An approach to extract content-specific coverage patterns

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

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

by

Atmakuri Venkata Trinath
200902005
venkatatrinath.atmakuri@research.iiit.ac.in

Center for Data Engineering
International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2015

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "An approach to extract content-specific coverage patterns" by Atmakuri Venkata Trinath, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. P. Krishna Reddy
Center of Data Engineering
IIIT, Hyderabad

Dedicated to my parents
Mrs. Sabitha Atmakuri and Mr. Narendra Kumar Atmakuri,
for their everlasting love and support

Acknowledgments

This research would not be successful without the help of many individuals. I want to acknowledge the efforts of individuals who constantly motivated and helped me during my research. First of all, I would like to express my gratitude to my thesis adviser Prof. P. Krishna Reddy, from the bottom of my heart for having patience and belief in me. He continuously encouraged me by shaping my research and by filling the gaps of my knowledge to attain solutions for the problems. He always make time for brain storming sessions of the research. His continuous review process before submitting every research paper made me understand the way of presenting ideas in the paper. Finally, with his invaluable and constant effort, he made me understand how to purse research.

I would like to thank my colleagues P. Gowtham Srinivas, M. Kumara Swamy, Satheesh Kumar, Ammar Yasir and R. Uday Kiran at IT in Agriculture lab and Center for Data Engineering lab at IIIT Hyderabad. The learning environment created in the lab with these people made my work easier and interesting. I would like to appreciate the efforts of Gowtham and Kumara Swamy who guided me in improving my writing skills. I would also like to appreciate the efforts of Satheesh in giving different perspectives of the problem. I would like to thank all my friends for making my life peaceful and memorable at IIIT Hyderabad.

I have no words to express my gratitude to my parents Narendra Kumar, Sabitha and my brother Harsha Vardhan for their unconditional love, support, and continuous motivation through out my life. Finally, I thank almighty for supporting me at all times in my life.

Abstract

Online advertising is evolved as a major form of advertising in this century. It uses internet as a medium to provide advertisements to consumers. The major problem in online advertising is to provide advertisements to relevant set of users. Sponsored search, contextual advertising, behavioral targeting and banner advertising are different modes of online advertising. Banner advertising is one of the important mode of online advertising. The three major entities involved in banner advertising are advertiser, publisher and visitor. Advertiser is interested in endorsing his products through banner advertisements. Publisher manages a website by selling the advertisement space to multiple advertisers. Visitors visit web pages of their interest containing banner advertisements. The aim of an advertiser is to spread his advertisement to a certain percentage of relevant visitors. On the other hand, to generate more revenue for a given website publisher has to manage coverage demands of multiple advertisers by providing appropriate sets of relevant web pages. The major issues in banner advertising are efficient auction system to sell advertisement space, scheduling banner advertisements on a web page for maximizing the revenue, pricing models to be followed for banner advertisements, targeted banner advertisements based on the user behavior, matching banner advertisements with the web pages, allocating relevant sets of web pages to multiple advertisers with corresponding coverage demands and finding appropriate set of users interested in particular content of web pages.

In this thesis, we focus on extracting relevant sets of web pages for efficient placement of banner advertisements based on the click-through data of a website and key words describing every web page assuming similar visitors behaviour. The following two factors are important in placing banner advertisements on web pages of a website. a) Percentage of users visiting a web page. b) Relevance of banner advertisement to the content of web page. In the literature, model of *coverage patterns* and algorithms to mine *coverage patterns* from transactional data are proposed. The usefulness of these patterns is demonstrated by applying it to banner advertisement placement problem. For a given website, *coverage pattern* is a set of web pages visited by certain percentage of visitors. Publisher can extract these patterns from the click-through data of a website to meet the demands of multiple advertisers.

