid for the head keywords to reach more users. This creates a high demand for the head query keywords and little to no demand for the tail query keywords [5]. The long tail phenomenon also makes it quite difficult to capture the relevant keywords from the long tail. The above stated factors result in under-utilization of a significant amount of the ad space provided by tail queries in sponsored search which is identified as the research issue.

In this paper, we propose an approach to exploit the long tail of the search query keywords for sponsored search. We propose that instead of bidding on search query keywords, advertisers should bid upon high level concepts. The motivation for bidding on concepts is inspired from the trends in advertising on social media[1]. In social media advertising, advertisers target concepts beyond keywords such as photography, reading, travelling, lifestyle, etc. In sponsored search, bidding on concepts will result in capitalization of ad space of the tail queries as all the keywords would be considered based on the relevancy rather than frequency. Bidding on concepts instead of keywords would also ensure that the advertisers do not have to retrieve all the search keywords from the long tail.

In this paper, we propose an allocation mechanism for sponsored by considering concepts as bidding units rather than search keywords. We propose that during ad campaign creation, an advertiser is shown a taxonomy based on the content of the ad and is asked to select a concept in the shown taxonomy that seems to be the most relevant to the product. We propose an approximate allocation between the nodes of the taxonomy and the advertisers. To acknowledge the long tail phenomenon, we extract knowledge from search query logs using the notion of Coverage Patterns (CPs). In the literature, approaches to extract CPs have been proposed. Given a database of transactions, a CP is a set of items such that it covers a certain percentage of transactions having given overlap ratio [1, 4]. By extending the notion of CPs, an effort has been made in the literature to propose allocation approach to improve the performance of Adwords [3] and display advertising by assuming that an advertiser requests a set of keywords [11]. In this paper, taking query logs and taxonomy as input, we propose a new approach to extract the knowledge of level-wise CPs and use the corresponding framework to allocate incoming queries to ads based on the high-level concepts requested by the advertisers. The proposed approach is compared against traditional sponsored search model. Experiments on the real world data set of AOL search query logs show the improvement in ad space utilization and reach of advertisers.

The remainder of this paper is organized as follows: in Section 2, we review the related work in the context of coverage patterns and long tail advertising in search engines; in Section 3 we discuss the background on coverage patterns and sponsored search; in Section 4 we discuss the basic idea followed by the proposed approach in Section 5; experiments are discussed in Section 6, followed by conclusions and a discussion on future work in Section 7.

[1] ads.twitter.com

2 Related Work

In the literature, challenges of tail queries in sponsored have been primarily addressed by means of query expansion [5–7]. In [6], the authors formulated a taxonomy based model to classify search queries, specifically tail queries. Organic clicks were used as blind feedback mechanism to learn the model. The authors explored its feasibility on search advertisement relevance. In another study [7], the authors expand search queries by adding multiple features including category of retrieved web pages and salient named entities. Furthermore, the authors propose an approach in [5] to expand tail queries in real time using an inverted index build from head and torso *expanded* queries. Using the expanded queries, the authors show improvement in ad retrieval.

Sponsored search has been also explored from the perspective of revenue optimization. In [9], it was modelled as an online bipartite matching problem such that advertisers are one set of disjoint vertices and queries are the other disjoint set. They developed an algorithm for advertisement allocation of incoming queries to optimize the revenue of the search engine. This bipartite approach is a high level architecture for Adwords, Google's sponsored search. A more detailed survey of the related literature [12] explains multiple models of bipartite graph matching with its context as Adwords, including algorithms from display ads and welfare maximization.

The model of coverage patterns has been proposed in the literature in the form of an apriori style approach proposed in [1] followed by a pattern growth approach in [4]. Coverage Patterns have been employed in improvement of delivering guaranteed contracts in display advertising [11] and in coverage of more advertisers in Adwords [3].

In this paper, we propose a framework to capitalize the long tail of search queries. We extend the bipartite model discussed in [9, 12] into an end-to-end approach. The proposed approach is different from [5] as the authors propose to capture tail queries using a taxonomy by generalizing a query into a taxonomy node. However, the taxonomy was not exposed to the advertisers and was only employed internally whereas in this paper, we propose a mechanism to allow advertisers to bid on concepts by showing a taxonomy related to their ads. In [3], the authors used coverage patterns to group similar keywords but the model to group keywords was employed by abstracting similar keywords only to a single concept rather than a hierarchical relationship of taxonomy, as proposed in this paper. It should be noted that the previous approaches [3, 5–7] have emphasized on keyword analysis or query expansion where in this paper we present an alternative approach of bidding on concepts rather than keywords in sponsored search.

3 Background: Sponsored Search and Coverage Patterns

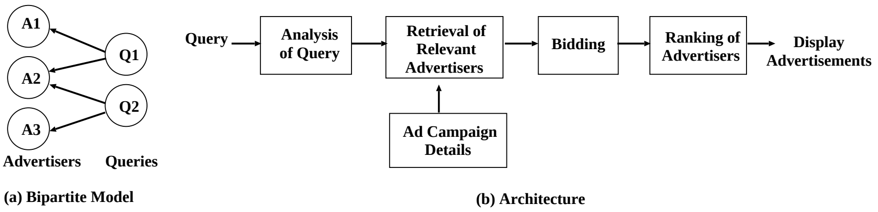
In this section, we briefly explain the sponsored search framework and notion of coverage patterns.

