Exploiting ad space of long tail queries through concept-based bidding in sponsored search

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science (by Research)
in
Computer Science and Engineering

by

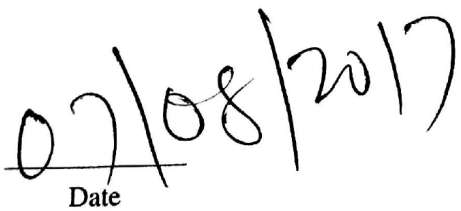
Amar Budhiraja
201303009
amar.budhiraja@research.iiit.ac.in
amar.budhiraja1@gmail.com

International Institute of Information Technology
Hyderabad - 500 032, INDIA
July 2017

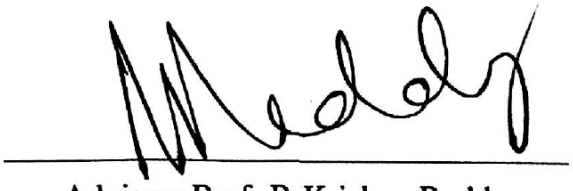
International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Exploiting ad space of long tail queries through concept-based bidding in sponsored search" by Amar Budhiraja, has been carried out under my supervision and is not submitted elsewhere for a degree.



Date



Advisor: Prof. P. Krishna Reddy
Data Science and Analytics Center
Kohli Center on Intelligent Systems
IIIT, Hyderabad

To my family

Acknowledgments

I would like to express my sincere gratitude to my advisor Prof. P. Krishna Reddy for his continuous guidance, patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. The door to Prof. Reddy's office was always open whenever I ran into a troubled spot or had a question about my research or writing. He consistently allowed this thesis to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to thank my colleagues Raghav Kalyanasundaram, Sai Kavya Vaddadi, Mamatha Alugubelly, Amulya Kotni and Lakshmi Gangumalla for their constant support and for providing a positive learning environment. I am specially grateful to Sai Kavya Vaddadi for the brain storming sessions during my peak research months.

I also take this opportunity to thank Google Research India for providing travel grants to attend International Conference in Data Science and Advanced Analytics, 2015 at Paris, France. Furthermore, I would like to thank Microsoft Research India for supporting my travel to International Conference in Database Systems for Advanced Applications, 2017 at Suzhou, China for presenting my work at the conference.

Last but not the least, I express my deepest gratitude toward my parents, Sardar Harbans Singh and Bibi Paramjeet Kaur and my elder brother, Maninder Singh, for their unconditional support in all my endeavours.

Abstract

Sponsored search is one of the most dominant modes of online advertising on the web. In sponsored search, advertisers bid on relevant keywords to advertise 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 determined by 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. The ecosystem of sponsored search has three main stakeholders - search engine, advertisers and users. Search engines aim to maximize the revenue by showing ads. Advertisers want to maximize the reach of their service or increase the sales. Users see advertisements on the search results page. Some of the key research challenges in sponsored search are query-ad relevance, click-through-rate prediction, optimal auction design, optimum utilization of ad space of rare (tail) queries and click-fraud detection.

In this thesis, related to sponsored search, we have investigated approaches to exploit the ad space of tail queries. It is well established that search queries tend to follow a heavy-tailed Zipf distribution wherein a large fraction of queries occur too infrequently. Such infrequent query set is called *long tail*. Advertising on long tail queries is challenging as long tail queries are encountered rarely which makes them hard 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 or no demand for the tail query keywords. The long tail phenomenon also makes it quite difficult for an advertiser 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.

We propose two approaches for the utilization of ad space of tail queries by proposing that instead of bidding on keywords, advertisers should bid upon high level concepts. In the first approach, we propose bidding on concepts by organizing concepts into a two level taxonomy. In the second approach, we propose a generalized approach by considering organization of concepts into a multi-level taxonomy.

In the first approach, we have proposed an improved framework to cover more advertisers by exploiting the notions of coverage and concept taxonomy based on the log data of search queries. The proposed framework allows to form the distinct groups of keywords such that each group of keywords could be allocated to meet the demands of the advertiser. We model each search session as a transaction

and queries occurring in a session form the items of the transaction. Coverage patterns are then mined from these session-based transactions. The coverage patterns fuse tail keywords together into multiple groups to maximize the unique visitors and two-level concept taxonomy ensures that groups are meaningful. A comprehensive framework has been proposed to map the extracted coverage patterns and the demands of the advertisers to allocate incoming queries to advertisers. We have conducted experiments on a real world dataset of AOL web search engine and found out that it is possible to meet the demands of more advertisers with the proposed approach. It was found out that the proposed approach is able to provide diverse but meaningful group of keywords which allows the advertiser to display advertisement to appropriate users based on the requirements.

In the second approach, we propose that advertisers should bid upon high level concepts represented by a multi-level taxonomy instead of search keywords during ad space auctions. Advertisers are free to bid on any node of the taxonomy. Bidding on any node in the taxonomy provides more flexibility to the advertisers to target the potential consumers. However, to allocate children nodes of bidding nodes to advertisers, the flat model of coverage patterns cannot be used. To address the issues of inter-dependency of concepts on each other, we exploit search query logs and a taxonomy to extract level-wise coverage patterns. We further propose an end-to-end architecture which takes search query logs, taxonomy and advertising demands as inputs and allocates an incoming search query to the advertisers. 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.

Overall, we have proposed two approaches for improving the utilization of ad space of long tail queries. Both approaches organize concepts into taxonomy and exploit the knowledge of coverage patterns extracted from search session log data. We have proposed an improved approach, which is easy to adopt, by organizing concepts into two-level taxonomy and a generalized approach by organizing the concepts into a multi-level taxonomy. Based on the results on real world data, we conclude that there is an opportunity to improve ad space utilization of long tail queries in sponsored search.

Contents

Chapter Page

1 Introduction . . . 1
1.1 Introduction to Online Advertising . . . 1
1.1.1 History of Online Advertising . . . 1
1.1.2 Types of Online Advertising . . . 2
1.1.2.1 Display Advertising . . . 2
1.1.2.2 Social Media Advertising . . . 3
1.1.2.3 Sponsored Search . . . 3
1.1.2.4 Email Advertising . . . 4
1.1.2.5 Online Classified Ads . . . 4
1.2 Sponsored Search . . . 4
1.3 Research Issues in Sponsored Search . . . 5
1.4 Research Gap . . . 6
1.5 Overview of Proposed Approaches . . . 7
1.6 Contributions of Thesis . . . 8
1.7 Organization of Thesis . . . 9

2 Related Work . . . 10
2.1 Sponsored Search . . . 10
2.2 Long tail advertising in sponsored search . . . 11
2.2.1 Query Expansion . . . 11
2.2.2 Query Reformulation . . . 12
2.3 Coverage Patterns . . . 13
2.4 How Proposed Approaches Are Different . . . 14

3 Background: Sponsored Search and Coverage Patterns . . . 15
3.1 Sponsored Search Framework . . . 15
3.2 Overview of Coverage Patterns Model . . . 16
3.2.1 Overview of approaches to mine Coverage Patterns . . . 18
3.2.2 CMine Algorithm with Example . . . 19
3.3 Summary . . . 20

4 Exploiting ad space of tail queries using a two level taxonomy and coverage patterns . . . 21
4.1 Motivation and Basic Idea . . . 21
4.2 Proposed Model . . . 23
4.3 Proposed Framework . . . 23

4.3.1 Allocation of concepts to advertisers . . . 24
4.3.1.1 Conversion of Query Logs to Taxonomic Transactions . . . 25
4.3.1.2 Extraction of Coverage Patterns . . . 25
4.3.1.3 Estimation of Number of Impressions of Advertisers . . . 28
4.3.1.4 Matching of Coverage Patterns and Advertisers . . . 29
4.4 Experiments . . . 34
4.4.1 Dataset . . . 34
4.4.2 Implementation Methodology . . . 35
4.4.2.1 Sponsored Search system . . . 35
4.4.2.2 Proposed Approach . . . 36
4.4.3 Performance Metrics . . . 37
4.4.4 Results . . . 37
4.5 Discussion: Assumptions and Limitations . . . 38
4.6 Summary . . . 40

5 Exploiting ad space of tail queries using a multi-level taxonomy and coverage patterns . . . 41
5.1 Motivation and Basic Idea . . . 41
5.2 Proposed Model . . . 42
5.2.1 T-Cmine: An approach to extract 'level-wise' coverage patterns . . . 43
5.3 Proposed framework . . . 44
5.3.1 Allocation of concepts to advertisers . . . 46
5.4 Experiments . . . 49
5.4.1 Dataset . . . 49
5.4.2 Implementation Methodology . . . 50
5.4.3 Performance Metrics: . . . 50
5.4.4 Results . . . 51
5.5 Discussion: Assumptions and Limitations . . . 52
5.6 Comparison of the proposed approaches . . . 53
5.7 Summary . . . 54

6 Conclusion and Future Work . . . 56

7 Publications . . . 58

List of Figures

Figure		Page
1.1	Types of Online Advertising based on mechanims of delivery	3
1.2	Example of query's results page from Google showing sponsored search results	4
1.3	Long tail distribution of AOL search query logs with a short head of frequent keywords and long tail of infrequent keywords is shown in (a). Log-Log distribution of AOL search query logs in shown in (b). Log-Log distribution of a perfect long tail should align to a straight line.	7
3.1	Sponsored Search Model	15
3.2	Sponsored Search Architecture as discussed in [56]	16
3.3	Working of CMine algorithm. The term 'I' is an acronym for the item set (or web pages).	19
4.1	Sponsored search bidding: keywords based bidding and concept-based bidding	22
4.2	Proposed model	23
4.3	Example of a concept taxonomy	23
4.4	Proposed Architecture	24
4.5	Allocation of concepts to advertisers	25
4.6	Algorithm to Match Coverage Patterns and Advertisements	33
4.7	Performance with Respect to Coverage of Advertisers	38
4.8	Performance with respect to Diversity	39
5.1	Example 2: Example of T-Cmine	45
5.2	Proposed Sponsored Search Model and Architecture	45
5.3	Allocation of concepts to advertisers	46
5.4	Example Allocation	47
5.5	Performance with respect to utilization of ad space	51
5.6	Performance with respect to reach of advertisers	52

List of Tables

Table		Page
3.1	Transactional database, D	17
4.1	Sample Sessions	26
4.2	Taxonomic Transactions	26
4.3	Relative Frequencies of Sub-Concepts	28
4.4	Extracted Coverage Patterns	28
4.5	Table Showing Advertisements Details	29
4.6	Coverage Pattern and Advertiser Matching	32
4.7	Table Showing Allotment Details	32
4.8	Dataset Statistics	35
4.9	Advertisers Dataset Statistics	36
5.1	Search Query Dataset Statistics	50
5.2	Comparison of the proposed approaches	54

Chapter 1

Introduction

1.1 Introduction to Online Advertising

In online advertising, advertisers distribute promotional marketing messages to potential consumers through the Internet. Already a multi-billion dollar industry, online advertising is still showing rapid growth. This growth in the online advertising industry is attributed to multiple factors including considerably cheap and more targeted advertising compared to its offline counterparts, world wide coverage of the Internet, the velocity of ad delivery and quick feedback mechanisms.

1.1.1 History of Online Advertising

The notion of online advertising started in 1994 with the first clickable ad by Wired magazine's online counterpart, HotWired. HotWired devised a plan to set aside portions of its website to sell space to advertisers, similar to how ad space is sold in a print magazine. That is how banner advertising was born. Banner advertising quickly grew as a prominent method for websites to keep their content ungated and free for users. In 1995, ad servers were introduced and the concept of targeting relevant websites was started. Multiple services started operating which would help advertisers select websites where they could potentially find the desired consumer demographics. Around 1998, ad networks were formed and user targeting started becoming more prominent.

With search engines steadily gaining popularity in late 90s, following years also saw the rise of search engine advertising. Advertisers who were looking to create ads that were more targeted turned to sponsored search. Initially Paid Placement Model (PPM) was offered for search engine advertising where an advertiser would directly bid upon the keywords and based solely on the bid, the ad slot would be assigned. Google modified PPM by emphasizing more on user experience and hence, an ad's position on the query results page was determined by not only the bid of the advertiser but also by the relevance of the ad to the query and other features like Click-Through-Rate.

The next break-through came into display advertising. Around 2007, demand side and supply side platforms were introduced where the idea of Real Time Bidding (RTB) was put forward. RTB allows

advertisers to bid in real time for incoming ad slots. The bidding is usually done automatically but can also be done manually. RTB allowed advertisers to be more specific with respect to the targeted potential consumers and at the same time, it allowed publishers to sell most of the ad space as one impression is sold at a time.

Meanwhile, the social media ecosystem started developing around early 2000s with the advent of LinkedIn and Facebook and social media advertising started around 2005. Facebook allowed user level targeting and was a unique player in the market to accurately target demographics. Social media advertising later advanced to sponsored posts. Very recently, Facebook also started mobile ads on its mobile applications in order to target consumers in real time based on the location and other hyper-demographics.

Aside from the Internet explosion, the mobile era was steadily developing since early 90s. Around 2000, the Mobile Marketing Association was founded for the development of mobile ads and the first SMS ad was served in the same year by a Finnish news provider by offering free news headlines sponsored by advertising. This initiative also led to other experiments in mobile marketing and with the advent of smart phones in late 2000s, mobile advertising started growing at a very accelerated rate. Now, mobile advertising includes all forms of web advertising including sponsored search and banner ads.

The current status of online advertising is targeting 'potential' consumers who would make a transaction and not just click the ad. In order to achieve this, almost all players in the market are trying to achieve behavioral targeting at a very granular level by employing multiple data sources like chat logs, social media activities, local information, etc. A key role in behavioral targeting is played by mobile phones as they are a real time source of user activity. In fact, in the latest report by Internet Advertising Bureau (IAB), they state that mobile advertising has the highest growth in revenue in the past decade [41].

1.1.2 Types of Online Advertising

Based mechanism of delivery, online advertising is classified into the following categories - *Display Advertising*, *Social Media Advertising*, *Sponsored Search*, *Email advertising* and *Online Classified Ads*.

1.1.2.1 Display Advertising

Display advertising is the graphical information that appears next to content on web pages. This can include rich content beyond text such as logos, pictures and even videos. In display advertising, advertisers can either opt for guaranteed contracts (i.e. the publisher will show the number of ads as determined by the contract for a fixed price) or real time bidding (i.e. where each impression is sold individually via auctions through ad servers). Display advertising can be further classified into pop-up ads and banner ads. Pop-up are generally same as banner ads except they pop-up on the screen suddenly to grab attention.

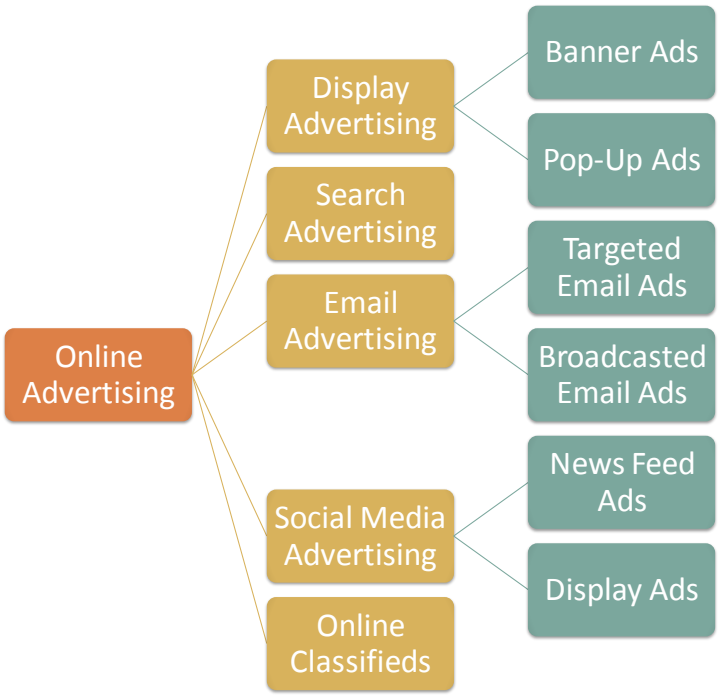


Figure 1.1: Types of Online Advertising based on mechanims of delivery

1.1.2.2 Social Media Advertising

Social media advertising has been defined as "any piece of online content designed with a persuasive intent and distributed through social media which enables the users to access, share and engage with it" [7]. Social media advertising can also be further classified into two types: news feed ads and display ads. News feed ads are the ads that are incorporated along with organic content in the news feed of the users. These are also popularly known as 'sponsored posts'. Display ads in social media are the ones which are shown outside the news feed, such as on the right side of Facebook.

1.1.2.3 Sponsored Search

Sponsored search refers to advertising on search engines. When a user poses a query to the search engine, ads are shown along with the search results. Search advertising is referred to as sponsored search because advertisers have to pay to be displayed on the search results page whereas organic content is freely displayed. Figure 1.2 shows a snapshot of search results containing ads and organic content.

In the next section, sponsored search will be discussed in more detail as this thesis contributes to long tail advertising in sponsored search.