However, the model of *coverage patterns* does not consider the relevance of web pages to the advertisement. This may lead to placing banner advertisement on irrelevant web pages. In this thesis, we propose a model of *content-specific coverage patterns* to capture the relevance of advertisement to web pages along with the percentage of visitors. In general a web page contains content related to multiple topics and different users visit the web page with varied topical interests. The research issue here is

to find the percentage of users interested in particular content of web pages in pattern. In this thesis, we propose a notion of *topic coverage* which captures the fraction of visitors interested in particular topic/keyword of web pages in the pattern. If every web page is labelled with all the keywords relevant to the content of web page, *topic coverage* of every keyword can be computed. As the *coverage patterns* are extracted based only on the percentage of visitors, most of them are irrelevant to the advertiser. The research issue is to extract the patterns of web pages with certain percentage of visitors and are relevant to the advertiser. In this thesis, we propose the notion of *content-specific coverage patterns* which defines the set of web pages visited by certain percentage of users interested in particular content of the web pages. We also propose the methodology to extract *content-specific coverage patterns* from the click-through data given keywords describing every web page. Experimental results on real world datasets show that the proposed model of *content-specific coverage patterns* extracts relevant set of patterns efficiently over the model of *coverage patterns*. The results also show that the proposed approach generates all the patterns relevant to the topics given by the advertiser while the model of *coverage patterns* generates some irrelevant patterns. Finally, we conclude that by bidding for slots on the web pages of *content-specific coverage patterns*, advertisers will have relevant set of expected number of visitors.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data mining has been employed in e-commerce for effective management of inventory, increasing the sales of products and in capturing the interests of customers. E-commerce uses internet as a medium to sell products. Internet monetization deals with the techniques used in converting the website traffic into revenue. Social networking, search engine, blogs and community websites are different sources of information on internet. Internet monetization techniques like advertising are used to link this information with the sales of products on e-commerce websites. Also, the number of users using internet is increasing at a rapid rate in the current century. The revenue of a website depends upon the number of users visiting the website. As more number of users visiting the website, advertisers can use this opportunity for targeting their products to these users by placing advertisements in these websites.

According to the Internet Advertising Bureau (IAB) internet advertising revenue report conducted by PricewaterhouseCoopers (PwC), annual revenues in United States alone for the year 2013 is $42.8 billion [1]. This is $6.2 billion (17%) higher than the revenue in the year 2012. Internet is becoming one of the major medium for advertising with the continuous increase in advertising revenue. Online advertising drives internet economy and provides sustainability to e-commerce websites. It also serves as the major source of revenue for search engines. Online advertising is preferred when compared to offline advertising because it is faster to diffuse information to a set of targeted users irrespective of geographical location. Initially, online advertising is similar to offline advertising where graphical advertisements are dominated over textual advertisements. Later, different modes like sponsored search advertising, contextual advertising and behavioral targeting are incorporated into online advertising. According to IAB report [1], the online advertising revenue due to banner advertisements is 19% of total revenue. Banner advertising is one of the important mode of online advertising. In this thesis, we address the issues related to banner advertisement placement. We attempt to solve the problem of banner advertising in providing multiple sets of web pages based on the percentage of visitors and topics given by the advertiser.

The organisation of this chapter is as follows. In this chapter, we first present the overview of different modes in online advertising in section 1.1. Next, we discuss the background and issues involved in

banner advertising. We present the overview of proposed approach in section 1.2. Finally, the contributions and the organisation of thesis are presented in section 1.3 and section 1.4 respectively.

1.1 Overview of the Modes in Online Advertising

Sponsored search advertising, contextual advertising (CA), behavioural targeting (BT) and banner advertising are different modes of online advertising (Figure 1.1). In this section, we discuss about the overview of different modes in online advertising.