3.1 Sponsored Search Background

The standard model for sponsored search is a bipartite model as shown in Fig. 1(a) such that each incoming query is matched to an advertiser based on certain constraints. These constraints are defined by multiple parameters including relevance score, bid of the advertiser on the query keywords and remaining budget of the advertiser.

The architecture of sponsored search for advertisement allotment has four main steps [12] as shown in Fig. 1(b):

1. **Analysis of Query:** In the first step, the query is analysed to extract important parameters such as session information to better serve advertisements.
2. **Retrieval of Relevant Advertisers:** Based on the query keywords and other query parameters learnt from Step 1, relevant advertisers are retrieved from ad campaigns which are to be considered for displaying alongside organic results.
3. **Bidding:** Due to competition among advertisers, incoming queries are allotted to advertisers through auctions such that advertisers bid for placing their ads on the query page. These bids can either be static or can be done in real time.
4. **Ranking of Advertisers:** Once the advertisers bid on a query, their bids are scaled according to a factor called *Quality Score*. The *Quality Score* is computed based on the parameters related to the respective advertisement. This includes expected *Click Through Rate (CTR)*, display URL's past *CTR*, quality of the landing page, remaining budget and advertisement/search relevance apart from several other parameters.



(a) Bipartite Model

(b) Architecture

Fig. 1: Sponsored Search: Model and Architecture

3.2 Coverage Patterns

In this section, we briefly explain about the notion of coverage patterns [1, 4]. Let $W = \{w_1, w_2...w_n\}$ be set of web pages and D be a set of transactions such that each transaction T is a set of web pages $T \subseteq W$. X is a pattern of web pages such that $X \subseteq W$ and $X = \{w_p, ...w_q, w_r\}$ where $1 \leq p \leq q \leq r$. T^{w_i} denotes a set of transactions containing the web page w_i and its cardinality is denoted $|T^{w_i}|$.

The fraction of transactions containing a web page w_i is called as the *Relative Frequency* of w_i and is calculated as $RF(w_i) = \frac{|T^{w_i}|}{|D|}$. A web page is considered frequent if it has a relative frequency greater than the threshold value, *minRF*. Coverage Set of a pattern $X = \{w_p, ...w_q, w_r\}$, $CSet(X)$ is a set of all transactions that contain at least one web page from the patterns i.e $CSet(X) = T^{w_p} \cup ...T^{w_q} \cup T^{w_r}$ such that $|T^{w_p}| > ... > |T^{w_q}| > |T^{w_r}|$. Coverage Support, $CS(X)$ is the ratio of size of $CSet(X)$ to size of D i.e., $CS(X) = \frac{|CSet(X)|}{|D|}$. Overlap ratio of a pattern X, $OR(X)$ is the ratio of number of transactions that are common between $X - w^r$ and w^r to the number of transactions in w_r i.e, OR(X) = $\frac{CSet(X-w^r) \cap CSet(w^r)}{CSet(w^r)}$.

A pattern is interesting if it has a high CS and low OR. A high CS value indicates more number of visitors and a low OR value means less repetitions amongst the visitors. Hence, a pattern is said to be interesting if $CS(X) > minCS(X)$, $OR(X) < maxOR$ and $RF(w^i) > minRF \ \forall \ w^i \in X$.

Table 1: Sample Transactions

TID	1	2	3	4	5	6	7	8	9	10
Pages	{a,b,c}	{a,c,e}	{a,c,e}	{a,c,d}	{b,d,f}	{b,d}	{b,d}	{b,e}	{b,e}	{a,b}

Example 1: To explain the notion of coverage patterns, we will consider a transactional database $|D|$ shown in Table 1. Let us assume the minRF to be 0.2, minCS to be 0.3 and maxOR to be 0.5. From Table 1, the number of transactions having a, T^a is 5, T^b is 7 and f, T^f is 1 . So, RF for a is 0.5, for b is 0.7 and for f is 0.1, f will be removed. On the other hand, a and b satisfy the constraint of minRF and therefore, {b,a} is a candidate set. The order of items in a coverage pattern is in decreasing order of the relative frequency and hence, the pattern is {b,a} and not, {a,b}. The Coverage Set for {b,a} is {1,2,3,4,5,6,7,8,9,10} and $|CSet\{b, a\}|$ is 10. So, coverage support of {b,a} is $\frac{10}{10}$ is 1 which is greater than $minCS$. The transactions containing {b,a} together is {1,10} and $T^a = 5$, so the overlap ratio is $\frac{2}{5} = 0.4 < maxOR$ and hence, {b,a} is a coverage pattern.