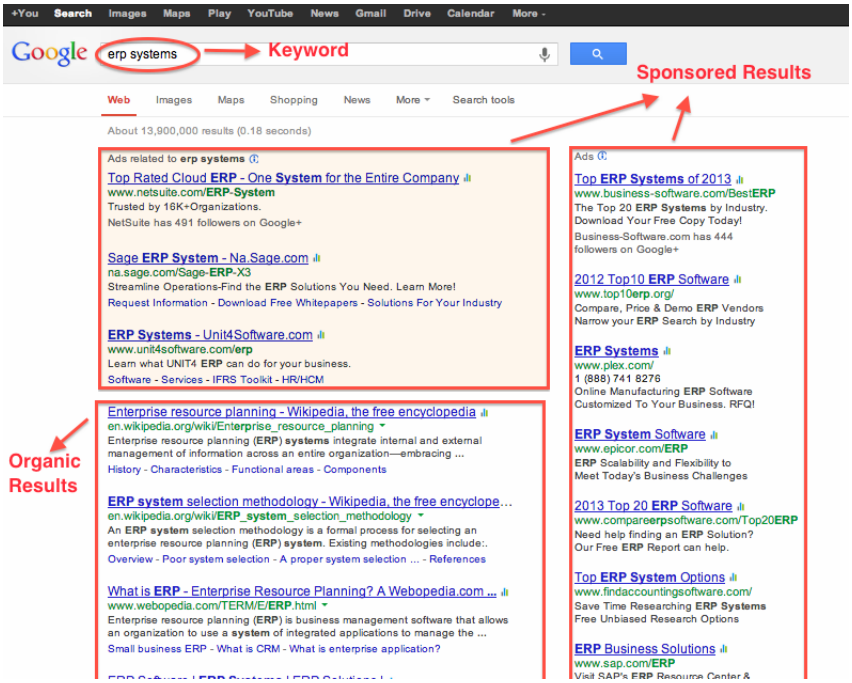


Figure 1.2: Example of query's results page from Google showing sponsored search results

1.1.2.4 Email Advertising

Email advertising refers to sending a commercial message to one or more potential customers. Email advertising generally involves catchy subjects and rich media to attract costumers. Generally, the goal of email advertising is to build loyalty, trust, or brand awareness. Email advertising can also be further classified into targeted email advertising and broadcast email advertising.

In targeted email advertising, promotions are sent to only specific a set of people who match the advertiser's criteria whereas in broadcast email marketing, emails are sent to everyone whose address is present in the email database.

1.1.2.5 Online Classified Ads

Online classified ads are ads which are posted online in categorical listings. Examples include online yellow pages and job portals. Quickr and Craigslist are some of the most prominent online classified ad providers.

1.2 Sponsored Search

The model of search engine advertising is more popularly known as sponsored search. When a user poses a query to the search engine, ads are shown along with organic search results. Figure 1.2 shows a snapshot of sponsored search from Google.

In sponsored search, advertisers bid upon search keywords which they deem relevant to their product. When a user queries the search engine, all the advertisers who chose to bid upon the keywords contained in query are considered as candidate advertisers to be shown on the query's results page. The candidate advertisers are then ranked according to the bid amount and other parameters, to be displayed on the query's results page.

In the sponsored search ecosystem, there are three major stakeholders - *Search Engine Users*, *Advertisers* and *Search Engine itself* (as the publisher of the ads). The goals of the stakeholders are defined as follows.

(i). **Search Engine Users:** Users visit search engines to meet their information requirements.

(ii). **Advertisers:** Advertisers aim to promote their product through ads.

(iii). **Search Engine:** Search engines wish to earn revenue through sponsored search.

Sponsored search plays a very important role in the sustainability of the search engine as the search engine model of knowledge sharing is free of cost. Hence, most of the major search engines in the market aim at improving revenue from sponsored search.

1.3 Research Issues in Sponsored Search

Following are some of the key research issues in sponsored search.

(i). **Ad-query relevance:** In sponsored search, it is of utmost importance that the displayed ads are related to the search query. If an ad is irrelevant to the search query, it will affect the user experience and the search engine may end up loosing the user. In the literature, attempts have been made to use traditional IR methods [61] and machine learning models [37] to evaluate if an ad is relevant to a given query.

(ii). **CTR prediction:** Click-Through-Rate (CTR) measures how many times an ad has been clicked compared to the number of times it was shown. Accurately predicting CTR is important as the search engine is only paid by the advertiser when an ad is clicked, and not when the ad is displayed. Hence, in order to make earn more revenue, high CTR valued ads are ranked higher than low CTR valued ads. Machine learning methods have been extensively discussed in the literature for predicting CTR [35, 63].

(iii). **Optimal auction design:** A search engine faces revenue maximization problem when it designs its auction mechanism. Typically, each query can be matched to multiple advertisers and each advertiser can be matched to multiple queries. Therefore, search engines aim at designing auctions to maximize this matching between ad slots and advertisers. Optimal auction design has often been discussed in literature by modeling ads and queries as an online bi-partite graph where the ad set is known but search queries come online [55, 56].

(iv). **Click frauds:** Click-fraud involves taking on the role of a consumer either directly (personally) or indirectly (software-supported). In the case of click-fraud, the goal of the attacker is an increase of the advertisers' costs by artificially increasing the number of clicks per ad. Click-fraud is frequently used with the intention to significantly downgrade a competitor's ranking and improving one's own search engine ranking [57]. It is very important to identify click frauds as it could lead to advertiser dissatisfaction and the search engine could potentially loose a revenue source.

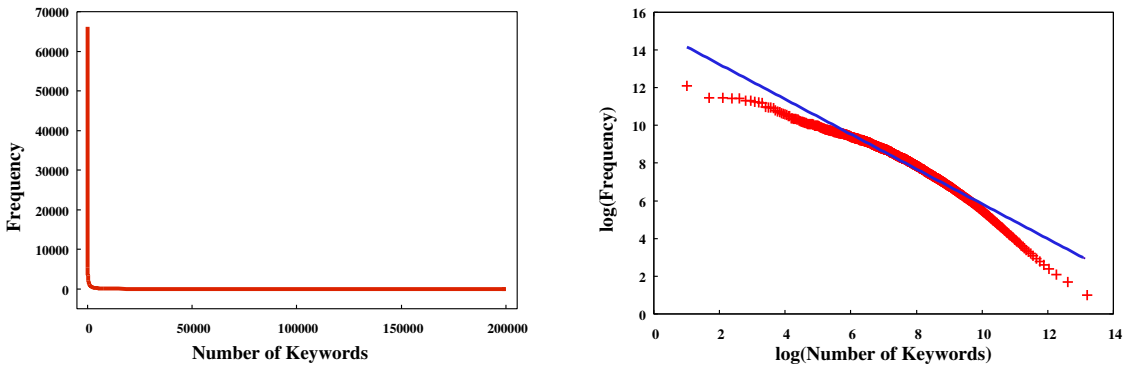
(v). **Bid phrase suggestions**: Search engines aim at easing the process of keyword auctions for the advertisers by providing bid phrase suggestions. Bid phrase suggestions help advertisers in selecting relevant keywords quickly without being a domain expert. Bid phrase suggestions have been achieved in the literature by using landing page of the ad [62], random forest classification of previously seen bid phrases [6], semantic word graphs [2] and Wikipedia [42].

(vi). **Utilization of ad space of rare queries**: Search queries follow a long tail distribution where a large fraction of search queries occur very infrequently. The difficulty of rare (infrequent) queries is largely caused by the inherent data sparsity problem due less number of organic and sponsored clicks. Allocating ads to rare queries is a well identified research issue and has been mostly addressed by the means of query expansion [11] or query reformulation [66]. This research issue is discussed in more detail in the next section.

1.4 Research Gap

It is well established that volume distribution of search queries follows the power law [11]. The *head* and *torso* of the distribution consists of the most frequent queries while the *tail* part is composed of the rare queries. Although, individually rare, tail queries form a significant fraction of the search volume which makes tail queries important for the advertising revenue [11]. Figure 1.3 (a) shows the long tail distribution for search queries of the AOL query dataset and Figure 1.3 (b) shows the log-log distribution of AOL search query logs. (Log-log distribution of a perfect long tail should align to a straight line.)

It has also been noted that head keywords are more desirable to advertisers because of their individually high volume in the search distribution [64]. It was further stated that during keyword auctions, the desirability for head keywords leads to a very high competition for the head query keywords while very less competition for the tail query keywords [64]. The long tail phenomenon also makes it quite difficult for an advertiser 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.



(a) Long tail distribution of AOL search query logs

(b) Log-Log distribution of AOL search query logs

Figure 1.3: Long tail distribution of AOL search query logs with a short head of frequent keywords and long tail of infrequent keywords is shown in (a). Log-Log distribution of AOL search query logs in shown in (b). Log-Log distribution of a perfect long tail should align to a straight line.

1.5 Overview of Proposed Approaches

In this thesis, we have studied the issue of under-utilization of ad space of tail queries in sponsored search. We propose that advertisers should bid upon high levels concepts instead of keywords in the ad space auctions. In the first approach, we consider that the concepts are organized in a two level taxonomy. In the second approach, we propose a generalized approach by considering organization of concepts in a multi-level taxonomy.

In the first approach, we have investigated bidding on concepts using a two level taxonomy. Advertisers can bid only on the first level and sets of children nodes of the bid nodes are allocated to the advertisers. The notion of coverage patterns has been employed to form these groups of children nodes. In the literature, a coverage pattern has been defined as a set of items which *covers* a minimum fraction of transactions and has a maximal *overlap* of transactions within the items in the pattern. To extract coverage patterns, we utilize the notion of search sessions by considering the queries typed in the same session as a transaction. We further transform each transaction by replacing each query with the first and second level nodes as per the bidding taxonomy. Thus, each transaction is translated into taxonomy nodes. Coverage patterns are extracted from these taxonomy-based search transactions. The extracted coverage patterns help in identifying mutually exclusive sets of taxonomy nodes which ensure certain coverage along with a maximum overlap of transactions, thereby reducing the repetition of nodes in the same transaction. An approximate matching is performed between the extracted coverage patterns and advertising demands. Using the matching, we have defined an end-to-end architecture to allocate ads to an incoming query. Experiments on the real world dataset of AOL search query logs show that the proposed approach can help in exploiting ad space of long tail queries.

In the second approach, we generalize bidding in ad space auctions from a two level taxonomy to a multi-level taxonomy. Advertisers are free to bid on any node in the taxonomy. Bidding on any node

in the taxonomy provides more flexibility to the advertisers to target potential consumers. During ad campaign creation, an advertiser is shown a taxonomy based on his/her ad content and is asked to select a node which he/she deems to be most relevant to the product being advertised. A set of children nodes of the bid node is allocated to the advertisers, similar to the first approach. However, each node in the taxonomy is composed of its descendants. For example, if we consider a node *Shopping*, it will be composed of children nodes like *Books, Fashion* and *Electronics*. Thus, each transaction which consists of *Books* will also consist of the node *Shopping* as that transaction also belongs to *Shopping*. If we model each search session as a transaction of taxonomy nodes, as proposed in the first approach, and mine coverage patterns, we will get coverage patterns where items are dependent upon each other. Thus the flat model of coverage patterns cannot be employed to allocate children nodes of bidden nodes to advertisers. We extend the model of coverage patterns by proposing the notion of *level-wise coverage patterns* which takes a taxonomy and search query logs as input and outputs coverage patterns amongst items which are at the same level in the taxonomy. Thus, the notion of level-wise coverage patterns helps in resolving the interdependence of nodes in coverage patterns. Using the notion of *level-wise coverage patterns* and *bidding on high level concepts through a multi-level taxonomy*, we propose an end-to-end framework which takes search query logs, bidding taxonomy and advertising demands as inputs and allocates advertisements to an incoming query. Experiments results on AOL search query logs show that the proposed approach can help in better utilization of space for sponsored search.

In both approaches, we propose to bid upon concepts instead of keywords in sponsored search auctions. The notion of concept-based bidding will result in capitalization of ad space of the tail queries as each concept will be composed of both head and tail keywords. Hence, each keyword would be considered for bidding based on its relevancy rather than frequency.

1.6 Contributions of Thesis

The contributions of the thesis are as follows:

(i). **An approach to bid on concepts in sponsored search:** In this thesis, we propose to sell high level concepts (using a two level taxonomy) in the ad space auctions instead of individual keywords. We argue that high order logical concepts are composed of head and tail keywords alike and bidding on such concepts would result in the capitalization of ad space of search keywords irrespective of their frequency.

(ii). **An approach to bid on a taxonomy in sponsored search:** In this approach, we propose to use a multi-level taxonomy for ad space auctions to provide more flexibility to the advertisers to target potential consumers.

(iii). **Algorithms and optimization objectives:** Algorithms and optimization objectives for the above approaches are also a part of this thesis.

(iv). **Experimental results:** Experiments have been conducted on a real world dataset of AOL search query logs to compare the performance of proposed approaches with the traditional sponsored search approach.

(v). **Literature survey:** The thesis provide a detailed survey on long tail advertising in sponsored search and on coverage patterns.

1.7 Organization of Thesis

The rest of the thesis is organized as follows:

(i). In Chapter 2, the Related Work for the sponsored search mechanisms and long tail advertising has been covered. We also discuss related literature of coverage patterns, as coverage patterns constitute to be a main part of the proposed approaches.

(ii). In Chapter 3, we discuss the sponsored search architecture and an overview of coverage patterns.

(iii). In Chapter 4, the first approach for long tail advertising in sponsored search has been proposed where we propose to extend the model of bidding on keywords in sponsored search to bidding on higher level concepts through a two level taxonomy.

(iv). In Chapter 5, the second approach for long tail advertising in sponsored search has been proposed where we generalize the first approach to use multi-level taxonomy during ad space auctions.

(v). In Chapter 6, we discuss conclusions and directions to future work.

Chapter 2

Related Work

In this chapter, we discuss some of key papers related to sponsored search, followed by the relevant work done towards long tail queries and coverage patterns.

2.1 Sponsored Search

The most important research issue in sponsored search is the relevance of an ad for a given query. In the past, attempts have been made to use the traditional IR approaches to improve ad query relevance for sponsored search [61]. The authors in [61] compared traditional methods - BM25, modified TF-IDF, language models and translation models for ad-query relevance. The authors further modified the translation models by augmenting it with the probability that the ad will be clicked for a query-ad pair through a mixture model. In another study [37], the authors modelled ad-query relevance as a machine learning problem. The authors proposed to use 19 initial features including common n-grams between ad copy and query and cosine similarity between the ad and the query. The author further improved the model by adding 'click' features (like frequency of clicks and position when the ad was clicked). For new ads, the authors proposed to leverage higher order click aggregations such as ad campaign, ad category, etc.

Sponsored search has also been studied from the perspective of revenue maximization. It has been often modeled as an online bipartite graph where one set of the graph (the ads) is known where as the other set is coming online (the queries) [55, 56]. The goal of revenue maximization is to match incoming queries to existing ads to maximize the revenue. In [56], the authors framed an algorithm considering the same bipartite model such that there are 'n' bidders, each with a daily budget, 'b' and the search engine is paid only when a user clicks on the ad. The algorithm gives a competitive ratio of $1 - 1/e$ as 'b' tends to infinity and also gives $1 - 1/e$ as the lower bound. More detailed survey on further optimization on the revenue is discussed in [55] with respect to bidding assumptions, display advertising and submodular welfare maximization.

Another key issue in sponsored search is CTR prediction to improve the revenue. In [35], the authors proposed a new online Bayesian algorithm for predicting CTR. The proposed algorithm is based on

regression that maps input features to probabilities. It maintains Gaussian beliefs over weights of the model and performs Gaussian online updates derived from approximate message passing. The authors empirically show that the algorithm performs better than the considered baselines on Bing Ads. In [63], the authors attempted to model the problem of CTR estimate for new ads as a machine learning problem. The authors selected and engineered features to train a logistic regression model. The features used in the model included *term wise CTR* of the ad and CTRs of *related ads*.

Apart from the above mentioned issues, research has also been carried to prevent click frauds [52], understanding bidding behaviors [46] and auction design [44]. A detailed survey on sponsored search and its research issues is provided by Trimponias *et al.* [75].

2.2 Long tail advertising in sponsored search

In the literature, the challenges of long tail queries are mainly addressed through query expansion or query reformulations. In this section, we review the related literature with respect to the same two approaches for addressing long tail challenges in sponsored search.

2.2.1 Query Expansion

In [12], the authors propose a taxonomy based classification method for query classification to leverage additional information for rare queries. The authors proposed to build the taxonomy by using the indexed documents and classifying them through a centroid based method and manually labelling the centroids into a taxonomy through human annotators. The authors proposed to take the search results returned by the query as the external data to classify the query into one of the classes. The conditional probability of query, q, belonging to a class, C in the taxonomy is the joint probability of the query belonging to the returned search result, $P(d|q)$, and probability of the search result belong to the class, C, $P(C|d)$. The authors maximize the *relevance* between query and first 'k' organic documents to perform query classification.

The authors in [13] propose to use web relevance feedback for the expansion of queries for sponsored search. It was proposed to augment the query by three means - terms from search results of the query, external taxonomic features based on the query and lexicon features extracted from the indexed documents of the search engine. The authors proposed to extract important terms from the search results using TF-IDF, with a logarithm term frequency and IDF computed over the ad corpus. For lexicon extraction, the authors employed an inhouse tool which will identify lexicons during crawling and indexing. For taxonomy based features, the authors used a modification of their previous work [12]. The taxonomy in this paper was built using ads titles which was populated by means of human annotators.

The authors in [11] further extended the work in [12, 13]. For the expansion of tail queries, the authors proposed to build an inverted index of the head queries and their features and use it as a look-up table during the query time. Query features (unigrams, phrases and taxonomy classes) were computed

as stated in [13]. The authors used this feature vector to do a TF-IDF which was built on the ad corpus. The authors validated their approach on 400 head queries and 400 tail queries and stated significant improvement over baselines.

Another direction to expand rare queries was taken by using the notion of feedback from the search logs [65]. The authors proposed to use the two bi-partite graphs - one graph of clicked URLs and another graph of skipped URLs for each query. The authors showed that clicked URLs and skipped URLs have similar click patterns for rare queries. The authors used the signals from the skip graph to smoothen the click graph. An optimized random walk on the click graph was framed by leveraging skip-graph to correlate the URLs for similar queries. The authors employed gradient descent to optimize the parameters of the random walk for long tail queries. The proposed model was highly scalable as it only uses the click and skip frequencies for each query and URL pair.

2.2.2 Query Reformulation

Apart from query expansion, another approach to address the challenges of tail queries is by means of query reformulation, either in real time or at the end of search result page as query suggestions.

In [86], the authors have attempted to perform a matching from tail queries to head queries. The authors modified the relevance function, $f(q, d)$, for fetching better quality results for tail queries. The authors argued that using the notion of collaborative filtering from recommendation systems, related head queries could be potentially used to improve the ads and search results of tail queries. The authors augmented the relevance function for a document and a query for organic content retrieval by adding an extra term based on similar head queries, where similarity between two queries was measured by considering three types of features: string matching, term matching and semantic matching.. The authors minimized the risk on the augmented objective function with respect to the training data to give query reformulations for tail queries.

The authors in [77] proposed to leverage Entity Linking in order to match head and tail queries. Using AOL search query logs, the authors extracted two sets of queries - head and tail. In the two sets, the authors tried to identify *spots* (possible Entities) and then used an open source tool, Dexter to identify the entities using the word-n-grams of the search queries. The authors finally compared the overlap ratios between the head and tail partitions to understand the performance of the proposed approach. In a more recent study [40], the authors have tried to use Knowledge Bases (KB) such as Yago [72] for query suggestions for long tail queries. The authors proposed to do entity linking on the search queries and use the entities to extract more knowledge from the KB. More formally, the authors attempted to integrate the Query-Flow graph (the graph which associate queries which have been seen in the same sessions) and the KB of entities. In the first step, entities are identified in the search query. In the next step, the entities are expanded by means of the KB and assigned a weight. Using the expanded entities from the second step, for each entity a Personalized Page Rank [18] is performed on the Query-Flow graph to fetch related queries as the query suggestions.

In another front, authors have tried to use templates to provide query recommendations for the tail queries [73]. The approach was motivated by the fact that most distinct queries follow similar patterns. For example, "Los Angeles hotels", "New York hotels" and "Paris hotels", could be generalized to "city-name hotels". Hence, if "Paris restaurants" is a query recommendation for "Paris hotels" and "Chicago restaurants" is the recommendation for "Chicago hotels", then based on certain confidence, a template rule can be defined as *city-name hotels* $\implies$ *city-name restaurants*. The authors proposed to construct such rules by means of a hierarchy over entities. The hierarchy was built using WordNet 3.0 hypernymy hierarchy and the Wikipedia category hierarchy. The authors suggest to use every unigram, bigram and trigram to be replacement candidates for the entities.

Very recently, authors in [66] employed deep learning for query suggestions. The papers claims that while past approaches can only suggest previously seen queries as query suggestions, the proposed approach is capable of generating *synthetic suggestions* as per the need. Their work is particularly important for long tail queries as these queries are seen very less and even uniquely, and hence, matching such queries to head queries is not always successful. The authors modified Recurrent Neural Networks (RNNs) into Hierarchical Recurrent Encoder-Decoder. Queries are fed word by word to an RNN and query sessions are fed query by query parallely using another set of RNNs. The training objective of the approach is maximizing the log-likelihood of a session in predicting the next query in the session after seeing the previous ones. The proposed approach showed an increase of 5.6% over the standard baselines on long tail queries sampled from AOL search query logs.

2.3 Coverage Patterns

Coverage patterns is the key pillar of the approaches proposed in this thesis. Hence, in this section we review the work done in coverage patterns.

The model of coverage patterns was proposed in [68, 71]. In [71], the notion of coverage patterns was proposed by defining Coverage Support (CS), minimum Relative Frequency (RF) and Overlap Ratio (OR). For a given coverage pattern $\{X_1, X_2, ...X_n\}$, the authors defined CS to be the fraction of transactions containing the items in the coverage pattern. OR was defined as the fraction of transactions which were common in the items contained in the coverage pattern. Minimum RF was defined to remove any items whose frequency was small enough to be ignored while mining interesting patterns. Based on the concepts defined in [71], the authors in [68] proposed an Apriori style algorithm to mine coverage patterns through the downward closure property of OR. The authors also incorporated bit level computations to improve the performance. The extraction of coverage patterns was further improved in [70] by using a pattern growth technique, Coverage Pattern Projected Growth (CPPG) which is inspired from Frequent Pattern Projected Growth (FPPG).

The utility of coverage patterns has been discussed in [47]. In this study, coverage patterns have been applied to display advertising. The authors proposed to leverage click logs of a website to extract

coverage patterns. The extracted coverage patterns were then used to perform an intelligent matching between guaranteed contracts of the advertisers and the web pages to optimize the publisher's revenue.

2.4 How Proposed Approaches Are Different

In this thesis, we have proposed two approaches for long tail advertising in sponsored search. Earlier approaches to address long tail query challenges in sponsored search involve either query expansion [11–13, 65] or query reformulation [40, 66, 73, 77, 86]. In this thesis, we propose to bid on high level concepts in ad space auctions instead of individual keywords. In the first approach, we propose use a two level taxonomy for auctioning ad space in search keywords. In the second approach, we generalize it to use a multi-level taxonomy to provide more flexibility to the advertisers. Both proposed approaches argue to move beyond individual keyword bidding in search engine advertising. Compared to the earlier approaches where taxonomies or external knowledge have been used ***internally*** for query expansion or query reformulations, in our approach ***the taxonomy is directly exposed to the advertisers*** during the ad space auctions to target potential consumers. An advertiser is in power to chose which set of concept(s) he/she wants to bid upon.

Also, the notion of coverage patterns proposed in the literature [70, 71] is only for flat transactions. We extend this notion of coverage patterns to *level-wise coverage patterns*, when a hierarchical relationship exists over the items of the transactional database.

Chapter 3

Background: Sponsored Search and Coverage Patterns

In this chapter, we will discuss the background on how sponsored search works followed by the notion of coverage patterns, which is the key pillar in the proposed approaches.

3.1 Sponsored Search Framework

In sponsored search, an advertiser has a daily budget and bids on a set of keywords. For every query that is fired, the set of advertisers who have bid upon any keywords contained in the query are ranked on the basis of their bids and top-k advertisers are chosen to be displayed as the sponsored results.

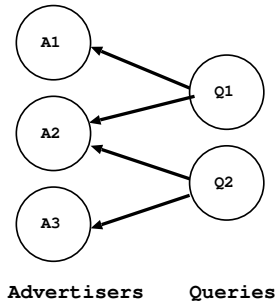


Figure 3.1: Sponsored Search Model

Mehta *et al.* [56] framed the problem of sponsored search as a bipartite matching problem. An instance of the same is shown in Figure 3.1. In this model, the left side is the set of advertisers and the right side is the set of incoming queries which are to be matched to the advertisers. When a query is fired by a user, it is to be matched to the relevant advertisers. The advertisers are then ranked and their advertisements are displayed along with query results.

Figure 3.2 shows the framework of sponsored search by considering single query as an input and a list of candidate advertisers to be displayed with the results as the output. The framework contains the following steps.

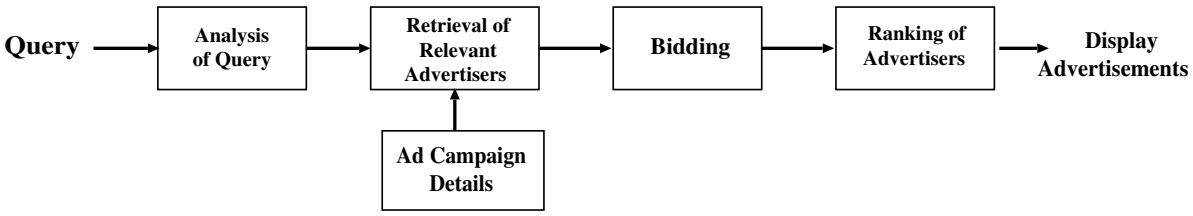


Figure 3.2: Sponsored Search Architecture as discussed in [56]

(i). *Analysis of Query*: In this part, a search query is analyzed for the purpose of identification of information which is useful for showing advertisements on search results page. This information could be in form of topic of the query, previous queries in the session and topics explored in the session apart from several other parameters.

(ii). *Retrieval of Relevant Advertisers*: In this step, relevant advertisements are retrieved from the created ad campaigns. This is performed by employing appropriate similarity measures between the bid keywords and the query keywords.

(iii). *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's results page. These bids can either be static or can be done in real time.

(iv). *Ranking of Advertisers*: Once the advertisers bid on a query, their bids are scaled according to a factor called *Quality Score*. *Quality Score* is computed based on the parameters related to the respective advertisement. This includes expected *CTR (Click Through Rate)*, display URL's past *CTR*, quality of the landing page, remaining budget and ad-query relevance apart from several other parameters [38].

3.2 Overview of Coverage Patterns Model

The model of coverage patterns was proposed in [70, 71] for efficient banner advertising. The click-through data of a website represents the navigational patterns of multiple visitors. The model of coverage patterns is used to mine these click-stream logs to extract interesting combination of web pages for advertising. Let $W = \{w_1, w_2, w_3 \cdots, w_n\}$ be the set of web page identifiers and D is the set of transactions, where each transaction T is the set of web pages such that $T \subseteq W$. A set of web pages $X \subseteq W$ i.e., $X = \{w_p, \cdots, w_q, w_r\}, 1 \leq p \leq q \leq r \leq n$ is a pattern. Every web page w_i is related to multiple topics, $\{tp_1, tp_2, tp_3, \cdots tp_m\}$, where m is the number of topics in w_i. Set of transactions containing web page w_i is denoted as T^{w_i} and its cardinality is denoted as $|T^{w_i}|$. Transactional database, D is given in Table 3.1.

Table 3.1: Transactional database, D

TID	1	2	3	4	5	6	7	8	9	10
Web pages	b, a, c	a, c, e	a, c, e	a, c, d	b, d, f	b, d	b, d	b, e	b, e	b, a

Web pages are said to be potential web pages for placing advertisement when they occur in more number of transactions. This means that web pages are having more number of visitors. This is captured by the aspect of *relative frequency*.

Definition 1 *(Relative frequency (RF)) and Frequent web page) The RF of a web page w_i, denoted by $RF(w_i)$, is equal to the ratio of number of transactions that contain w_i to D, i.e., $RF(w_i) = \frac{|T^{w_i}|}{|D|}$. Let the term 'minimum relative frequency (minRF)' indicate user-specified threshold value. A web page is frequent if it is no less than minimum threshold frequency, $minRF$, i.e., RF(w_i) $\geq minRF$.*

Given a pattern, it is interesting to find the users visiting at least one of the web pages in the pattern. It is interesting because if we place the advertisement on all the web pages in the pattern, it guarantees the delivery of the advertisement to the users visiting at least one of the web pages in the pattern. This is captured by the aspect of *coverage set*.

Definition 2 *(Coverage set and* coverage support (CS) *of a pattern $X = \{w_p, \cdots, w_q, w_r\}$, $1 \leq p, q, r \leq n$) The set of distinct transaction ids containing at least one web page of X is called coverage set of pattern X and is denoted as $CSet(X)$. Therefore, $CSet(X) = T^{w_p} \cup \cdots \cup T^{w_q} \cup T^{w_r}$. The ratio of the size of the $CSet(X)$ to D is called the coverage-support of pattern X and is denoted as $CS(X)$, i.e., $CS(X) = \frac{|CSet(X)|}{|D|}$.*

Given a pattern, adding a new item which co-occur with any of the items in the dataset may not increase the coverage support significantly. This is not interesting from the banner advertisement perspective as the same user is visiting these web pages having same advertisement. The new pattern formed by adding the new item is interesting it there is minimum overlap between the *coverage sets* of web pages. This is captured by the aspect of *overlap-ratio*.

Definition 3 *(Overlap ratio (OR) of a pattern.) OR of a pattern $X = \{w_p, \cdots, w_q, w_r\}$, where $1 \leq p, q, r \leq n$ and $|T^{w_p}| \geq \cdots \geq |T^{w_q}| \geq |T^{w_r}|$, is the ratio of the number of transactions common in $X - \{w_r\}$ and $\{w_r\}$ to the number of transactions in w_r, i.e., $OR(X) = \frac{|(Cset(X - \{w_r\})) \cap (Cset\{w_r\})|}{|Cset\{w_r\}|}$.*

The items in the pattern are ordered in descending order of frequency to satisfy the property of *sorted downward closure property* which is later helpful in mining coverage patterns.

A pattern X is said to be *non-overlap pattern* if $OR(X)$ is no greater than $maxOR$ and $\forall w_i \in X$, $RF(w_i) \geq minRF$. An item 'a' is said to be *non-overlap item* with respect to X, if $OR(X, a)$ is no greater than $maxOR$.

A *coverage pattern* is said to be interesting if it has high CS and low OR. It is interesting because having high CS means showing the advertisement to many users and having low OR means decreasing the repetitive display of advertisement. i.e., a pattern X is said to be interesting if $CS(X) \geq minCS$, $OR(X) \leq maxOR$, and $RF(w_i) \geq minRF\ \forall w_i \in X$. A coverage pattern X having $CS(X) = a\%$ and $OR(X) = b\%$ is denoted as follows:

$$X \quad [CS = a\%,\ OR = b\%]$$

Example 1 *From Table 3.1, the relative frequency of 'a' i.e.,* $RF(a) = \frac{|T^a|}{|D|} = \frac{5}{10} = 0.5$. *If the user-specified* $minRF = 0.5$*, then 'a' is called a frequent web page because* $RF(a) \geq minRF$. *Similarly for 'b' relative frequency is* $RF(b) = \frac{|T^b|}{|D|} = \frac{7}{10} = 0.7$. *'b' is called a frequent web page because* $RF(b) \geq minRF$. *The set of web pages 'a' and 'b' i.e.,* $\{a, b\}$ *is a pattern. The set of* $TIDs$ *containing the web page 'a' i.e.,* $T^a = \{1, 2, 3, 4, 10\}$. *Similarly,* $T^b = \{1, 5, 6, 7, 8, 9, 10\}$. *The coverage set of* $\{a, b\}$ *i.e.,* $CSet(\{a, b\}) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. *Therefore, coverage support of* $\{a, b\}$ *i.e.,* $CS(\{a, b\}) = \frac{|CSet(\{a,b\})|}{|D|} = \frac{10}{10} = 1$. *The* $OR(\{a, b\}) = \frac{|CSet(b) \cap CSet(a)|}{|CSet(a)|} = \frac{2}{10} = 0.2$. *If* $minRF = 0.4$*,* $minCS = 0.7$ *and* $maxOR = 0.5$*, then the pattern* $\{a, b\}$ *is a coverage pattern. It is because* $RF(a) \geq minRF$*,* $RF(b) \geq minRF$*,* $CS(\{a, b\}) \geq minCS$ *and* $OR(\{a, b\}) \leq maxOR$. *This pattern is written as follows:*

$$\{a, b\} \quad [CS = 1\ (= 100\%),\ OR = 0.1\ (= 10\%)]$$

In this rest of this section, we first give an overview of the approaches proposed for extracted coverage patterns followed by an overview of the CMine algorithm.

3.2.1 Overview of approaches to mine Coverage Patterns

The problem statement of mining coverage patterns is as follows. Given a transactional database D, set of web pages W (or items), and the user-specified *minimum relative frequency* ($minRF$), *minimum coverage support* ($minCS$) and *maximum overlap ratio* ($maxOR$), discover the complete set of coverage patterns in the database that satisfy *minRF*, *minCS* and *maxOR* thresholds. A naive approach is exponential because for a given website of 'n' web pages, we have to check for $minCS$ and $maxOR$ for all $(2^n - 1)$ sets of web pages.

An Apriori like approach, *CMine* which is based on multiple pass, candidate test and generation approach is proposed in [68]. It uses the fact that every non-overlap pattern is coverage pattern and by using the sorted closure property of overlap ratio the search space is minimized. Later, a pattern-growth approach, *coverage pattern projected growth (CPPG)* based on the notion of *non-overlap projection* is proposed in [69]. An enhanced coverage pattern projected growth approach, *ECPPG* is proposed in [70] by using the notion of *upward closure property* of coverage patterns.

The overview of *CMine* algorithm is as follows.

3.2.2 CMine Algorithm with Example

Let F be a set of frequent items, C_k be a set of candidate k-patterns, L_k be a set of coverage k-patterns and NO_k be a set of non-overlap k-patterns. The algorithm *CMine* begins with a scan of database and discovers set of all frequent web pages (denoted as F) and coverage 1-patterns (denoted as C_1). Non-overlap patterns (denoted as NO_1) will be the set of all frequent 1 items. Next, the items in NO_1 are sorted in descending order of their frequencies. Using NO_1 as the seed set, candidate patterns C_2 are generated by combining $NO_1 \bowtie NO_1$. From C_2, the patterns that satisfy $minCS$ and $maxOR$ are generated as L_2. Simultaneously, all candidate 2-patterns that satisfy $maxOR$ constraints are generated as non-overlap 2-patterns, NO_2. Since *overlap ratio* satisfies sorted closure property, C_3 is generated by combining $NO_2 \bowtie NO_2$. This process is repeated until no new coverage patterns are found or no new candidate patterns can be generated. The proposed algorithm uses bitwise OR and AND operations to find the *coverage support* and *overlap ratio* of a pattern, respectively. So, single scan of database is sufficient to find the bit strings of all single web pages and to extract the complete set of coverage patterns.