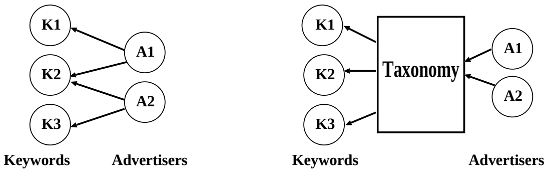
Thus, coverage patterns helps in extracting multiple sets of mutually exclusive subsets of items corresponding to coverage support and overlap ratio. In the literature, it has been demonstrated that coverage patterns can help in covering more advertisers and improve the diversity of viewers of individual ads [3, 11].

4 Basic Idea

The long tail phenomenon of search queries makes them unpredictable for sponsored search which is identified as the research issue. Advertisers also tend to target head query keywords during keyword auctions in order to cover more eye balls. However, this leads to a high demand for head keywords while little to no demand for the tail keywords. This imbalance in demand results in underutilization of ad space of the tail keywords. Hence, an opportunity has been identified to capitalize this long tail of search query keywords.

We propose that for sponsored search, advertisers should bid upon high level concepts instead of specific keywords. Bidding on high level concepts will result in capitalization of the ad space of the tail keywords as the keywords would be considered to be allocated based on relevancy rather than frequency.

In the proposed approach, we achieve bidding on concepts such that an advertiser would be shown a taxonomy based on his/her advertisement content. The advertiser is then asked to select a node in the taxonomy which he/she deems the most relevant for the advertisement. For example, an advertiser like *Amazon.com* would be shown at taxonomy of *Shopping* and based on the advertisement, the advertiser can select the appropriate node. If the advertisement is of books, the advertiser would select the node *Books* in the *Shopping* taxonomy or if the ad is related to clothing, the advertiser would choose to bid upon *Clothing or Fashion.* Thus, we propose to add a middle layer of concepts through a taxonomy during the bidding process such that an advertiser would chose a concept which would ultimately translate to a set of keywords, compared to the present approach where the advertiser is responsible for selecting all the desired keywords (Fig. 2).



(a) Keyword Bidding Mechanism

(b) Concept Bidding Mechanism

Fig. 2: Sponsored search bidding: keywords based bidding and concept based bidding

We propose an estimated allocation model based on the concept bidding such that groups immediate children nodes of bidding node are allocated to advertisers. Allocation of only children nodes of the bidding node is done to ensure that the allocation mechanism should consider the amount of generalization requested by the advertiser. For example, an advertiser who chose to bid upon *Shopping* should not be allotted something like {*Outwear, Skirts, Shirts*} as he would like to show his ad to a larger audience consisting of *Books, Clothing, Electronics, etc.*

To create such combinations of children nodes, we employ the notion of Coverage Patterns (CPs) such that CPs are extracted from the query logs and a matching is performed between the CPs at each node and the corresponding advertisers. When a query is posed by a user to the search engine, it is classified into these concepts according to the taxonomy and the advertisers who have been allocated any of these concepts are eligible to be ranked for the query.

To address the issues of allocation for a multiple level taxonomy, we propose an approach to extract CPs with respect to the taxonomy followed by an allocation approach for advertisers using the extracted coverage patterns.

4.1 T-Cmine: Extraction of Coverage Patterns with respect to a Taxonomy

In [2], an approach to extract generalized frequent patterns has been proposed. Similarly, we propose a methodology to extract coverage patterns involving the nodes of taxonomy by extending Cmine algorithm [4]. For a given transactional database $\mathcal{D}$ and the taxonomy $\mathcal{T}$ which relates the items of $\mathcal{D}$, we modify each transaction by appending the ancestors of each item in the transaction to the transaction. If we apply Cmine to this modified dataset several coverage patterns containing high-level as well as low level items would be extracted. Such patterns may not be useful for ad allocation. We are interested in the coverage patterns which contains the items at the same level and satisfy the following property.

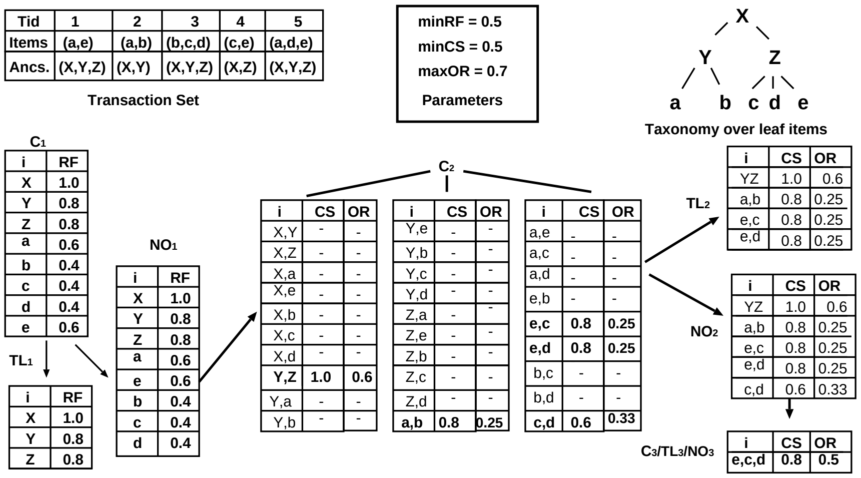
$$CP = \{c \mid c \in (\mathcal{I} \cup \mathcal{T}) \ \& \ \forall \ c \ parent(c) = P\} \tag{1}$$

Here, CP is coverage pattern containing items c such that all items belong to the same parent P. To extract level-wise coverage patterns, we propose the T-Cmine algorithm which is as follows.