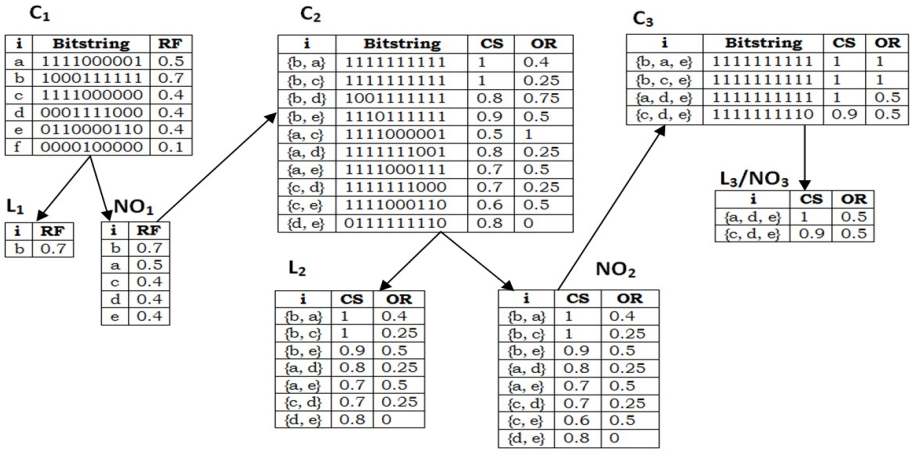


Figure 3.3: Working of CMine algorithm. The term 'I' is an acronym for the item set (or web pages).

We now explain the working of *CMine* algorithm using the transactional database, D, shown in Table 3.1 for the user-specified $minRF$, $minCS$ and $maxOR$ as 0.4, 0.7 and 0.5, respectively. We use Figure 3.3 to illustrate the *CMine* algorithm for finding coverage patterns in D.

The algorithm *CMine* scans all the transactions to generate bit string $|B^{w_i}|$ and relative frequencies (RF) of each web page w_i. $RF(w_i) = \frac{|B^{w_i}|}{|T|}$. $|B^{w_i}|$ denotes the number of 1's in the bit string. Each web page, $w_i \; \epsilon \; T$ which has a relative frequency no less than 0.4 is a member of the set of candidate 1-pattern, C_1. From C_1, the set of coverage 1-patterns, L_1, are discovered if their frequencies are greater than or equal to $minCS$. Simultaneously, set of non-overlap 1-patterns, NO_1, are discovered if candidate 1-patterns have relative support greater than or equal to $minRF$ and finally the web pages in NO_1 are sorted in the decreasing order of their frequencies. To discover the set of coverage 2-patterns, L_2, the algorithm computes the join of $NO_1 \bowtie NO_1$ to generate a candidate set of 2-patterns, C_2. Next, *overlap ratio* and *coverage ratio* for each candidate pattern is computed. *Coverage support* is computed by boolean OR operation and *overlap ratio* is computed by boolean AND operation. For example, $CS(\{b, a\}) = \frac{|B^b \vee B^a|}{|T|} = \frac{|1111111111|}{10} = 1.0$ and $OR(\{b, a\}) = \frac{|B^b \wedge B^a|}{|B^a|} = \frac{|10000001|}{10} = \frac{2}{10} = 0.4$.

The columns titled 'CS' and 'OR' respectively show the *coverage support* and *overlap ratio* for the patterns. The set of candidate 2-patterns that satisfy $maxOR$ are discovered as non-overlap 2-patterns, denoted as NO_2. Simultaneously, the set of candidate 2-patterns that satisfy both $minCS$ and $maxOR$ are discovered as coverage 2-patterns. Next, C_3 is generated by $NO_2 \bowtie NO_2$. We discover non-overlap 3-patterns, NO_3, and coverage 3-patterns, L_3 in the same manner that is stated above. The algorithm stops as no more candidate 4-patterns can be generated from non-overlap 3-patterns.

3.3 Summary

In this chapter, we discuss how sponsored search works followed by an overview of coverage patterns. We discussed the model of coverage patterns followed by approaches to extract coverage patterns. Cmine approach was discussed in detail.

Chapter 4

Exploiting ad space of tail queries using a two level taxonomy and coverage patterns

In this chapter, we propose the first approach to long tail advertising in sponsored search. We propose that advertisers should bid upon high level concepts instead of keywords in ad space auctions. We formulate the problem as an optimization objective to maximize the revenue of the search engine and give an approximate algorithm for it along with an end-to-end framework.

4.1 Motivation and Basic Idea

Search queries follow a long tail distribution with a small head of frequent queries and a long tail of infrequent queries. Advertising on tail queries is hard mainly due to the data sparsity problem as the frequency of tail queries is very less. Also, it has been observed that advertisers tend to bid for head keywords in the ad space auctions because of the individually high reach of head keywords. Further, the long tail of search queries makes it quite difficult for the advertisers to identify all the relevant keywords for their products. The stated factors result in significant amount of under-utilization of ad space of the tail queries which is identified as the research gap.

In this chapter, we propose an approach to exploit the ad space of tail keywords. We propose that instead of bidding on keywords advertisers should bid upon high level concepts in ad space auctions. The motivation behind bidding on concepts is inspired from the fact that each concept can be logically composed of multiple keywords, head and tail alike. The reach of an individual tail query keyword may not be desirable. However, the reach of a set of keywords containing both head and tail query keywords or just tail keywords can be considered equivalent to head keywords. For example, the word "airline" is very popular in the AOL search query logs but the words like { first, jack, pennsylvania, mount, wildlife } are not as popular. A group of such keywords could be created to compete with the head keywords in ad space auctions.

The key idea of this approach is to sell groups of keywords in ad space auctions. The keywords are grouped together through high level concepts. We propose that in ad space auctions of search

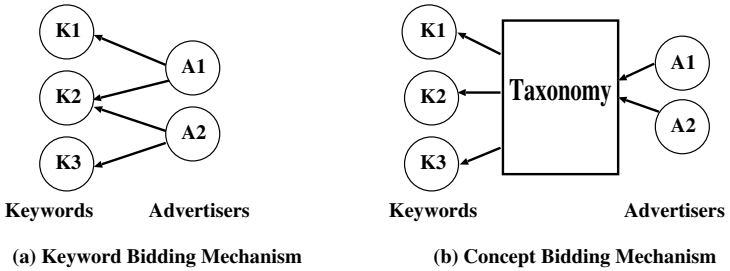


Figure 4.1: Sponsored search bidding: keywords based bidding and concept-based bidding

keywords, advertisers should bid upon high level concepts instead of individual keywords. Bidding on higher level concepts is achieved through a two-level taxonomy. During ad space auctions, advertisers are shown the first level of the taxonomy and are asked to select a node which seems to be the most relevant to their product. For example, if an advertiser wants to bid upon shopping, he/she will select the concept *Shopping* from the first level nodes. Thus, we propose to add a middle layer of concepts through a taxonomy during the bidding process such that an advertiser would choose 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. Figure 4.1 (a) shows the keywords based approach where advertisers bid upon keywords where advertiser A_1 chose to bid upon keywords K_1 and K_2 and advertiser A_2 chose to bid upon keywords K_2 and K_3. Figure 4.1 (b) shows the concept-based bidding approach where advertisers bid on concepts through a taxonomy which are composed of search keywords.

In the proposed approach, each advertiser selects a node to bid upon and groups of children nodes of bidding nodes are assigned to the advertisers. To form such groups of children nodes, we mine coverage patterns from session-based transactions derived from search query logs. Each search session is modelled as a transaction with queries occurring in the session as the items. Each query is further replaced by the corresponding first and second level nodes from the bidding taxonomy. Coverage patterns are extracted from these session based transactions using the approach proposed in [70]. Compared to the arbitrary grouping on nodes, the extracted coverage patterns would help in identifying sets of nodes of the children nodes which ensure maximal overlap, thereby reducing the repetition of ads in the same session. We propose to perform a matching between the extracted coverage patterns and the advertising demands. We propose to model the matching as an optimization objective to maximize the revenue of the search engine and propose an end-to-end framework to realize the matching and to allow the advertisers to bid on a two-level taxonomy.

The rest of the chapter is structured as follows. In Section 4.2, the proposed model is discussed. The proposed framework in discussed in Section 4.3 including the optimization objectives and algorithms followed by experiments in Section 4.4, and assumptions and limitations in Section 4.5. Section 4.6 summarizes the chapter.

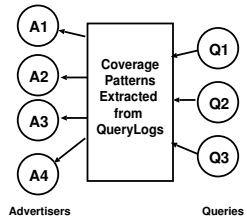


Figure 4.2: Proposed model

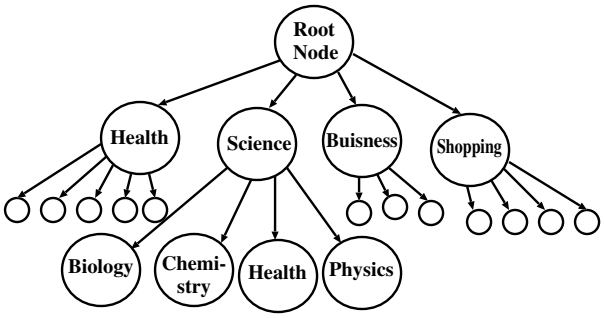


Figure 4.3: Example of a concept taxonomy

4.2 Proposed Model

When a query is fired by a user to the search engine, the query is first classified according to the taxonomy and the advertisers who have been assigned the corresponding nodes are the candidate advertisers for the query. Thus, in the proposed model, a middle layer of coverage patterns has been added between keywords and advertisers as shown in Figure 4.2. The left most side of the diagram shows one disjoint set as the advertisers and the right side has another disjoint set as queries with coverage patterns of taxonomy nodes in the middle.

Through the proposed model, it is possible to exploit more ad space of the tail queries as each advertiser is allocated a set of concepts and each concept would contain a mix of head and tail keywords. Thus, all the keywords, head or tail, would be considered for advertising based on their relevancy rather than frequency.

Using the proposed model, we have defined an end-to-end framework for allocating incoming queries to advertisers. The proposed framework is discussed in the next section.

4.3 Proposed Framework

The proposed framework is shown in Figure 4.4. It has the following steps.

(i). ***Analysis of Query***: This step is same as the original sponsored search systems approach with an addition to classify the incoming query using a taxonomy. We have proposed the approach by considering a two-level taxonomy. A dummy root node is level zero, nodes at level one are termed

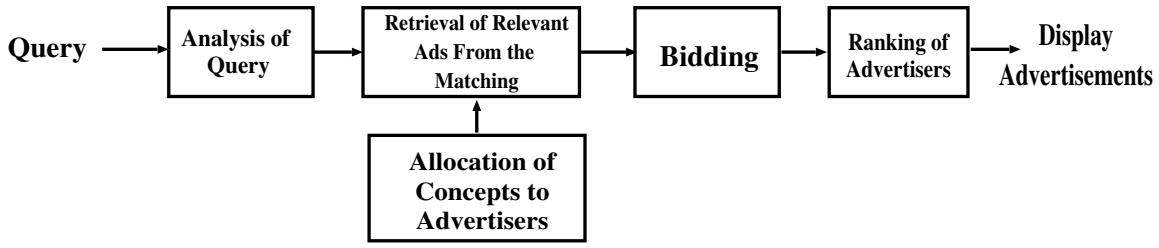


Figure 4.4: Proposed Architecture

as concepts and nodes at level two are sub-concepts of their parent concept as shown in Figure 4.3. Given a query, we extract the corresponding concept and sub-concept pair from the concept taxonomy along with other information as required by the sponsored search system. (Note that we propose that the advertiser should bid upon a node in the taxonomy. In this study, we investigate the setting such that advertisers are only bidding on the nodes in level 1. It is possible to extend the proposed approach by considering a taxonomy of higher number of levels with appropriate modifications, which will be explored in the next chapter.)

(ii). ***Retrieve Relevant Ads from the Matching***: This step identifies the relevant advertisers for a given query. Compared to the proposed approach, we have modified it such that it takes search query and the matching between advertisers & coverage patterns (which is the output of the allocation of concepts to advertisers component) as inputs. Using the matching, we retrieve advertisers who have been allotted the coverage pattern containing the sub-concept of the incoming query. In the next subsection, we will explain the component of allocation of concepts to advertisers and how the matching is computed.

(iii). ***Bidding***: Ad space on each search query is sold by means of auctions. Each advertiser bids for the ad slots. This bid can be static or dynamic. (This step is the same as the standard sponsored search approach.)

(iv). ***Ranking of Advertisers***: Based on the bid amount and other relevant parameters like CTR and ad-query relevance, advertisers are ranked to be shown on the query's results page. (This step is also same as the standard sponsored search approach.)

In the remaining section, we will discuss how to allocate concepts to advertisers and compute the matching between advertisers and coverage pattern.

4.3.1 Allocation of concepts to advertisers

In this sub-section, we explain the process of allocation of the concepts of taxonomy to advertisers. The process of allocation is divided into the following steps, as shown in Figure 4.5.

(i). *Conversion of Query Logs to Taxonomic Transactions*

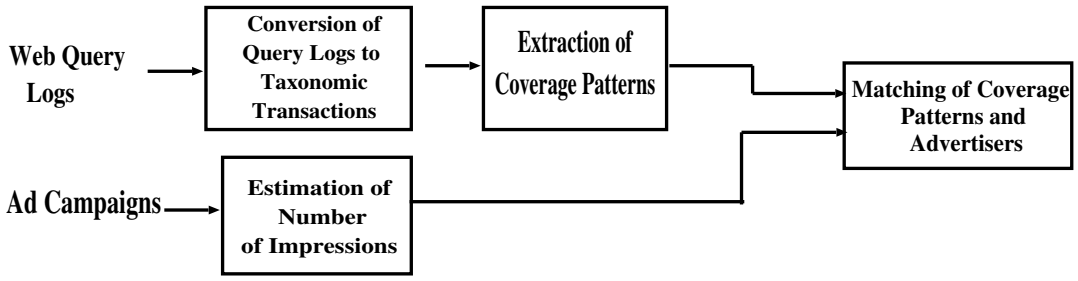


Figure 4.5: Allocation of concepts to advertisers

(ii). *Extraction of Coverage Patterns*

(iii). *Estimation of Number of Impressions (of Advertisers)*

(iv). *Matching Coverage Patterns and Advertisers*

In the rest of the subsection, we explain each step in detail.

4.3.1.1 Conversion of Query Logs to Taxonomic Transactions

We extract search sessions from the given web query logs. The search sessions are transformed into session based transactions. Queries in the same session form the items of the corresponding session transaction. We further transform each transaction by replacing each query with it's concept and sub-concept pair using the taxonomy. The transformed transaction is termed as 'taxonomic transaction'.

Example 2: We explain all the steps of the framework through a running example by considering a set of advertisers who chose to bid on the concept of *Science*. *Biology, Chemistry, Technology, Environment, Physics* and *Agriculture* are the sub-concepts of *Science*. Conversion of query logs into transactions is shown in Table 4.1 and Table 4.2. Table 4.1 shows the sessions extracted from the query logs. These queries are classified into concepts and sub-concepts according to the taxonomy. For example, in session with session–id 1, there are three queries of concept *Science* out of which one belongs to the sub-concept *Biology* and two belong to the sub-concept *Chemistry*. To create a transaction, we used the session–id, concept and the sub-concepts that occur in the session. For the first session, it will be $< 1, Science, \{Biology, Chemistry\} >$. Similarly, the rest of the query logs are converted to taxonomic transactions to mine coverage patterns.

4.3.1.2 Extraction of Coverage Patterns

Extraction of coverage patterns from the search query posed two main challenges which are stated as below.

Table 4.1: Sample Sessions

Session-id	Query	Concept	Sub-concept
1	chromatography	Science	Biology
1	o-toulene	Science	Chemistry
1	dimethylene	Science	Chemistry
2	wood boring bees	Science	Agriculture
2	anthrenus carpet	Science	Biology
3	artficail intelligence	Science	Technology
4	tree symbolism	Science	Environment
5	internship computer science	Science	Technology
5	ipod prices	Science	Technology
6	health plan	Science	Biology

Table 4.2: Taxonomic Transactions

Sr. No	Concept	Sub-concepts
1	Science	{Biology, Chemistry}
2	Science	{Agriculture, Biology}
3	Science	{Technology}
4	Science	{Enviornment}
5	Science	{Technology}
6	Science	{Biology}

(A) Coverage patterns are mined from the perspective of unique visitors considering each transaction as one visitor. However, in order to leverage coverage patterns in computational advertising, the knowledge of coverage patterns needs to be converted to either impressions or clicks.

We propose to convert number of visitors of a coverage pattern into impressions. We will later use this knowledge of impressions to match coverage patterns to advertisers.

To estimate the number of impressions from the coverage patterns, we used set theory. From the definition of coverage patterns, a coverage set of a coverage pattern X is the union of all the transactions of each item in X and overlap ratio is the ratio of number of common transactions between the first $n-1$ items of coverage pattern X and the last item, n^{th} item to the number of transactions having the n^{th} candidate item. Thus, OR is roughly half of the intersection ratio between X and the candidate item.

One key assumption that we have made in the estimation is that we don't take into account the frequency of each item in a transaction and consider it one, if present and zero otherwise. From overlap ratio and coverage support, the task was to find the number of times both items have occurred together in the transaction dataset. In set theory, the following theorem gives the total number of items having both A and B.

$$n(A) + n(B) = n(A \cup B) + n(A \cap B) \tag{4.1}$$

Our requirement is to estimate $n(A) + n(B)$ where coverage support would be $n(A \cup B)$ and $\frac{OverlapRatio}{2}$ would be $n(A \cap B)$. Hence, number of impressions by a given coverage pattern, X is computed as followed.

$$Impressions(X) = (CS(X) + \frac{OR(X)}{2}) * |D| \tag{4.2}$$

where $|D|$ is the total number of transactions or this case, number of sessions.

(B) In the proposed approach, we allocated coverage patterns of taxonomy nodes to the advertisers. Given a single advertiser, which coverage pattern to allocate is the challenge we faced. To resolve the issue of selection of coverage patterns, we propose to rank coverage patterns. A ranking amongst coverage patterns could be established by multiple methods including only overlap ratio (to reduce repetition), only coverage support (to get the most number of impressions) or a function of both overlap ratio and coverage. In this approach, we determine the ranking of a coverage pattern to be the number of unique visitors that a coverage pattern can provide. A higher value of unique visitors implies a better coverage pattern. We calculate the ranking parameter of a coverage pattern as the difference between the Coverage Support and Overlap Ratio as shown in Equation 4.3.

$$RP(X) = CS(X) \ - \ OR(X) \tag{4.3}$$

A sorting algorithm of $O(nlogn)$ would be sufficient as the extraction and ranking of the patterns can happen offline.

In continuation of the example 2, we assume the size of the dataset to be of 1000 taxonomic transactions. Table 4.3 shows relative frequencies of the six considered sub-concepts of *Science* and Table 4.4 shows coverage patterns mined from the transactions.

Table 4.4 shows the extracted coverage patterns with their estimated number of impressions. For example, for the pattern {*Agri, Phy*}, the value of CS is 0.6 and that of OR is 0.17 and hence, the number of impressions for {*Agri, Phy*} is $(0.6 + \frac{0.17}{2}) * 1000 = 685$.

In Table 4.4, the coverage pattern {*Agri, Env*} is least interesting because the difference of coverage support and overlap ratio of the pattern is 0.1. It also indicates that the fraction of repeated visitors of this coverage pattern is at least 0.35. Table 4.4 shows sorted coverage patterns according to the rank using the same parameter. The value of difference of coverage support and overlap ration for {*Agri,Phy*}, {*Bio, Chem*}, {*Chem, Env*} and {*Agri, Env*} are 0.43, 0.42, 0.25 and 0.1 respectively and hence, they are ranked in that order.

Table 4.3: Relative Frequencies of Sub-Concepts

Sub-concept	Bio	Tech	Phy	Chem	Env	Agri
RF	0.4	0.3	0.3	0.45	0.35	0.45

Table 4.4: Extracted Coverage Patterns

Pattern	CS	OR	Visitors	Concept	Rank
{Agri, Phy}	0.6	0.17	685	Science	1
{Bio, Tech}	0.65	0.23	765	Science	2
{Chem, Env}	0.5	0.25	625	Science	3
{Agri, Env}	0.45	0.35	625	Science	4

4.3.1.3 Estimation of Number of Impressions of Advertisers

In this sub-section, we calculate the number of impressions required by an advertiser in the most optimal situation. An advertiser has a daily budget which is the maximum amount that can be spent per day. A minimum bid is also defined during creation of ad campaigns. Using the bid and budget we can calculate the number of clicks required on the advertisement to exhaust the advertiser's daily budget.

$$Clicks\ Needed = \frac{Budget}{Bid} \tag{4.4}$$

To estimate the number of impressions from the maximum clicks, we employ the idea from Click-Through Rate (CTR). CTR is defined as follows.

$$CTR = \frac{Clicks}{Impressions} \times 100 \tag{4.5}$$

For this study, we assume that we have the knowledge of CTR for an advertisement, either predicted or calculated from the history of the advertisement. Given the CTR information, we can estimate the number of impressions required to achieve the number of clicks to exhaust an advertiser's budget.

$$Number\ of\ Impressions = \frac{Clicks\ Needed}{CTR} \times 100 \tag{4.6}$$

We can rewrite Equation 4.6 using Equation 4.4 as follows.

$$Number\ of\ Impressions = \frac{Budget}{Bid} \times \frac{1}{CTR} \times 100 \tag{4.7}$$

We used Equation 4.7 to estimate the required impressions by each advertisement, which is shown in column 6 of Table 4.5.

Table 4.5: Table Showing Advertisements Details

Ad	Bid	Budget	CTR	Concept	Impressions
A1	0.40	5.00	5.0	Science	250
A2	0.50	4.00	3.5	Science	228
A3	0.90	10.00	7.5	Science	149
A4	1.5	10.00	5.0	Science	133
A5	2	18.00	6.5	Science	138
A6	1	12.00	6.0	Science	200
A7	0.5	4.00	3.5	Science	228
A8	2.5	25.00	7.0	Science	142
A9	1.5	18.00	7.5	Science	160

4.3.1.4 Matching of Coverage Patterns and Advertisers

In this section, we frame the matching that is to be performed between coverage patterns and the advertisers in terms of impressions provided by the coverage patterns and the impressions needed by the advertisers. In the first subsection, we describe the objective that we intend to minimize followed by an algorithm for the same.

(A) **Formation of Objective Function:**

We first define an objective for the matching and then provide an algorithm for the same later.

The goal of a search engine is to increase the revenue of the sponsored search with respect to a set of constraints. In the Cost-Per-Click (CPC) model, the search engine is paid the bid amount of the advertiser only when a user clicks on the ad. Hence, to maximize its revenue, a search engine must try to maximize the number of clicks. Thus, the objective could be framed in terms of maximizing the revenue which is the sum of clicks times bid of the ads shown, as stated in Equation 4.8.

$$Max\ Revenue = \sum (bid_{ij} \times click_{ij}) \tag{4.8a}$$

$$s.t.\ \sum (bid_{ij} \times click_{ij}) <= budget_i\ \ \forall Ad_i \in Ads \tag{4.8b}$$

Equation 4.8a captures that the total revenue of the sponsored search is maximizing the number of clicks on an ad of advertiser A_i for which the advertiser pays the bid amount bid_{ij} according to the CPC. Equation 4.8b states the constraints that the advertisers can't exceed the daily budget for the day. From the above equation, we can say that the revenue of the system depends directly upon the number of relevant clicks.

$$Revenue \propto clicks \tag{4.9}$$

Without loss of generality, one can claim that the number of clicks is directly proportional to the number of relevant impressions. (A click is also dependent on other factors including ad quality, pacing of ads, etc. The assumption here is that more impressions implies more clicks.) Hence, revenue is also proportional to impressions shown (Equation 4.10).

$$Revenue \propto clicks \propto (Impressions\ Shown) \tag{4.10}$$

Therefore, if the number of impressions shown are more, an increase in revenue is expected. Conversely, we can argue that if we minimize the number of unused impressions, we can increase the revenue. Hence, in the proposed approach, we forecast the supply in the form of coverage patterns and perform an approximate matching between forecasted supply and demand of advertisers. We also have the demands of the advertisers in number of impressions. Therefore, the goal of the matching becomes to minimize this difference between forecasted supply and the demand of the advertisers, which is as follows.

$$Min\ Z = (Estimated\ Supply - Demand) \tag{4.11}$$

Equation 4.11 can be refined by considering the problem with respect to number of impressions. *Estimated Supply (ES)* is in the form of coverage patterns which can be calculated as follows.

$$ES\ = \sum_{i=1}^{m}(Impressions(Node_i)) \tag{4.12}$$

where $Impressions(Node_i)$ indicates the number of impressions supplied by a node of the taxonomy. As stated earlier, in this study, we have limited our scope to only a two level taxonomy. Thus, Equation 4.12 will be used at the nodes of the second level nodes as they will compose the extracted coverage patterns.

Demand (D) can be expressed as the sum of the demands of all the advertisers. Using Equation 4.7, the expression of D is as follows.

$$D = \sum_{i=1}^{n}(\frac{Budget_i}{Bid_i} \times \frac{1}{CTR_i} \times 100) \tag{4.13}$$

However, a coverage pattern is a collection of sub-concepts which are high level nodes in the taxonomy encompassing several keywords. Therefore, the number of impressions provided by a coverage pattern would be remarkably greater than any individual advertiser's requirement. Hence, the matching between coverage patterns and advertisers is one-to-many matching. In this regard, we use Equation 4.13 to re-frame the Equation 4.11 as follows.

$$Min\ Z' = \sum_{j}^{m}(Imp(CP(X_j))$$

$$-\sum_{i}^{n}(\frac{Budget_i}{Bid_i} \times \frac{1}{CTR_i} \times 100 \times X_{ij})) \tag{4.14a}$$

$$s.t\ \sum(bid_{ij} \times click_{ij}) <= budget_i\ \ \forall Ad_i \in Ads \tag{4.14b}$$

Equation 4.14a states the objective to minimize the difference between estimated supply and demand in a one-to-many coverage patterns to advertiser matching such that difference of impressions provided by a coverage pattern (X_j) and the sum of impressions required by the advertisers allotted to the coverage pattern is minimized. Equation 4.14b determines the constraint that the advertisers cannot spend more than their daily budget.

(B) **Algorithm for Matching:**

Using the objective function defined in the last subsection, we propose a greedy algorithm based on *High Water Mark* algorithm as discussed in [10]. The algorithm takes as input, a list of advertisers Ad_list, a list coverage patterns, CP_list, a complete list of Sub-concepts, SCs and a greedy parameter ϵ and returns a matching of advertisers and coverage patterns in the form of a list.

Since it is a one-to-many matching between coverage patterns and advertisers, iterating over coverage patterns is a more appropriate approach than iterating over advertisers. For every coverage pattern, the algorithm loops on the advertiser list to decide which advertisers should the coverage pattern be allocated to. The decision whether an advertiser would be matched to a coverage pattern is taken according to the ϵ value which determines the upper bound on the amount of relaxation with respect to the number of impressions required by the advertiser. The decision whether a coverage pattern is to be assigned to an advertiser or not is determined by the following

$$Ad_i.imps - cp.cvg < \epsilon \times Ad_i.imps \tag{4.15}$$

If the difference between the number of impressions required by an advertiser, $Ad_i.imps$ and the provided coverage of the coverage pattern, $cp.cvg$ is lesser than the product of ϵ and impressions required, $Ad_i.imps$, then the coverage pattern is allocated to the respective advertiser. With each coverage pattern, looping over entire set of advertisers ensures that the coverage pattern would be exhausted once it has iterated over the entire list. Thus, a better coverage pattern would be completely allocated before the using the next one. If a coverage pattern is assigned to even one advertiser, it is to be noted that any other coverage pattern having any sub-concept from the

Table 4.6: Coverage Pattern and Advertiser Matching

Coverage Pattern	Advertisers
{Agri, Phy}	$\{A_1, A_2, A3\}$
{Bio, Tech}	$\{A_4, A_5, A_6, A_7\}$
{Chem, Env}	$\{A_8, A_9\}$

Table 4.7: Table Showing Allotment Details

Session-Id	Query	Concept	SubConcept	Ad
1	snail reproduction	Science	Biology	A5
1	snail shells	Science	Biology	A4
1	chemical composition of snail shell	Science	Chemistry	A8
2	program for einstien equation	Science	Technology	A5
2	e mc2	Science	Physics	A3
3	heat energy reactions	Science	Chemistry	A8
4	computer science internships	Science	Technology	A5
4	ipod price	Science	Technology	A4
5	god particle	Science	Physics	A3

allotted coverage patterns can't be used. The last part of the algorithm removes all the coverage patterns whose any sub-concept has been allotted during the current iteration. The algorithm terminates when all the advertisers have been either assigned or if the algorithm has iterated over all the patterns. In case, if any advertiser is still left to be allocated a coverage pattern, we assign the sub-concepts that have not been assigned to any other advertisers, SCs_left in the form of a single pattern.

Continuing example 2, the best coverage pattern in Table 4.4 will be used first - {Agri, Phy}. The pattern will iterate through the advertisers to see if the number of impressions needed can be satisfied by the coverage pattern. We use the value of ϵ to be 0.1 which means that the relaxation upper bound is 10% of the impressions required by the advertiser. It is backed up by the reasoning that 10% relaxation in the upper bound would not cost anything substantial to the publisher because the total number of impressions is quite large.

The pattern {*Agri, Phy*} was exhausted and matched to A_1, A_2 and A_3 . All the patterns containing either {*Agri*} or {*Phy*} need to be removed and hence, the coverage pattern {*Agri, Env*} was removed. Similarly, the next coverage pattern is matched to advertisers and the process is continued until all the advertisers are matched to a coverage pattern.

In the rest of the section, we will show how advertisements are displayed when a query is fired, given the advertisers and coverage patterns matching as shown in Table 4.6. For each incoming query, the concept and sub-concept pair are identified using the concept taxonomy proposed earlier. From the coverage patterns and advertiser matching, the advertisers matched to the coverage pattern containing the query's sub-concept are retrieved.

Figure 4.6: Algorithm to Match Coverage Patterns and Advertisements

Input: Advertisers List(Ad_list), Coverage Patterns List(CP_list), Sub-Concepts List (SCs), Greedy Parameter(ϵ)
Output: List of Advertisers Allotted Coverage Patterns($Allotted_Ads$)
Method:

$Alloted\_Ads \leftarrow \{\}$
$SCs\_alloted \leftarrow \{\}$
for each CP in CP_sorted **do**
  $num\_of\_ads = len(Ad\_List)$
  if $(len(Allotted\_Ads) == num\_of\_ads)$ **then**
    $break$
  end if
  $flag \leftarrow False$
  for each Ad in Ad_list **do**
    if $(Ad.imps - CP.cvg < \epsilon \times Ad.imps)$ **then**
      $flag \leftarrow True$
      $Ad.CP \leftarrow CP$
      $SCs\_alloted \leftarrow (SCs\_alloted \cup CP.SCs)$
      $CP.cvg \leftarrow (CP.cvg - Ad.imps)$
      $Alloted\_Ads \leftarrow Alloted\_Ads + Ad$
      $Ad\_list.remove(ad)$
    end if
  end for
  $temp \leftarrow CP\_List$
  if $(flag)$ **then**
    for each $pattern$ in $temp$ **do**
      if $(pattern.SCs \cap CP.SCs \neq \phi)$ **then**
        $CP\_list.remove(pattern)$
      end if
    end for
  end if
end for
if $(len(Ad\_list) > 0)$ **then**
  $SCs\_left \leftarrow SCs - SCs\_alloted$
  for each Ad in Ad_list **do**
    $Alloted\_Ads \leftarrow Alloted\_Ads + Ad$
    $Ad.CP \leftarrow SCs\_left$
  end for
end if
Output $Alloted\_Ads$.

In sponsored search advertising, ranking is done based on the scaled bid of the advertiser. Scaling of the bid is achieved by multiplying the actual bid with *Quality Score*. As mentioned earlier, *Quality Score* is composed of several factors that influence the optimality of the query and advertisement matching. For the sake of this study, we assume *Quality Score* to be equal to product of *Remaining Budget* of the advertiser with other factors like quality of landing page, relevance of ad to the query, etc. In this study, we consider *Quality Score* to be the *Remaining Budget* with a constraint that the same ad will not be repeated in the same web search session.

For the continuing example, lets assume five sessions for illustration purposes as shown in Table 4.7. For each query that is fired, it is classified into a concept and sub-concept pair according to the hierarchy. For the first query, it is $\{$*Science, Biology*$\}$ and the corresponding coverage pattern is $\{$Bio, Tech$\}$ which has been alloted to advertisers A_4, A_5, A_6 and A_7. Since this is the start of the session, no ads have been shown. So, *NRF* is one for all the advertisers and *Quality Score* is equal to the budget because no clicks have occurred. For the first query the ranking of the advertisers is as follows $A_5 > A_4 > A_6 > A_7$ and hence, A_5 is displayed along with query results. The next query also has the same pair of concept and sub-concept extracted from the hierarchy i.e. $\{$Science, Biology$\}$. However, now since A_5 has been displayed in the same session earlier, the value of NRF for A_5 would become zero and making the rankings as follows $A_4 > A_6 > A_7 > A_5$ and A_4 would be displayed. This process would be followed queries come in. Table 4.7 shows a complete example of five sessions of how allotment is done.

It is to be noted that the contribution of the proposed approach is primarily towards allocation of advertisements to queries. Through the above explanation, we show how the proposed approach is independent of the display measures such as *Quality Score*. Therefore, it can be easily integrated with the existing system.

4.4 Experiments

In this section we compare the performance of the proposed approach with the framework of sponsored search. We explain dataset, methodology, performance metrics and results.

4.4.1 Dataset

We used the CABS120k08 [58] data corpus which is a collection of search queries from the AOL500k dataset, documents clicked, document rank, timestamps and user id. The CABS120k08 dataset models the web document as a unit. The data set also contains category, rank along with the relevant search queries for the document.

From the dataset, we extracted all the queries in the form:

$$< query,\ user - id,\ timestamp,\ category\ hierachy >$$

Table 4.8: Dataset Statistics

Category	No. of Sub-concepts	Sessions	Queries	Queries Per Session
Arts	15	7,107	15,317	2.15
Health	12	9,181	26,385	2.87
Society	16	6,471	13,223	2.04
Shopping	19	14,819	40,463	2.73
Total	62	37,578	95,388	2.54

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 based on the standard 30 minutes rule of the four popular categories – *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.

4.4.2 Implementation Methodology

In this section, we explain the methodology used to implement Sponsored Search and proposed approach.

4.4.2.1 Sponsored Search system

For implementation, we needed advertisers data containing bids according to queries dataset. Existing bidding datasets are either anonymized or have a lot of keywords which are not present in the dataset. Hence, we had to create the advertisers dataset. We extracted keywords from the query logs and lemmatized them using the NLTK library [54]. Relative frequency of the lemmatized words was calculated and a maximum bid of $10.00 was set for the most frequent keyword. Minimum bids for all the other keywords was then computed as a function of their frequency with respect to the frequency of the most occurring keyword.

$$MinBid(keyword) = \frac{Frequency(keyword)}{MaxFreqency} \times 10.00 \tag{4.16}$$

To select keywords for the advertisers, we developed a cumulative frequency index on the keywords. For each advertiser, a random number is selected with the total number of keywords as the upper bound. The keyword corresponding to that number is then extracted using the cumulative frequency index. The index plays an important role by ensuring that a keyword that has more frequency has more chances of getting selected. Once the first keyword is selected, similar keywords from the dataset are selected based on Wu-Palmer [78] within a threshold value of 0.80. Wu-Palmer similarity is calculated in accordance to the taxonomic depth of the two words.