Algorithm 1 T-Cmine: Algorithm to extract Coverage Patterns with respect to a Taxonomy

```
Input: $\mathcal{D}$, dataset of transactions; $\mathcal{T}$, Taxonomy defined over items of $\mathcal{D}$;
Compute $\mathcal{D}*$ from $\mathcal{T}$ by appending ancestors to $\mathcal{D}$;
$TL_1 := \{frequent1\ itemsets\}$;
$NO_1 := \{frequent1\ itemsets\}$;
$C_2 := NO_1 \bowtie NO_1$;
$TL_2$ := Remove any patterns from $C_2$ which contain items other than sister nodes;
$TL_2$ := Remove any patterns from $TL_k$ which do not satisfy $minCS$, $maxOR$ property;
$NO_2$ := Remove any patterns from $TL_k$ which do not satisfy $maxOR$ property;
$k := 3$
while $TL_{k-1} \neq \phi$ do
    $C_k := NO_{k-1} \bowtie NO_{k-1}$;
    $TL_k$ := Remove any patterns from $TL_k$ which do not satisfy $minCS$, $maxOR$ property;
    $NO_k$ := Remove any patterns from $TL_k$ which do not satisfy $maxOR$ property;
end
```

The proposed algorithm takes the dataset $\mathcal{D}$ and a taxonomy $\mathcal{T}$ that defines the relationship between the items of the $\mathcal{D}$. The algorithm first adds ancestors of each item in a transaction to the transaction. Then, the first set of CPs (TL_1) is calculated by getting the frequent items for which relative frequency is greater than *minRF*. The same set (TL_1) is also considered as Non-Overlapping Patterns set (NO_1). Using the (NO_1), candidate-2 coverage patterns are computed in the same way as Cmine algorithm. We prune all the patterns which contains other than sister nodes as stated Equation 1. From the pruned set, we extract patterns which satisfy both minCS and maxOR property which are the Coverage Patterns of length 2 (TL_2). In the next step, non-overlapping patterns (NO_2) are generated by sorting them in order of CS and removing any CPs which don't satisfy maxOR criteria. Note that the pruning step is only required at for $k = 2$ as once the patterns containing any non-sister nodes are removed, there will be no non-overlapping patterns that can be generated that contain non-sister nodes in a CP. From $k = 3$, for k^{th} iteration of the algorithm, first candidate CPs, C_k are generated by joining NO_{k-1} patterns. From C_k, any patterns which do not satisfy the minCS or maxOR are not considered to generate CPs of length k, TL_k. From C_k, patterns which do not satisfy the maxOR or contain non-sisters nodes are removed and the remaining are sorted according to coverage support to generate non-overlapping sets of items of length k, NO_k. It should be noted that OR follows a 'sorted' downward closure property [4], and hence, the item sets of candidate sets, C_k are sorted to obtain the corresponding non-overlapping sets NO_k. An example of the algorithm is also shown in Fig. 3.



Tid	1	2	3	4	5
Items	(a,e)	(a,b)	(b,c,d)	(c,e)	(a,d,e)
Ancs.	(X,Y,Z)	(X,Y)	(X,Y,Z)	(X,Z)	(X,Y,Z)

Transaction Set

minRF = 0.5
minCS = 0.5
maxOR = 0.7
Parameters

C1

i	RF
X	1.0
Y	0.8
Z	0.8
a	0.6
b	0.4
c	0.4
d	0.4
e	0.6

TL1

i	RF
X	1.0
Y	0.8
Z	0.8

NO1

i	RF
X	1.0
Y	0.8
Z	0.8
a	0.6
e	0.6
b	0.4
c	0.4
d	0.4

C2

i	CS	OR
X,Y	-	-
X,Z	-	-
X,a	-	-
X,e	-	-
X,b	-	-
X,c	-	-
X,d	-	-
Y,Z	**1.0**	**0.6**
Y,a	-	-
Y,b	-	-

i	CS	OR
Y,e	-	-
Y,b	-	-
Y,c	-	-
Y,d	-	-
Z,a	-	-
Z,e	-	-
Z,b	-	-
Z,c	-	-
Z,d	-	-
a,b	**0.8**	**0.25**

i	CS	OR
a,e	-	-
a,c	-	-
a,d	-	-
e,b	-	-
e,c	**0.8**	**0.25**
e,d	**0.8**	**0.25**
b,c	-	-
b,d	-	-
c,d	**0.6**	**0.33**