Table 4.9: Advertisers Dataset Statistics

Category	Avg Budget	Max Budget	Min Budget	Avg CTR	Max CTR	Min CTR
Arts	$30.58	$42.18	$21.70	6.89	6.98	6.7
Health	$20.79	$35.07	$3.39	6.85	6.96	6.52
Society	$39.77	$68.65	$11.16	6.86	7.0	6.4
Shopping	$8.06	$10.85	$6.37	6.86	7.0	6.63
Overall	$29.75	$79.56	$3.39	6.85	7.0	6.35

Another key point that should be mentioned here is the number of advertisers in the dataset. For experiments, twenty sets of advertisers were considered, five for each category with number of advertisers as $10, 20, 30, 40$ and 50. The upper bound of 50 for each category was decided after analyzing the web query logs such that we are able to meet the boundary condition of satisfying at least 90% of the total advertisers to achieve a comparison of both approaches. For the creation of dataset, the value of daily budget and CTR is decided at random. The value of CTR was kept between 6.0% to 7.0% while for daily budget it was $0.00 to $100.00. The reason for keeping CTR higher than usual is validate experiments from the perspective of budget exhaustion. The analysis of the created dataset is shown in Table 4.9.

To simulate sponsored search, we used the existing query logs with the generated advertiser dataset. The session boundary for the experiments has been set according to the 30 minutes difference rule [19]. For each query, we iterate through all the advertisers in the dataset and again compute Wu-Palmer similarity within the threshold of 0.90. The bid for each keyword for all the candidate advertisers is generated by adding a random value between 0.00 to half of the minimum bid of the keyword. The value of *Quality Score* for the sake of experiments in this study has been kept as a product of *Remaining Budget* and a NRF. It is to be noted here that we simulated the Pay Per Click model using CTR such that for every X impressions, we assumed there would be one click where the value of X is as follows

$$X = ceil(100.0/int(ceil(CTR))) \tag{4.17}$$

4.4.2.2 Proposed Approach

For the proposed approach, we keep a middle layer of coverage patterns between incoming queries and advertisers. We used the same advertisement dataset from the previous experiment with a modification that the bid on a concept is the average of the bid on the keywords. For the simulation purposes, we scaled the bid according the same *Quality Score* and used the same concept of number of impressions and CTR as in previous experiment to ensure clicks in the advertisements.

4.4.3 Performance Metrics

We have employed two performance metrics, one is to evaluate the efficiency to cover more advertisers and another is to evaluate the diversity aspect.

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)*. It can be observed that high AS indicates more utilization of a session, which in turn indicates better utilization of ad space.

$$AS = \frac{SUAS}{NSA} \tag{4.18}$$

To evaluate the diversity, we want to evaluate the number of sessions allotted to each advertisement. We calculate the value of Sessions per Advertisement (SA) which is the ratio of *Number of Advertisements of all Sessions (NAS)* to *Number of Advertisements (NA)*. Higher value of metric implies more number of unique eye balls and thus, increasing the chances of the advertisement being clicked by *diverse* users.

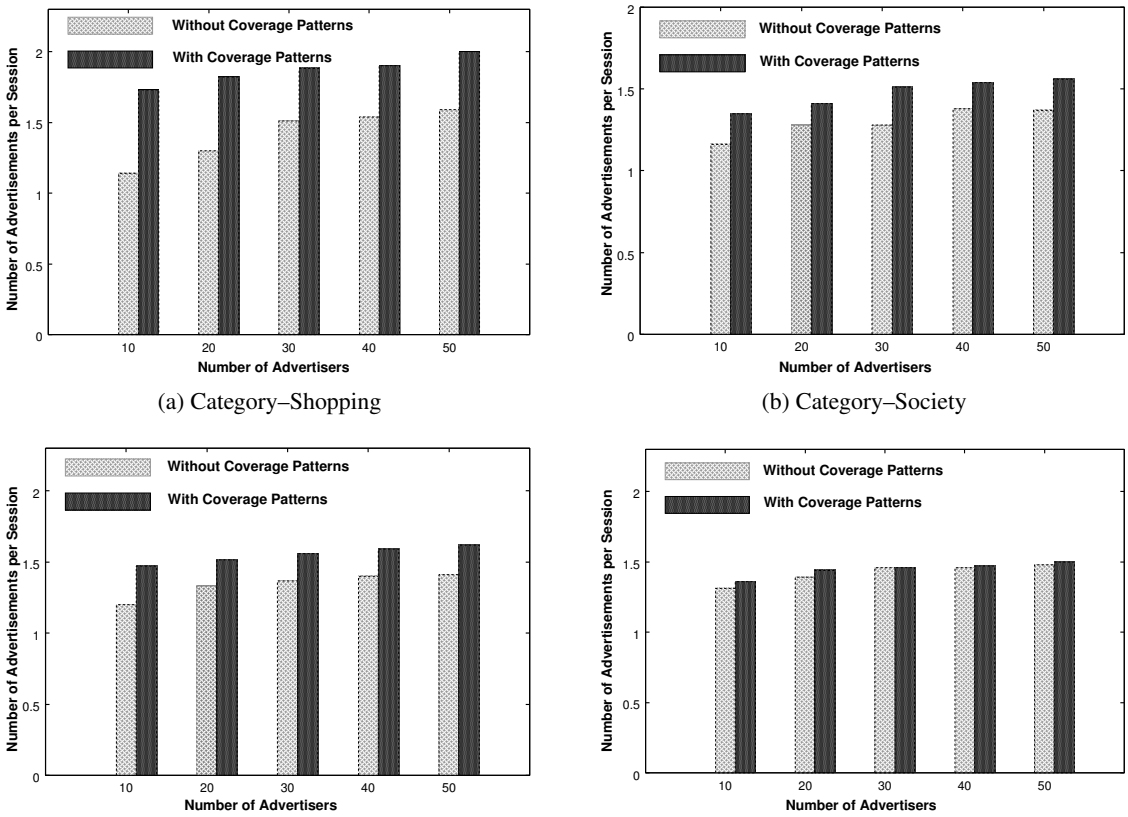
$$SA = \frac{NAS}{NA} \tag{4.19}$$

4.4.4 Results

Figure 4.7 show a comparison of sponsored search architecture with and without coverage patterns. It can be observed that there is a significant improvement in the performance of AS for each category. For *Shopping, Arts, Society* and *Health*, with respect to utilization of sessions, an increase of 33.08%, 15.81%, 14% and 1.89% was realized with respect to utilization of sessions. Overall, we observed an increase of 16.20% in the utilization of sessions. This is due to the fact that the allocation of advertisements with the proposed approach is able to cover more advertisers by grouping the infrequent keywords. It can be observed that the average session length is 2.5 queries. Even though the session length is small, the experiment results show that the knowledge of coverage patterns is able to improve the performance. We expect that in class of applications with large session length, the knowledge of coverage patterns could improve the performance significantly.

It can be also observed that the performance of the proposed approach under all categories is not the same. Especially, for *Health* category there is no significant performance improvement. This is due to the fact that the coverage patters extracted for this category have high coverage support and high overlap ratio. Therefore, the number of unique eyeballs is relatively less compared to other categories.

Figure 4.8 shows the performance of two approaches regarding diversity. It can be observed that there is a significant improvement in performance of SA for each category. In the category of *Shopping* an increase of 26.13% in the diversity was noted while for the categories of *Arts*, *Society* and *Health*, it was 6.62% , 21.98% and 14.85% respectively. On an average, we noticed an increase of 17.39% in



(a) Category–Shopping

(b) Category–Society

(c) Category–Arts

(d) Category–health

Figure 4.7: Performance with Respect to Coverage of Advertisers

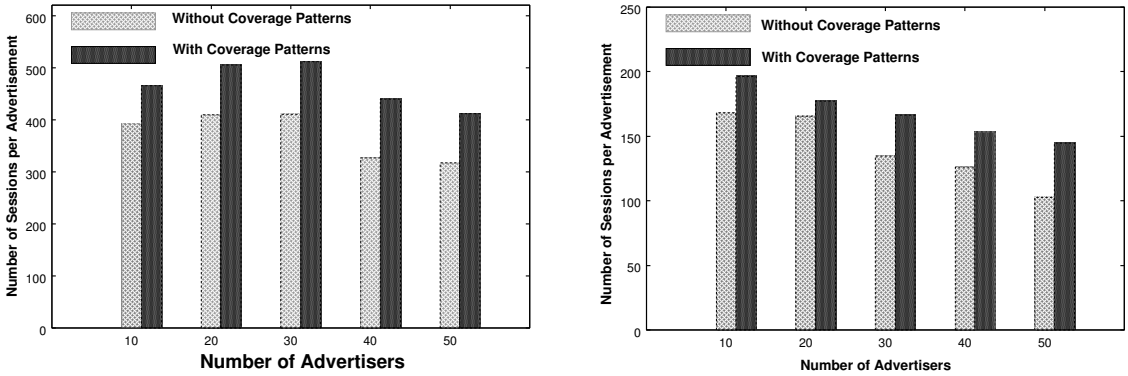
the diversity in the proposed approach. The variation in the diversity performance occurs due to search behavioral patterns pertaining to each category.

4.5 Discussion: Assumptions and Limitations

The proposed approach has the following assumptions and limitations:

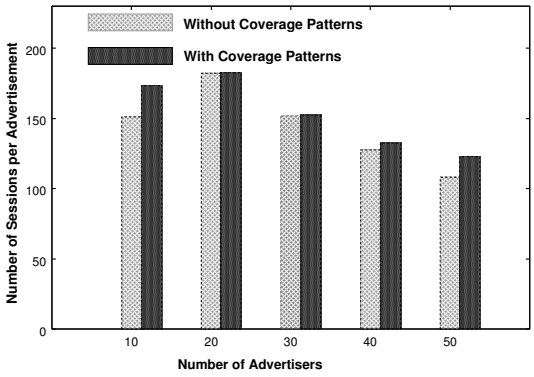
(i). We have assumed that bidding on concepts is more natural compared to bidding on keywords as concepts communicate integrated ideas which is not the case with individual keywords. Our motivation to bid on concepts also comes from social media advertising where advertisers select demographic concepts like photography, reading, travelling, lifestyle, etc. The approach of bidding on concepts has been quite successful in social media advertising.

(ii). It is assumed that the transactions formed from web query logs can be used to identify the set of concepts that cover a given percentage of visitor population. Such knowledge could be used to

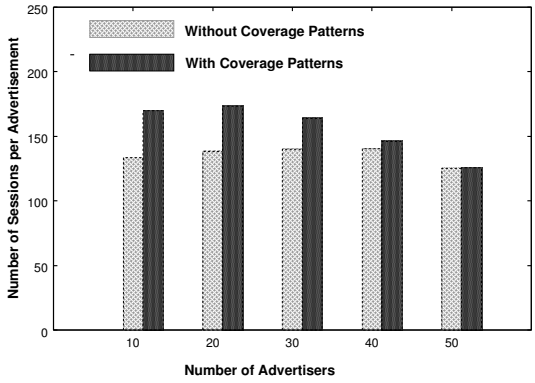


(a) Category–Shopping

(b) Category–Society



(c) Category–Arts



(d) Category–Health

Figure 4.8: Performance with respect to Diversity

place the advertisement assuming similar search behavior. The related issues will be investigated as a part of future work.

(iii). We have also made the assumption that each query and ad pair within the same high level concept are relevant to each other. Overall, this assumption may be true but more analysis is needed on how to do targeted advertising similar to the approach of bidding on keywords. Moreover, we also assume that if an advertiser is bidding on a high level node in the taxonomy, he/she is interested in showing his/her ad on a query which belongs to any descendant of the bidding node.

(iv). There is a need to understand the interplay between more relevant queries and increase in reach of audiences. More reach of an ad might also mean that ad is being shown to irrelevant users, thereby compromising the user experience and brand perception.

(v). Bidding is only performed on the first level of the taxonomy. This might not be suitable for all the advertisers as the advertising requirements for each advertiser is different.

4.6 Summary

In this chapter, we have discussed the first approach for long tail advertising in sponsored search. We proposed that advertisers should bid upon high level concepts instead of individual keywords in ad space auctions. Using session based transactions, we extracted coverage patterns from search query logs and performed a matching between extracted coverage patterns and advertising demands. We further proposed an end-to-end framework to use this matching to allocate ads to incoming search queries. Experimental results on a real world dataset also showed improvement with respect to utilization of ad space and reach of advertisements.

Chapter 5

Exploiting ad space of tail queries using a multi-level taxonomy and coverage patterns

In this chapter, we extend the notion of bidding on concepts from a two level taxonomy to a multi-level taxonomy. We propose the notion of *level-wise coverage patterns* to extract coverage patterns when a taxonomy is defined over the itemset. The notion of *level-wise coverage patterns* and taxonomy based bidding has been used to propose an end-to-end framework to allocate ads to incoming search queries for sponsored search.

5.1 Motivation and Basic Idea

Search queries follow a long tail distribution with a small head of frequent (head) queries and a long tail of infrequent queries. Advertising on tail queries is hard as they occur rarely. Moreover, during ad space space auctions, advertisers are biased towards targeting head query keywords because of the individual high reach of head queries. Also, it is quite difficult for an advertiser to identify all relevant keywords from the long tail. The stated factors lead to under-utilization of ad space of tail queries which is identified as the research issue in this thesis.

In the previous chapter, we have a proposed an approach to exploit the ad space of tail keywords. We have proposed that advertisers should bid upon concepts instead of keywords in ad space auctions. We proposed to use a two level taxonomy for bidding in ad space auctions where advertisers could only bid on the first level. Through experiments, we observed that bidding on concepts shows promise with respect to ad space utilization for sponsored search. The previous approach is easy to put in practice, however, bidding on only the first level of taxonomy may not be sufficient for certain advertisers and more flexibility may be desired for certain advertising demands.

In this chapter, we propose an approach of bidding on concepts in a multi-level taxonomy instead of keywords or a two level taxonomy. During the ad space auctions, an advertiser is shown a taxonomy based on the content of his/her ad. The advertiser is then asked to select a node in the taxonomy which he/she deems the most relevant for his/her product. For example, an advertiser like *Amazon.com* would

be shown at taxonomy of *Shopping* and based on the advertising requirements, the advertiser can select the appropriate node. If the advertisement is related to 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, the approach of bidding on a multi-level taxonomy gives more flexibility to the advertisers to target potential consumers compared to the first proposed approach.

Using the notion of bidding on a node in the taxonomy, we propose an allocation model 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, the notion of coverage patterns has been employed. Coverage patterns containing children nodes of a bidden node will help in identifying mutually exclusive sets of concepts which ensure a minimum coverage and also satisfy the requirement of a maximum overlap amongst the items in the CP, thereby, reducing repetition of ad slots in the queries belonging to the concepts of the same patterns. Using the extracted coverage patterns from the query logs, a matching is performed between the coverage patterns at each node and the corresponding advertisers.

However, in the literature the approaches proposed to extract coverage patterns [70] consider a flat transaction model which cannot be used to extract coverage patterns in this approach. Hence, we propose the notion of 'level-wise' coverage pattern extraction when a hierarchy is present over the itemset. Using this notion of 'level-wise' coverage patterns and bidding on multi-level taxonomy, we propose an end-to-end architecture to allocate ads to incoming queries.

In the next Section, we discuss the algorithm to extract 'level-wise' coverage patterns followed by the proposed framework in Section 5.3, experiments in Section 5.4 and a discussion in Section 5.5 followed by a comparision of the two proposed approaches in Section 5.6. Section 5.7 summarizes the chapter.

5.2 Proposed Model

In this section, we will briefly discuss the proposed model. Compared to the standard approach of keywords based allocation in sponsored search, we have proposed to add a middle layer of coverage patterns between the advertisers and search queries. When a query is posed by a user to the search engine, it is first classified into the taxonomy nodes. The advertisers who bid upon the corresponding nodes are considered as candidate advertisers for the query's results page. Based on their bid, query-ad relevance, CTR and other parameters, the advertisers are ranked and displayed on the results page.

To extract the coverage patterns, the existing model of coverage pattern extraction proposed in the literature [70] cannot be employed due to the interdependence of parent-child nodes on each other. For example, if we consider a node *Shopping* and its children *Books,Fashion* and *Electronics* then all the transactions of *Books,Fashion* and *Electronics* will also belong to *Shopping*. Therefore, the coverage of

a parent node in a transactional dataset is equal to number of transactions of its children nodes which can be stated by the following recursive relation,

$$C(N_i) = n(\bigcup(T(N_{i-1}) \mid \forall\, N_{i-1} :\ parent(N_{i-1}) = N_i)) \tag{5.1}$$

where $C(N_i)$ is the coverage of node N_i. The coverage of $C(N_i)$ is equal to the number of transactions of all children which is denoted by union of the transactions of the children. Equation 5.1 shows that there is a flow of coverage from the root node to the leaf nodes in the taxonomy. Hence, the flat model of extraction of coverage patterns cannot be employed here.

5.2.1 T-Cmine: An approach to extract 'level-wise' coverage patterns

In the proposed approach, since each session transaction contains multi-level items (query and all the taxonomy nodes related to it), the flat transaction model of coverage patterns cannot be used to extract coverage patters. In this section, we propose an approach that takes a dataset of transactions and a taxonomy defined over the items and outputs 'level-wise' coverage patterns.

In [67], 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 [70] (also discussed in Chapter 3). 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.

$$CP = \{c \mid c \in (\mathcal{I} \cup \mathcal{T})\ \&\ \forall\, c\ parent(c) = P\} \tag{5.2}$$

Here, a *level wise coverage pattern* 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.

The proposed algorithm takes a dataset $\mathcal{D}$ and a taxonomy $\mathcal{T}$ that defines the relationship amongst the items of $\mathcal{D}$. The algorithm first adds ancestors of each item in a transaction to the transaction. Then, the first set of coverage patterns (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 in Equation 5.2. 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 coverage patterns which don't satisfy maxOR criteria.

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

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 coverage pattern, 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 coverage patterns 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 [70], 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 Figure 5.1.

5.3 Proposed framework

In this section, we discuss the proposed model and the proposed framework as shown in Figure 5.2. We propose that in the ad space auctions, advertisers should bid upon a taxonomy instead of individual keywords. An advertiser is free to bid on any node in the taxonomy and a set of children nodes of the bidding node is allocated to the advertiser in the form of level-wise coverage patterns. When a user poses a query to the search engine, it is first classified by the bidding 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 a coverage pattern containing any of these concepts would be considered to be displayed on the results page of the query - *Harry Potter*. Thus, as compared to the standard sponsored search model of a bipartite graph between advertisers and queries as shown in Figure 3.1, we add a middle layer of coverage patterns between search queries and advertisers similar to the first approach, as shown in Figure 5.2 (a).

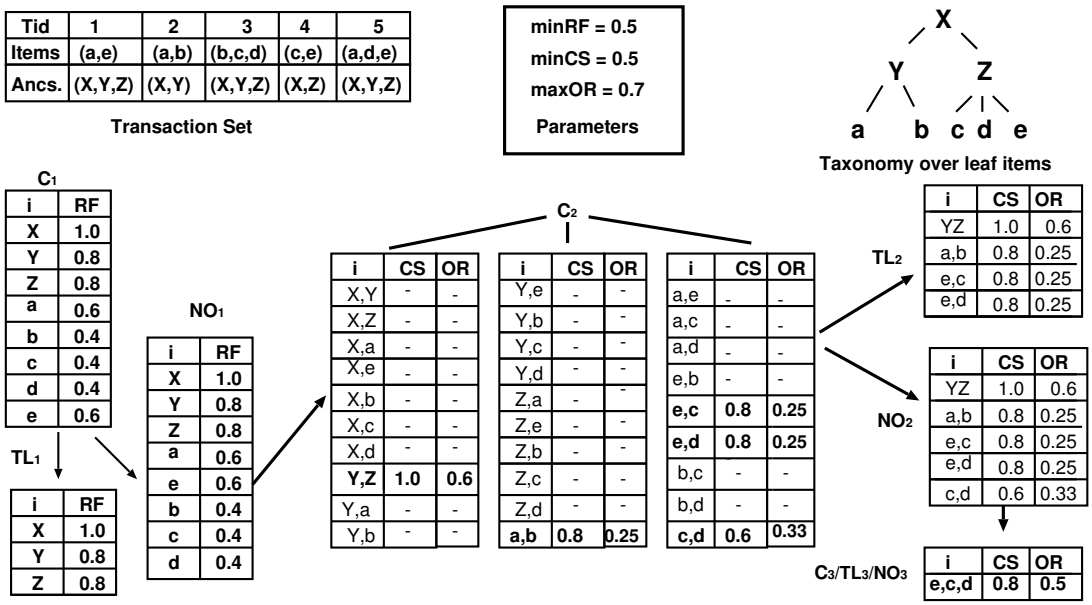


Figure 5.1: Example 2: Example of T-Cmine

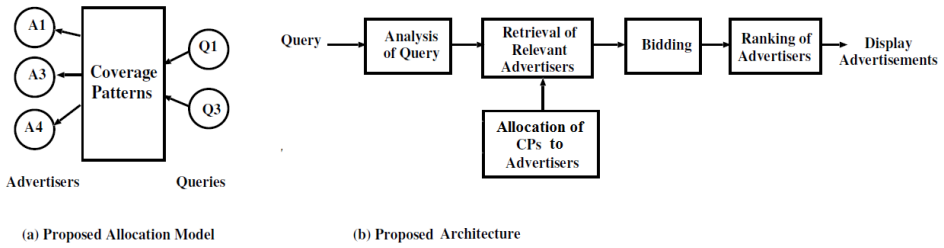


Figure 5.2: Proposed Sponsored Search Model and Architecture

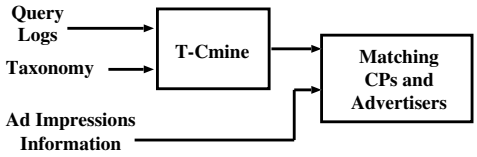


Figure 5.3: Allocation of concepts to advertisers

The sponsored search architecture has four major steps for query allocation to advertisers as discussed in Chapter 3. 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 coverage patterns. We discuss each step of the proposed architecture as follows.

(i). ***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*.

(ii). ***Retrieval of Relevant Advertisers***: The input to this step is the incoming search query, its classification into the taxonomy nodes, and a matching between coverage patterns and advertisers. The matching is the output from the component where allocation of concepts to advertisers is achieved. In the next subsection, we will discuss the allocation mechanism in detail.

(iii). ***Bidding***: Ad space on each search query is sold by means of auctions. Each advertiser bids for the ad slots. This bid can be static or dynamic. (This step is the same as the standard sponsored search approach.)

(iv). ***Ranking of Advertisers***: Based on the bid amount and other relevant parameters like CTR and ad-query relevance, advertisers are ranked to be shown on the query's results page. (This step is also same as the standard sponsored search approach.)

5.3.1 Allocation of concepts to advertisers

In this section, we explain how concepts are allocated to advertisers. 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 as shown in Chapter 4. The allocation process process has two main components as shown in Figure 5.3:

(i). **Extraction of coverage patterns using T-Cmine**: This step takes query logs and taxonomy as input and extracts coverage patterns as explained in Section 5.2.1.

(ii). **Matching coverage patterns and Advertisers**: In this step, we take the demands of the advertisers and the coverage patterns extracted from query logs and perform a matching between the

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

Figure 5.4: Example Allocation

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 have 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 coverage pattern if any of its descendants have been allocated to advertisers is to subtract the number of impressions allotted to the advertisers of the children of nodes contained in the respective coverage pattern. Equation 5.3 captures the necessary changes required to a coverage pattern such that for each node in the coverage pattern (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 coverage pattern. 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.

$$CP.imp = CP.imp - \sum_{k} \sum_{ij} A_{ij} \tag{5.3}$$

Example 3: In Figure 5.4, 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 coverage patterns pertaining to *Shopping* are shown in Table 1 of Fig 5.4. 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 coverage patterns containing these two nodes. For example, the coverage pattern {*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 1000 i.e. 1500 - (200 + 300) = 1000. In the next part of this section, the matching between coverage patterns is performed considering the proposed modification.

A matching is performed with advertisers as one side of the bipartite and coverage patterns 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 coverage patterns and advertisers allocated to them should be minimal. For example, if an advertiser demands 100 impressions and there are two coverage patterns with impressions 150 and 200 respectively, then we chose to allocate the coverage pattern with 150 impressions. A similar case can be made when the supply of coverage patterns is 50 and 75 impressions and demand by the advertisers is 100 impressions, then the coverage pattern with 75 impressions is chosen. We frame the objective function of the matching on the same notion which is as follows. Equation 5.4a aims at minimizing the difference between the allocated advertising and the coverage patterns. 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 5.4b lays out the constraint such that the sum of impressions of allocated advertisers does not exceed the impression provided by the coverage pattern to avoid the objective from going negative.

$$Min\ Z = \sum_{level=d}^{0} (\sum_{j} |CP_j.Impressions - \sum_{i}^{n} (Ad_{ij}.Impressions)|) \tag{5.4a}$$

$$s.t\ \ CP_j.Impressions >= \sum_{i=1}^{n} (Ad_{ij}.Impressions) \tag{5.4b}$$

Continuing Example 3 from Fig 5.4. From the last step, we have coverage patterns 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 coverage pattern {*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 coverage pattern {*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 coverage pattern, 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 coverage patterns and advertisers will be one-to-many as the number of impressions that can be covered by a coverage pattern is quite high 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 Figure 5.2, it will be first classified according to the taxonomy as *Shopping; Books; Fiction*. Advertisers who have been allotted a coverage pattern 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.)

5.4 Experiments

5.4.1 Dataset

For the experiments, we used the CABS120k08 [58] 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 as a transaction to extract coverage patterns by T-Cmine as sessions form the logical boundary of searching. Table 5.1 shows the statistics of the extracted dataset.

Table 5.1: 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

5.4.2 Implementation Methodology

The standard sponsored search approach mentioned in [56] 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.

5.4.3 Performance Metrics:

Two performance metrics have been employed to compare the keywords based approach [56] and the proposed concept-based bidding approach.

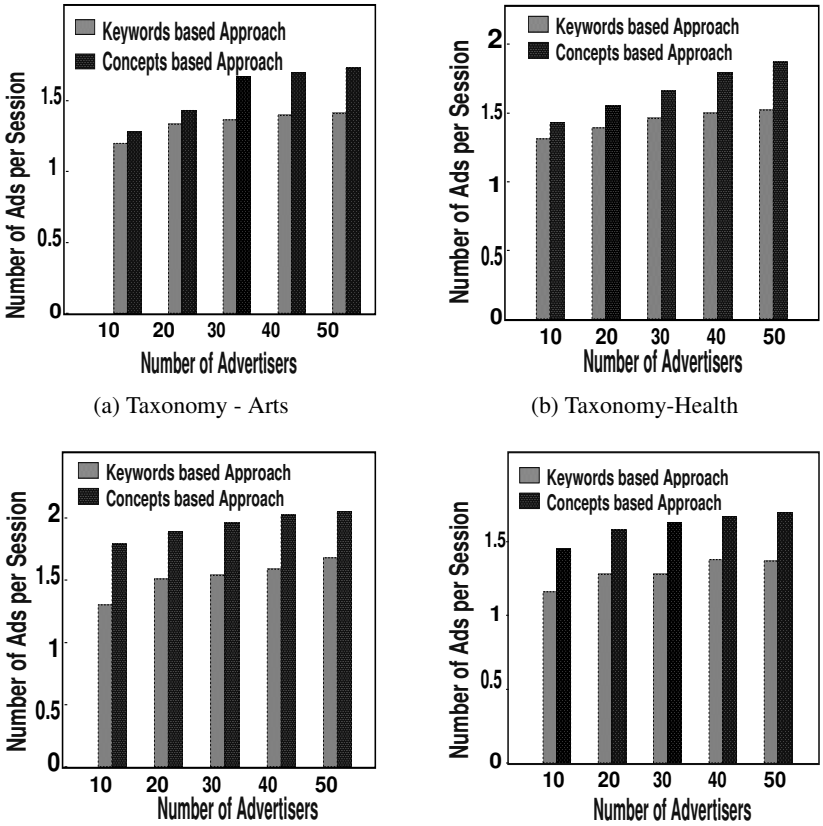
To evaluate the utilization of ad space, we calculate the average number of unique Advertisements per Session (AS). It is calculated as the ratio of *Sum of Unique Advertisements of all Sessions (SUAS)* and *Number of Sessions with Advertisements (NSA)*. High value of AS indicates more utilization of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

$$AS = \frac{SUAS}{NSA} \tag{5.5}$$

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.

$$SA = \frac{NAS}{NA} \tag{5.6}$$



(a) Taxonomy - Arts

(b) Taxonomy-Health

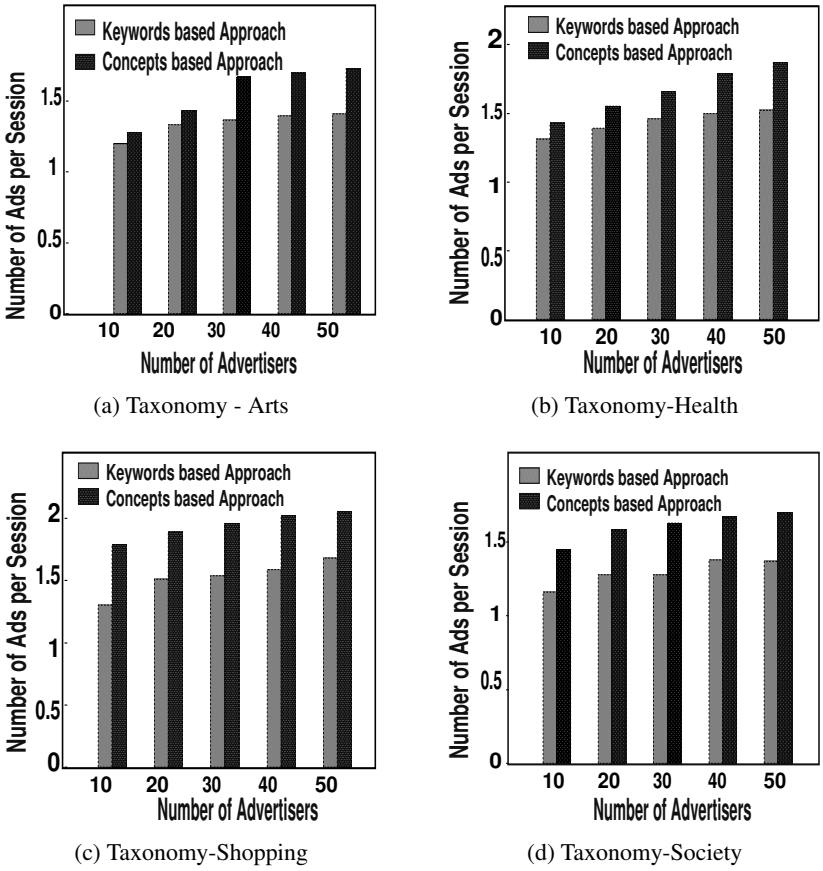
(c) Taxonomy-Shopping

(d) Taxonomy-Society

Figure 5.5: Performance with respect to utilization of ad space

5.4.4 Results

Figure 5.5 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



(a) Taxonomy - Arts

(b) Taxonomy-Health

(c) Taxonomy-Shopping

(d) Taxonomy-Society

Figure 5.6: Performance with respect to reach of advertisers

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.

Figure 5.6 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.

5.5 Discussion: Assumptions and Limitations

The proposed approach has the following assumptions and limitations:

(i). We have used an external taxonomy for the experiments and we have assumed that this taxonomy is adequate enough to express the advertising requirements of the advertisers. We have also assumed that the taxonomy is appropriate enough for translating search queries into logical concepts to efficiently match advertisements to incoming queries.

(ii). Another assumption that we have made is that bidding on a taxonomy would be more natural than bidding on keywords. However, this needs more investigation if an advertiser would find it convenient to bid on a taxonomy as a taxonomy could very broad or could be very deep.

(iii). On the coverage pattern front, we have assumed the distribution of taxonomy nodes for future search queries will remain the same or change less, and hence, coverage patterns extracted today can be used tomorrow.

(iv). The proposed approach is also not equipped to handle query ambiguity. If the search query is ambiguous, it will be classified into multiple orthogonal taxonomy nodes and the proposed is not capable of handling that. For example, the query 'Harry Potter' could belong to the node *Movies* as well as *Books*.

(v). We have also assumed keyword distribution within each node of the taxonomy is uniform i.e. no concept is too frequent or too infrequent. This may not be true and taxonomy concepts might *become popular overtime*.

5.6 Comparison of the proposed approaches

Through the study presented in this chapter, we were able to show that it is possible to leverage the ad space of tail queries through extending the notion of bidding on keywords to bidding on a taxonomy.

The approach proposed in this chapter is not quantitatively comparable to the approach covered in the previous chapter as the previous approach uses a flat-concept bidding mechanism and the sub-concepts are not exposed to the advertisers but in the second approach advertisers are free to bid upon any node in the taxonomy. For example, let's consider two advertisers - *Amazon* (an online shopping store) and *Crosswords* (an online book store). In the first approach, both of the advertisers would bid upon Shopping. However, in the second approach, *Amazon* would still bid upon Shopping but *Crossword* would bid on Books. Hence, in the approach proposed in this chapter, the advertisers can choose the desired level of generalization in terms of taxonomy nodes whereas in the previous approach, the assignment is done internally. Therefore, the approaches are not quantitatively comparable.

We discuss a qualitative comparison of both approaches in this section. Table 5.2 states the key differences between the two proposed approaches.

The first approach uses a two level taxonomy whereas the second approach a multi-level taxonomy for ad space auctions.

In first approach, advertisers are only allowed to bid on the first level of the taxonomy. The second approach gives flexibility to the advertisers to bid on any node in the taxonomy and target different levels generalizations as required.

The third key difference is with respect to pricing models. A pricing model is a mechanism to decide the minimum bid amount for a given ad slot. Pricing models would remain the same or have less modifications in the first approach as each advertiser is bidding on the same level of the taxonomy. However,

Table 5.2: Comparison of the proposed approaches

Feature	Approach 1	Approach 2
Taxonomy Structure	Flat high level concepts are exposed to advertisers for bidding	Entire taxonomy of concepts is exposed to advertisers
Flexibility	Less flexible with respect to targeting potential customers	More flexible with respect to targeting potential customers
Pricing Models	Same pricing model could be used as all advertisers are bidding on the same level	New pricing models need to be defined to address bidding at different levels
Truthful Auctions	Truthful auctions can be proven easily as bidding is done on the same level	Truthful auctions have to be analyzed and modelled as bidding is allowed at different granularity levels
Complexity	Easy to put in practice	Relatively difficult to put in practice

in the second approach advertisers can bid upon any level in the taxonomy, and thus, advertisers would bid upon different levels in the taxonomy for the same query's ad space. For example, for a query like *fiction books*, advertisers who bid upon *Shopping* and advertisers who chose to bid upon *Books*, which is a child of *Shopping*, are eligible to be shown on the results page for the query. So, pricing models are required to determine how the advertisers should be charged for a given ad slot when the bids are spread across the taxonomy.

The fourth difference in the proposed approaches is that of truthful auctions. A truthful auction aims to encourage its bidders to show their true valuations of the commodity during bidding. In the first approach, truthful auctions can be easily implemented as all advertisers are bidding at the same level of the taxonomy and all queries will be translated to the same concept for each advertiser. In the second approach, advertisers can bid on any level of the taxonomy, so it becomes crucial to make sure that the bidding is truthful. For example, if lower level nodes are generally cheaper according to the pricing model then an advertiser might bid upon several small lower nodes instead of bidding on the more relevant higher level to optimize his/her revenue.

Overall, the approach proposed in this chapter provides more flexibility to the advertisers at the cost of more complexity with respect to pricing model, auctions and allocation of advertisers to incoming queries.