TL2

i	CS	OR
YZ	1.0	0.6
a,b	0.8	0.25
e,c	0.8	0.25
e,d	0.8	0.25

NO2

i	CS	OR
YZ	1.0	0.6
a,b	0.8	0.25
e,c	0.8	0.25
e,d	0.8	0.25
c,d	0.6	0.33

C3/TL3/NO3

i	CS	OR
e,c,d	0.8	0.5

Fig. 3: Example 2: Example of T-Cmine

5 Proposed Approach

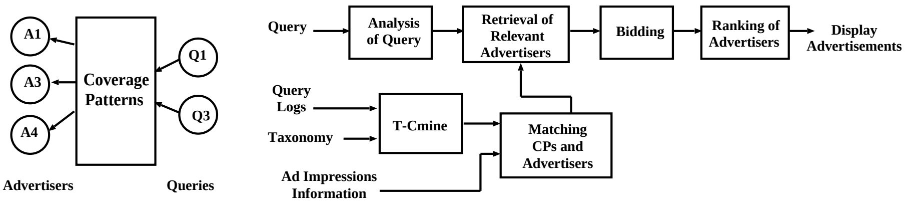
In this section, we discuss the proposed approach. In contrast to the sponsored search model of keyword based bidding, we proposed to add a middle layer of concepts during the bidding. Similarly, we also propose to add a middle layer to the allocation process such that when a user poses a query, it is first classified by a taxonomy into a set of nodes. For example, a query on *Harry Potter* would be classified into nodes *Shopping; Books; Fiction.* An advertiser who was allocated any of these concepts would be considered to be displayed on the query of *Harry Potter.* As compared to the standard sponsored search model of a bipartite graph between advertisers and queries as shown in Fig. 1 (a), we add a middle layer of CPs between search queries and advertisers as shown in Fig. 4 (a).

The sponsored search architecture has four major steps for query allocation to advertisers. The proposed architecture also has four major online steps for allocation of incoming queries to advertisers. But, in the proposed architecture, we also exploit the knowledge extracted from the query logs in the form of CPs. We discuss each step of the proposed architecture as follows.

1. **Query Analysis:** This step is same as the standard sponsored search architecture. But, we also extract the concepts of each incoming query. For

example, if the query is *Harry Potter* which belongs to the taxonomy *Shopping* then, it's concepts would be *Shopping; Books; Fiction.*

2. **Retrieval of Relevant Advertisers:** Based on the concepts inferred from the query in the first step, we retrieve advertisers from the matching of CPs. (In the next part of this subsection, we show how this matching of CPs and advertisers is achieved.)
3. **Bidding:** This step is same as the standard sponsored search architecture.
4. **Ranking of Advertisers:** This step is same as the standard sponsored search architecture.



(a) Proposed Allocation Model

(b) Proposed Allocation Architecture

Fig. 4: Proposed Sponsored Search Allocation Model and Architecture

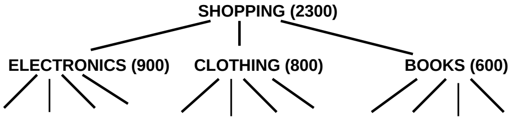
5.1 Matching CPs and Advertisers

In this section, we explain how the matching between CPs and advertisers is achieved. It should be noted that while considering this approach we assume the CPM (Cost Per Mille) payment mechanism, which can be easily extended to CPC (Cost Per Click) mechanism [3]. The matching process has two main components:

1. **Extraction of CPs using T-Cmine**: This step takes input of the query logs and the taxonomy and extract CPs as explained in Section 4.1.
2. **Matching CPs and Advertisers**: In this step, we take the demands of the advertisers and the CPs extracted from query logs and perform a matching between the two. An allocation protocol has been proposed such that specialized requests are processed before generalized. The reason for doing a specialized-to-generalized allocation is to acknowledge that an advertiser who bids on a lower level in the taxonomy has less options of allocation compared to the advertiser who bids on a higher level. For example, an advertiser who bids on the root node can be satisfied by any choice of children nodes. However, such an allocation poses a challenge where a coverage pattern containing a parent node has to be allocated given its descendants has been allocated to advertisers. Allocation at a node should take into account if any of its descendants have been allocated as coverage of a node is sum of coverage of its descendants. Hence, impressions of a node should be modified to take into consideration if any of its descendant nodes have been allocated to the advertisers. The necessary modification to a CP if any of its descendants have been allocated advertisers is to subtract the number of impressions allotted to the advertisers children of nodes contained in the respective CP.

Equation 2 captures the necessary changes required to a CP such for each node in the CP (denoted by k), count the impressions of allocated advertisers (denoted by j) of each descendant (denoted by i) and subtract it from total impressions of the CP. It should be noted that a coverage pattern is allocated to a set of advertisers if and only if it has enough impressions to satisfy the allocated advertisers. It may happen that advertisers are not allocated a coverage pattern if supply is greater than demand, and thus the following equation will never result in a negative value for the number of impressions of a coverage pattern.

$$CP.imp = CP.imp - \sum_{k} \sum_{ij} A_{ij} \tag{2}$$



Example Taxonomy

Ad ID	Node	Impressions
A1	Shopping	800
A2	Clothing	500
A3	Books	200
A4	Books	300
A5	Shopping	500

Table2: Example Impression Requests by Advertisers

Extracted CPs	Imp	Modified Imp
{Books, Clothing}	1400	400
{Electronics, Clothing}	1700	1200
{Books, Electronics}	1500	1000

Table1: Impressions provided by CPs beforeand after allocation at level 3

Ad Id	CPs
A1	{Books, Electronics}
A5	{Electronics, Clothing}

Table 3: Allocated CPs to Advertisers at Shopping

Fig. 5: Example Allocation

Example 3: In Fig. 5, we show an example allocation. We consider the top two levels of a taxonomy to show and consider advertisers who bid on the first three levels. Each advertiser bids on a node and has a demand of certain impressions at that node. Assuming allocation was done at level two i.e. for *Electronics, Clothing* and *Books*, we will show how it will be done for *Shopping.* The node *Shopping* has three children and CPs pertaining to *Shopping* are shown in Table 1 of Fig 5. However, as we know that allocations have been done for advertisers who chose to bid upon *Books* and *Clothing*, we need to adjust the impressions provided by the CPs containing these two nodes. For example, the CP {*Book, Clothing*} has 1400 initial impressions, but some advertisers were already allocated *Books* and *Clothing* during allocations at lower level(s). Hence, those impressions need to be subtracted i.e. 1400 - (500 + 200 + 300) = 400. Similarly, for {*Electronics, Clothing*}, the modified number of impressions is 1200 i.e. 1700 - 500 and that of {*Books, Electronics*} is 100 i.e. 1500 - (200 + 300) = 1000. In the next part of this section, the matching between CPs is performed considering the proposed modification.

A matching is performed with advertisers as one side of the bipartite and CPs as the other side. The matching is done at each node of the taxonomy where more than one advertisers choose to bid. In order to maximize the revenue, the matching should be performed in such a way that maximum number of impressions that can be provided by the coverage patterns should be allocated. We propose the matching as an optimization problem in the same respect such that the difference between the CPs and advertisers allocated to them should be minimal. For example, if an advertiser demands 100 impressions and there are two CPs with impressions 150 and 200 respectively, then we chose to allocate the CP with 150 impressions. A similar case can be made when the supply of CPs is 50 and 75 impressions and demand by the advertisers is 100 impressions, then the CP with 75 impressions is chosen. We frame the objective function of the matching on the same notion which is as follows. Equation 3a aims at minimizing the difference between the allocated advertising and the CPs. The objective function is such that for each advertiser Ad_{ij} who has been allocated the CP, CP_j the difference between the two is minimal. Equation 3b lays out the constraint such that the sum of impressions of allocated advertisers does not exceed the impression provided by the CP to avoid the objective from going negative.

$$Min\ Z = \sum_{level=d}^{0} (\sum_{j} |CP_j.Impressions - \sum_{i}^{n} (Ad_{ij}.Impressions)|) \tag{3a}$$

$$s.t\ \ CP_j.Impressions >= \sum_{i=1}^{n} (Ad_{ij}.Impressions) \tag{3b}$$

Continuing Example 3 from Fig 5. From the last step, we have CPs whose impressions have been updated according to allocations at their descendants. We show how the allocation is to be done for the node *Shopping.* Two advertisers A_1 and A_5 chose to bid on the node *Shopping.* In the proposed approach, we decide to serve the advertisers on a first-come-first-serve basis. For ad A_1, we select the CP {*Books, Electronics*} because it has the lesser difference compared to the other node. It should be noted that now the number of impressions covered by CP {*Books, Electronics*} has been reduced to 200 as A_1 has been allotted to it. Next, we look at ad A_5 and we see that out of the three CPs, only {*Electronics, Clothing*} has enough impressions to satisfy the advertiser and after this allocation, the number of impressions covered by {*Electronics, Clothing*} reduces to 500. Through the example, we wanted to demonstrate how the proposed specialized-to-generalized allocation would work for advertisers who bid on *Shopping* considering a set of advertisers bid on children of *Shopping* and hence, the results for only A_1 and A_5 are shown. It should be noted that the matching between CPs and advertisers will be one-to-many as the number of impressions that can be covered by a CP is much large compared to demands of a single advertiser.

Considering the allocation done for Example 3, let us say a query related to the taxonomy is fired say, *Harry Potter.* As shown in Fig. 4, it will be first classified according to the taxonomy as *Shopping; Books; Fiction.* Advertisers who have been allotted a CP containing any of these nodes are considered for being displayed on this query's results page i.e. A_1, A_3, A_4 and A_5 would be considered to be displayed. The decision on who out these four would be shown and in which order will be decided by the ranking mechanism which includes their bids, remaining budget etc. (As stated earlier, ranking and bidding are independent of the proposed approach.)

6 Experiments

6.1 Dataset

For the experiments, we used the CABS120k08 [10] dataset which is a collection of search queries from the AOL500k dataset along with the documents clicked, document rank, timestamps and user id. The dataset models the web document as a unit. The data set also contains the classification of the clicked document according to a concept taxonomy of four levels. From the dataset, we extracted all the queries in the form: $< query,\ user-id,\ timestamp,\ concept\ taxonomy >$. Concept taxonomy present in the data is a four level taxonomy including the root node. Without loss of generality, we assumed that the search queries related to the documents also have the same category as the web document. The case where the same document had multiple categories, the first one was arbitrarily selected. After extracting queries, we extracted sessions of four most popular taxonomies – *Arts, Health, Society* and *Shopping* from the dataset that had more than a single query with at least two sub-concepts of the same concept in the same session. Each session is used a transaction to extract coverage patterns by T-Cmine as sessions form the logical boundary of searching. Table 2 shows the statistics of the extracted dataset.