5.7 Summary

In this chapter, we extend the previous approach of bidding on concepts instead of keywords for the ad space auctions. We proposed to generalize the previous approach from a two level taxonomy

to a multi-level taxonomy. We also proposed the notion of 'level-wise' coverage patterns in order to extract coverage patterns when a taxonomy is defined over the itemset. Using bidding on multi-level taxonomy and the notion of level-wise coverage patterns, we formulate an end-to-end framework for allocating ads to incoming search queries. Experiments on AOL search query logs showed improvement in performance with respect to ad space utilization and reach of the advertisements.

Chapter 6

Conclusion and Future Work

In this chapter, we present the final conclusions and give possible directions for future work.

We have investigated the issue of under-utilization of ad space provided by tail keywords in sponsored search. We have proposed two approaches where we suggest that advertisers should bid upon high level concepts instead of keywords in the ad space auctions. In the first approach, we consider that the concepts are organized in a two level taxonomy. We assign groups of concepts to advertisers. These groups are formed by employing coverage patterns on search query transactions. Through experiments on a real world dataset, it can seen that there is an improvement in the ad space utilization and ads are more evenly distributed to the audience. In the second approach, we generalize the taxonomy from two level to multi-level in order to provide more flexibility to the advertisers during bidding on ad slots. We also propose the model of level-wise coverage patterns to assign nodes to advertisers. Experimental results show that the approach proposed allows more flexibility to the advertisers and helps in exploiting more ad space for sponsored search.

Overall, bidding on concepts has shown promise in improving the utilization of ad space of search queries for sponsored search. Since, each concept is composed of a mix of head and tail query keywords, each keyword is considered for advertising based on the relevancy rather than frequency.

The possible directions of future work as are as follows.

In the proposed approaches, external taxonomies have been employed to perform bidding. As a part of future work, we plan on investigating the construction of taxonomy for keyword auctions. We plan to answer if a taxonomy can be constructed from search query logs and bidding phrases of advertisers to suit the bidding requirements.

Coverage patterns have shown promising results with respect to distribution of ads to more unique eye balls. We plan on evaluating other methods of node grouping such as generalized frequent patterns or diverse frequent patterns.

Another important direction of future work is to look at the trade-off of the proposed approaches with respect to targeted advertising. Targeted advertising is a key factor of success in sponsored search and it is important to understand how targeted advertising will work with the bidding on groups of keywords

or bidding on nodes in a taxonomy because the query and the ad might belong to the same high level concept but may as well belong to different topics.

We also plan to look at truthful auctions in the case of taxonomic bidding. Compared the keywords bidding scenario where each advertiser bids on the keywords (lowest level of the taxonomy), in the proposed approaches the advertisers can bid on different levels of generalization in the taxonomy. For example, a large advertiser could bid upon multiple lower levels concepts, if those concepts are cheaper than the high level concept desired by the advertiser. Thus, a key issue is to make sure that an advertiser's bids are truthful while allowing the freedom to bid on any node in the taxonomy.

In the proposed approaches, we have assumed a first-come-first-serve advertiser model. As a part of future work, we intend to explore scheduling mechanisms for incoming advertisers to evaluate different allocation schemes for a stream of incoming advertisers.

On the coverage patterns front, we plan on looking at parallel approaches to extraction of coverage patterns. We also aim to investigate the notion of taxonomic coverage patterns for banner advertising and real time bidding.

Chapter 7

Publications

1. Budhiraja, Amar, and P. Krishna Reddy. "An approach to cover more advertisers in adwords." IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2015. IEEE, 2015. (Supported by Google Research India Student Travel Grants.)

2. Budhiraja, Amar, and P. Krishna Reddy. "An Improved Approach for Long Tail Advertising in Sponsored Search." International Conference on Database Systems for Advanced Applications (DASFAA), 2017. Springer, 2017. (Supported by Microsoft Research India Student Travel Grants.)

Bibliography

[1] Iab internet advertising revenue report. 2015.

[2] V. Abhishek and K. Hosanagar. Keyword generation for search engine advertising using semantic similarity between terms. In *Conference on Electronic commerce*, pages 89–94. ACM, 2007.

[3] G. Aggarwal, A. Goel, and R. Motwani. Truthful auctions for pricing search keywords. In *Conference on Electronic commerce*, pages 1–7. ACM, 2006.

[4] G. Aggarwal, J. Feldman, S. Muthukrishnan, and M. Pál. Sponsored search auctions with markovian users. In *International Workshop on Internet and Network Economics*, pages 621–628. Springer, 2008.

[5] G. Aggarwal, G. Goel, C. Karande, and A. Mehta. Online vertex-weighted bipartite matching and single-bid budgeted allocations. In *Symposium on Discrete Algorithms*, pages 1253–1264. SIAM, 2011.

[6] R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *International Conference on World Wide Web*, pages 13–24. ACM, 2013.

[7] S. Alhabash, J. Mundel, and S. A. Hussain. Social media advertising. *Digital Advertising: Theory and Research*, 2017.

[8] K. Asdemir. Bidding patterns in search engine auctions. In *Workshop on Sponsored Search Auctions*, 2006.

[9] V. Bharadwaj, P. Chen, W. Ma, C. Nagarajan, J. Tomlin, S. Vassilvitskii, E. Vee, and J. Yang. Shale: an efficient algorithm for allocation of guaranteed display advertising. In *International Conference on Knowledge Discovery and Data mining*, pages 1195–1203. ACM, 2012.

[10] V. Bharadwaj, P. Chen, W. Ma, C. Nagarajan, J. Tomlin, S. Vassilvitskii, E. Vee, and J. Yang. Shale: an efficient algorithm for allocation of guaranteed display advertising. In *International Conference on Knowledge Discovery and Data Mining*, pages 1195–1203. ACM, 2012.

[11] A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, D. Metzler, L. Riedel, and J. Yuan. Online expansion of rare queries for sponsored search. In *International Conference on World Wide Web*, pages 511–520. ACM, 2009.

[12] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using web knowledge. In *International Conference on Research and Development in Information Retrieval*, pages 231–238. ACM, 2007.

[13] A. Z. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel. Search advertising using web relevance feedback. In *International Conference on Information and Knowledge Management*, pages 1013–1022. ACM, 2008.

[14] A. Budhiraja and P. K. Reddy. An approach to cover more advertisers in adwords. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE, 2015.

[15] A. Budhiraja and P. K. Reddy. An improved approach for long tail advertising in sponsored search. In *International Conference on Database Systems for Advanced Applications*, pages 169–184. Springer, 2017.

[16] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In *International Conference on Knowledge Discovery and Data Mining*, pages 875–883. ACM, 2008.

[17] M. Cary, A. Das, B. Edelman, I. Giotis, K. Heimerl, A. R. Karlin, C. Mathieu, and M. Schwarz. Greedy bidding strategies for keyword auctions. In *Conference on Electronic commerce*, pages 262–271. ACM, 2007.

[18] S. Chakrabarti. Dynamic personalized pagerank in entity-relation graphs. In *International Conference on World Wide Web*, pages 571–580. ACM, 2007.

[19] M. Chau, X. Fang, and O. R. Liu Sheng. Analysis of the query logs of a web site search engine. *Journal of the American Society for Information Science and Technology*, 56(13):1363–1376, 2005.

[20] Y. Chen, G.-R. Xue, and Y. Yu. Advertising keyword suggestion based on concept hierarchy. In *International Conference on Web Search and Data Mining*, pages 251–260. ACM, 2008.

[21] Y. Chen, G.-R. Xue, and Y. Yu. Advertising keyword suggestion based on concept hierarchy. In *International Conference on Web Search and Data Mining*, pages 251–260. ACM, 2008.

[22] Y. Chen, P. Berkhin, B. Anderson, and N. R. Devanur. Real-time bidding algorithms for performance-based display ad allocation. In *International conference on Knowledge Discovery and Data mining*, pages 1307–1315. ACM, 2011.

[23] Y. Choi, M. Fontoura, E. Gabrilovich, V. Josifovski, M. Mediano, and B. Pang. Using landing pages for sponsored search ad selection. In *International conference on World Wide Web*, pages 251–260. ACM, 2010.

[24] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

[25] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *International Conference on World Wide Web*, pages 227–236. ACM, 2008.

[26] P. W. H. Coopers. Iab internet advertising revenue report, 2014.

[27] B. Edelman and M. Ostrovsky. Strategic bidder behavior in sponsored search auctions. *Decision support systems*, 43(1):192–198, 2007.

[28] J. Feng, H. K. Bhargava, and D. M. Pennock. Implementing sponsored search in web search engines: Computational evaluation of alternative mechanisms. *INFORMS Journal on Computing*, 19(1):137–148, 2007.

[29] A. Frank, A. Asuncion, et al. Uci machine learning repository. 2010.

[30] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal. Using the wisdom of the crowds for keyword generation. In *International Conference on World Wide Web*, pages 61–70. ACM, 2008.

[31] M. R. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified np-complete problems. In *Symposium on Theory of Computing*, pages 47–63. ACM, 1974.

[32] J. Garofalakis, P. Kappos, and D. Mourloukos. Web site optimization using page popularity. *IEEE Internet Computing*, 3(4):22–29, 1999.

[33] N. Gatti, A. Lazaric, and F. Trovò. A truthful learning mechanism for contextual multi-slot sponsored search auctions with externalities. In *Conference on Electronic Commerce*, pages 605–622. ACM, 2012.

[34] A. Ghose and S. Yang. An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55(10):1605–1622, 2009.

[35] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *International Conference on Machine Learning*, pages 13–20, 2010.

[36] Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. In *International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005)*, 2005.

[37] D. Hillard, S. Schroedl, E. Manavoglu, H. Raghavan, and C. Leggetter. Improving ad relevance in sponsored search. In *International Conference on Web Search and Data Mining*, pages 361–370. ACM, 2010.

[38] D. Horowitz and S. D. Kamvar. The anatomy of a large-scale social search engine. In *International Conference on World wide web*, pages 431–440. ACM, 2010.

[39] J. Hu, G. Wang, F. Lochovsky, J.-t. Sun, and Z. Chen. Understanding user's query intent with wikipedia. In *International Conference on World Wide Web*, pages 471–480. ACM, 2009.

[40] Z. Huang, B. Cautis, R. Cheng, and Y. Zheng. Kb-enabled query recommendation for long-tail queries. In *International on Conference on Information and Knowledge Management*, pages 2107–2112. ACM, 2016.

[41] U. IAB. Iab/pwc digital advertising revenue report q1 2016. iab, 2016.

[42] A. H. Jadidinejad and F. Mahmoudi. Advertising keyword suggestion using relevance-based language models from wikipedia rich articles. *Journal of Computer & Robotics*, 7(2):29–35, 2014.

[43] B. J. Jansen and T. Mullen. Sponsored search: an overview of the concept, history, and technology. *International Journal of Electronic Business*, 6(2):114–131, 2008.

[44] B. J. Jansen and S. Schuster. Bidding on the buying funnel for sponsored search and keyword advertising. *Journal of Electronic Commerce Research*, 12(1):1, 2011.

[45] J. Jansen. *Understanding sponsored search: Core elements of keyword advertising*. Cambridge University Press, 2011.

[46] Z. Katona and M. Sarvary. The race for sponsored links: Bidding patterns for search advertising. *Marketing Science*, 29(2):199–215, 2010.

[47] V. N. S. Kavya and P. K. Reddy. Coverage patterns-based approach to allocate advertisement slots for display advertising. In *International Conference on Web Engineering*, pages 152–169. Springer, 2016.

[48] B. Kitts and B. Leblanc. Optimal bidding on keyword auctions. *Electronic Markets*, 14(3):186–201, 2004.

[49] B. Kitts, J. Y. Zhang, G. Wu, W. Brandi, J. Beasley, K. Morrill, J. Ettedgui, S. Siddhartha, H. Yuan, F. Gao, et al. Click fraud detection: adversarial pattern recognition over 5 years at microsoft. In *Real World Data Mining Applications*, pages 181–201. Springer, 2015.

[50] S. Lahaie. An analysis of alternative slot auction designs for sponsored search. In *Conference on Electronic Commerce*, pages 218–227. ACM, 2006.

[51] K.-c. Lee, B. Orten, A. Dasdan, and W. Li. Estimating conversion rate in display advertising from past erformance data. In *International Conference on Knowledge Discovery and Data Mining*, pages 768–776. ACM, 2012.

[52] X. Li, M. Zhang, Y. Liu, S. Ma, Y. Jin, and L. Ru. Search engine click spam detection based on bipartite graph propagation. In *International Conference on Web search and data mining*, pages 93–102. ACM, 2014.

[53] W.-l. Lin and Y.-z. Liu. A novel website structure optimization model for more effective web navigation. In *Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on*, pages 36–41. IEEE, 2008.

[54] E. Loper and S. Bird. Natural language toolkit. 2005.

[55] A. Mehta. Online matching and ad allocation. *Theoretical Computer Science*, 8(4):265–368, 2012.

[56] A. Mehta, A. Saberi, U. Vazirani, and V. Vazirani. Adwords and generalized online matching. *Journal of the ACM*, 54(5):22, 2007.

[57] A. Mladenow, N. M. Novak, and C. Strauss. Online ad-fraud in search engine advertising campaigns. In *Information and Communication Technology-EurAsia Conference*, pages 109–118. Springer, 2015.

[58] M. G. Noll and C. Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *Web Intelligence and Intelligent Agent Technology*, volume 1, pages 640–647. IEEE, 2008.

[59] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175, 2013.

[60] C. Perlich, B. Dalessandro, R. Hook, O. Stitelman, T. Raeder, and F. Provost. Bid optimizing and inventory scoring in targeted online advertising. In *International Conference on Knowledge Discovery and Data Mining*, pages 804–812, 2012.

[61] H. Raghavan and R. Iyer. Evaluating vector-space and probabilistic models for query to ad matching. In *SIGIR Workshop on Information Retrieval in Advertising*, 2008.

[62] S. Ravi, A. Broder, E. Gabrilovich, V. Josifovski, S. Pandey, and B. Pang. Automatic generation of bid phrases for online advertising. In *International Conference on Web Search and Data Mining*, pages 341–350. ACM, 2010.

[63] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *International Conference on World Wide Web*, pages 521–530. ACM, 2007.

[64] O. J. Rutz and M. Trusov. Zooming in on paid search ads-a consumer-level model calibrated on aggregated data. *Marketing Science*, 30(5):789–800, 2011.

[65] Y. Song and L.-w. He. Optimal rare query suggestion with implicit user feedback. In *International Conference on World wide web*, pages 901–910. ACM, 2010.

[66] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *International on Conference on Information and Knowledge Management*, pages 553–562. ACM, 2015.

[67] R. Srikant and R. Agrawal. Mining generalized association rules. *Future generation computer systems*, 13(2-3):161–180, 1997.

[68] P. G. Srinivas, P. K. Reddy, S. Bhargav, R. U. Kiran, and D. S. Kumar. Discovering coverage patterns for banner advertisement placement. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 133–144. Springer, 2012.

[69] P. G. Srinivas, P. K. Reddy, and A. Trinath. Cppg: efficient mining of coverage patterns using projected pattern growth technique. In *Trends and Applications in Knowledge Discovery and Data Mining*, pages 319–329. Springer, 2013.

[70] P. G. Srinivas, P. K. Reddy, A. Trinath, S. Bhargav, and R. U. Kiran. Mining coverage patterns from transactional databases. *Journal of Intelligent Information Systems*, pages 1–17, 2014.

[71] B. Sripada, K. R. Polepalli, and U. K. Rage. Coverage patterns for efficient banner advertisement placement. In *International Conference Companion on World Wide Web*, pages 131–132. ACM, 2011.

[72] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *International Conference on World Wide Web*, pages 697–706. ACM, 2007.

[73] I. Szpektor, A. Gionis, and Y. Maarek. Improving recommendation for long-tail queries via templates. In *International Conference on World wide web*, pages 47–56. ACM, 2011.

[74] S. Thomaidou and M. Vazirgiannis. Multiword keyword recommendation system for online advertising. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 423–427. IEEE, 2011.

[75] G. Trimponias and D. Papadias. A survey on sponsored search advertising in large commercial search engines. 2013.

[76] A. Trinath, P. Gowtham Srinivas, and P. Krishna Reddy. Content specific coverage patterns for banner advertisement placement. In *International Conference on Data Science and Advanced Analytics*, pages 263–269. IEEE, 2014.

[77] M. Verma and D. Ceccarelli. Bringing head closer to the tail with entity linking. In *International Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 37–39. ACM, 2014.

[78] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[79] S. Yang and A. Ghose. Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Science*, 29(4):602–623, 2010.

[80] Y. Yang and F. Wang. *Budget constraints and optimization in sponsored search auctions*. Elsevier, 2013.

[81] S. Yuan, J. Wang, and X. Zhao. Real-time bidding for online advertising: measurement and analysis. In *International Workshop on Data Mining for Online Advertising*, page 3. ACM, 2013.

[82] H. Zaragoza, B. B. Cambazoglu, and R. Baeza-Yates. Web search solved?: all result rankings the same? In *International Conference on Information and Knowledge Management*, pages 529–538. ACM, 2010.

[83] W. Zhang, S. Yuan, and J. Wang. Optimal real-time bidding for display advertising. In *International Conference on Knowledge Discovery and Data Mining*, pages 1077–1086. ACM, 2014.

[84] W. Zhang, S. Yuan, and J. Wang. Optimal real-time bidding for display advertising. In *International Conference on Knowledge Discovery and Data mining*, pages 1077–1086. ACM, 2014.

[85] Y. Zhang, W. Zhang, B. Gao, X. Yuan, and T.-Y. Liu. Bid keyword suggestion in sponsored search based on competitiveness and relevance. *Information Processing & Management*, 50(4):508–523, 2014.

[86] K. Zhou, X. Li, and H. Zha. Collaborative ranking: improving the relevance for tail queries. In *International Conference on Information and Knowledge Management*, pages 1900–1904. ACM, 2012.