Table 2: Search Query Dataset Statistics

Taxonomy	Number of Nodes	Sessions	Queries
Arts	48	7,107	15,317
Health	59	9,181	26,385
Society	68	6,471	13,223
Shopping	79	14,819	40,463
Total	254	37,578	95,388

6.2 Implementation Methodology

The standard sponsored search approach mentioned in [9] is compared with the concept based bidding approach. We simulate advertising demands randomly in terms of impressions for five sets of advertisers having 10, 20, 30, 40 and 50 advertisers. For the standard keyword bidding, a keyword is selected as the seed for each advertiser such that the probability of selection of a keyword as the seed is proportional to its frequency in the dataset, in order to mimic the advertising demand. Followed by selection of a seed keyword, all keywords from the dataset are selected to be in the advertiser's campaign for which the Wu-Palmer similarity is more than 0.8. The number of requested impressions is randomly chosen between 100 and 1000. To simulate bid for each keyword, we consider the minimum bid as \$1.00, the maximum bid as \$ 10.00 and the actual bid for each keyword is considered as the function of its relative frequency between the minimum and maximum value. For the experimental setup, we assume the bid to be paid per hundred impressions instead of per 1000 impressions as in CPM model to analyse more number of requests. The bid amount here indicated how much the advertiser is willing to pay for 100 impressions. For the concept based bidding approach, bid of an advertiser on the concept is average of bids on all the keywords in his/her campaign.

6.2.1 Performance Metrics: Two performance metrics have been employed to compare the keywords based approach [9] and the proposed concept-based bidding approach.

To evaluate the utilization of ad space, we calculate the average number of unique Advertisements per Session (AS). It is calculated as the ratio of *Sum of Unique Advertisements of all Sessions (SUAS)* and *Number of Sessions with Advertisements (NSA)*. High value of AS indicates more utilization of a session, which in turn indicates covering of more advertisers.

$$AS = \frac{SUAS}{NSA} \tag{4}$$

We also measure the reach of each advertisement. Reach is defined as the number of users that view the ad. In this experiment, we consider reach of the ad with respect to the sessions instead of users as sessions define a logical boundary of tasks in search engines. To measure the reach, the value of Sessions per Advertisement (SA) is calculated which is the ratio of *Number of Unique Sessions for each Ad (NUSA)* to *Number of Advertisements (NA)*. A higher value of the metric implies the more number of unique eye balls and thus, increasing the chances of the advertisement being viewed by *diverse* users.

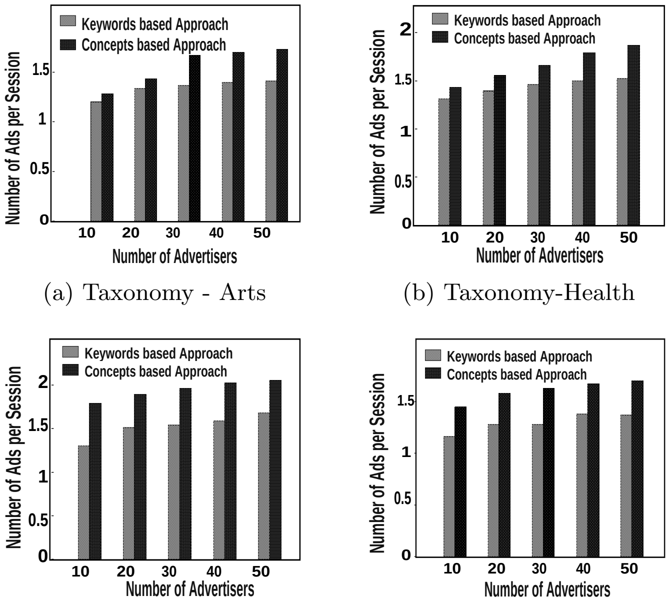
$$SA = \frac{NAS}{NA} \tag{5}$$

6.2.2 Results : Fig. 6 reports the results with respect to ad space utilization. A fair improvement is observed in concept based bidding mechanism. Average improvement is 19.81% across all four taxonomies and all sets of advertisers. For individual taxonomies, average improvement for *Arts* is 18.33%, *Health* is 13.74%, *Society* is 17.29% *Shopping* is 29.86%. The improvement for *Shopping* show the highest improvement by a significant margin compared to the other three taxonomies. This is because for *Shopping* taxonomy average length of a session as well as distribution of nodes was higher compared to the other three taxonomies. Hence, it was possible to extract more interesting coverage patterns in the category of *Shopping*. These results align in the same way for the next performance metric as well.

Fig. 7 shows the performance of two approaches with respect to reach of advertisements. An average improvement of 18% was observed. For individual taxonomies, improvement for *Arts* is 13.41%, *Health* is 14.83%, *Society* is 16.05% *Shopping* is 27.70%. The results for *Shopping* show significant improvements again because of the same reason as stated above.

7 Conclusions and Future Work

In this paper, we address the issue of advertising on long tail search queries in search engines. We propose that advertisers should bid upon high level concepts represented by a taxonomy instead of search keywords during ad space auctions. To address the issues of inter-dependency of concepts on each other, we exploit



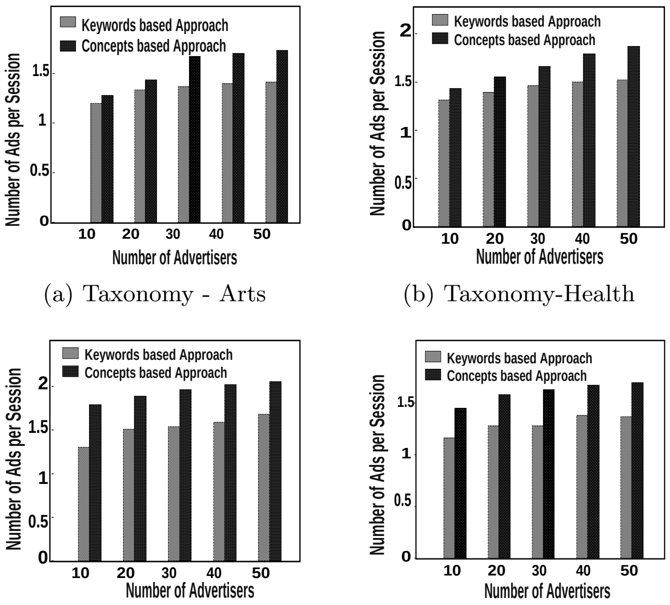
(a) Taxonomy - Arts

(b) Taxonomy-Health

(c) Taxonomy-Shopping

(d) Taxonomy-Society

Fig. 6: Performance with respect to utilization of ad space



(a) Taxonomy - Arts

(b) Taxonomy-Health

(c) Taxonomy-Shopping

(d) Taxonomy-Society

Fig. 7: Performance with respect to reach of advertisers

search query logs and a taxonomy to extract level-wise coverage patterns. The corresponding architecture is used to perform allocation of incoming queries to advertisers for sponsored search. Experiments on a real world dataset of AOL search query logs show improvement in performance with respect to ad space utilization and reach of the advertisements.

As a part of future work, we plan to analyse what is the trade-off between relevance and bidding on concepts in terms of targeted advertising. Also, in this paper, we assumed that a taxonomy exists over search query logs. We plan to investigate how different taxonomies would suit the problem and if it is possible to build a taxonomy to suit sponsored search so to avoid the long tail phenomenon amongst the nodes of the taxonomy. We also intend to look at truthful auctions for concept-based bidding as the advertisers are targeting same keywords but using different concepts.

Bibliography

[1] Srinivas, P. G., Reddy, P. K., Bhargav, S., Kiran, R. U., & Kumar, D. S. (2012). Discovering coverage patterns for banner advertisement placement. In Pacific-Asia Conference on Knowledge Discovery and Data Mining.

[2] Srikant, R., & Agrawal, R. (1997). Mining generalized association rules. Future Generation Computer Systems.

[3] Budhiraja, A., & Reddy, P. K. (2015). An approach to cover more advertisers in adwords. In Data Science and Advanced Analytics.

[4] Srinivas, P. G., Reddy, P. K., Trinath, A. V., Bhargav, S., & Kiran, R. U. (2015). Mining coverage patterns from transactional databases. Journal of Intelligent Information Systems.

[5] Broder, A., Ciccolo, P., Gabrilovich, E., Josifovski, V., Metzler, D., Riedel, L., & Yuan, J. (2009). Online expansion of rare queries for sponsored search. In International Conference on World Wide Web.

[6] Broder, A. Z., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V., & Zhang, T. (2007). Robust classification of rare queries using web knowledge. In International ACM SIGIR conference on Research and development in Information Retrieval.

[7] Broder, A. Z., Ciccolo, P., Fontoura, M., Gabrilovich, E., Josifovski, V., & Riedel, L. (2008). Search advertising using web relevance feedback. In Proceedings of the 17th ACM conference on Information and Knowledge management.

[8] Skiera, B., Eckert, J., & Hinz, O. (2010). An analysis of the importance of the long tail in search engine marketing. In Electronic Commerce Research and Applications.

[9] Mehta, Aranyak, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. (2007). Adwords and generalized online matching. Journal of the ACM (JACM).

[10] Noll, M. G., & Meinel, C. (2008). The metadata triumvirate: Social annotations, anchor texts and search queries. In Web Intelligence and Intelligent Agent Technology.

[11] Kavya, V. N. S., & Reddy, P. K. (2016). Coverage Patterns-Based Approach to Allocate Advertisement Slots for Display Advertising. In International Conference on Web Engineering.

[12] Mehta, A. (2012). Online matching and ad allocation. Theoretical Computer Science.