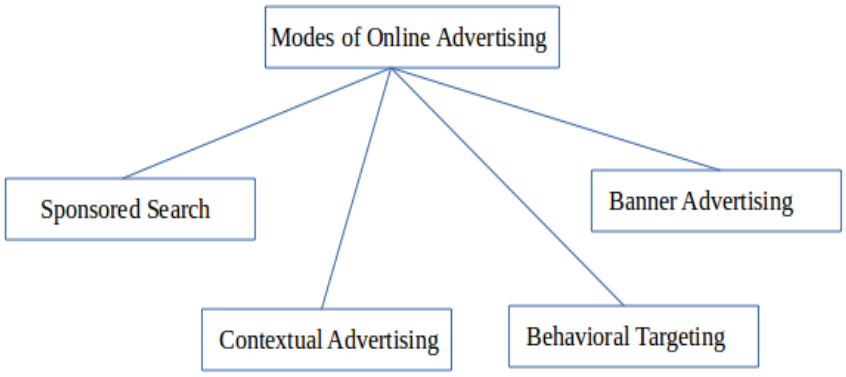

Figure 1.1 Modes of online advertising

1.1.1 Overview of Sponsored Search

Sponsored search advertising [2, 3] is a means of advertising where advertiser links are shown based on the keywords searched. The revenue of most of search engines is mainly due to sponsored search. Advertiser provides hyper-links which are annotated with keywords describing advertisement along with title and description. He also provides bids to specific keywords related to the advertisement. Bids generally consists of maximum price per keyword and can also include activation period, language, geographical and other constraints. Search engine runs manual and automated review process to make sure that the keywords given are related to the advertisement. When a user searches for a query, search engine searches for advertisements whose keywords are matching with the keywords specified in the query and displays them along with the search query results. When many advertisers bid for the same keywords, an electronic auction is run to determine the ranking order of advertisements based on the bids and click-through rate of advertisements. Search engine collects data regarding the clicks by the users and charges corresponding advertisers based on the number of clicks for his advertisement. Figure

1.2 shows the sponsored advertisements in the top and right boxes of web page based on the keywords ”bird” and ”houses” searched in the search engine..

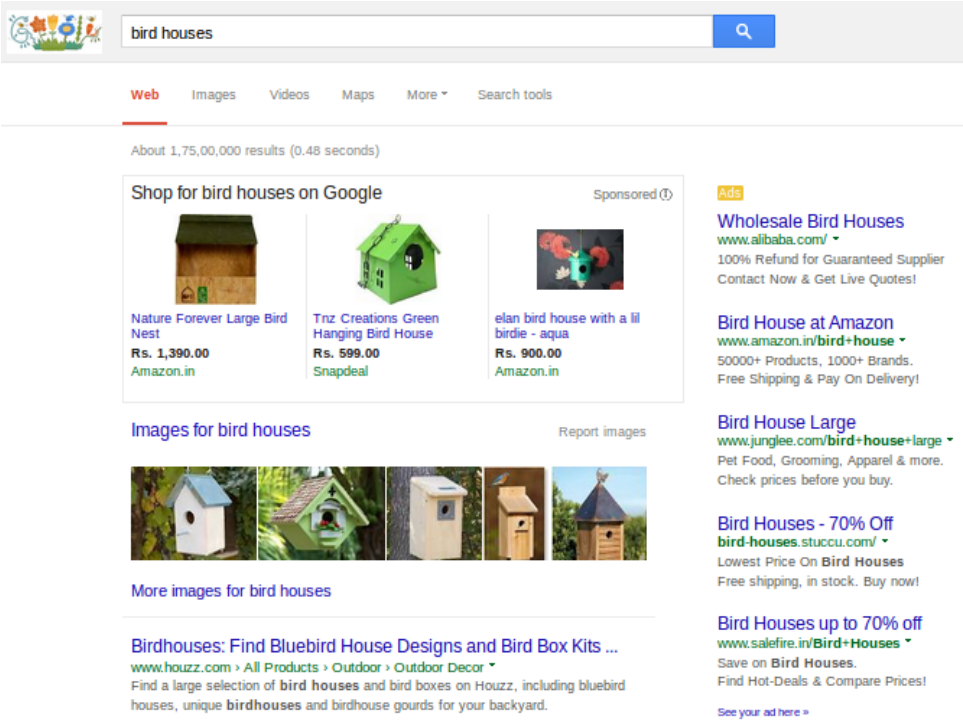

Figure 1.2 Top and right boxes shows the sponsored advertisements on searching for keywords ”bird” and ”houses” in Google search engine

1.1.2 Overview of Contextual Advertising

Contextual advertising [4] deals with the advertisements to be placed on the web pages of a website. It aims to place advertisements based on the relevance to the web page.

The following four entities are involved in *contextual advertising*. a) Publisher: *Publisher* owns web pages and tries to maximize advertising revenue by displaying advertisements and ensuring better user experience. b) Advertiser: *Advertiser* provides advertisements to endorse his products. c) Ad-network: *Ad network* decides which advertisements are to be placed on web pages by acting as a mediator between advertiser and publisher. d) Users: *Users* visit the web pages containing advertisements.

The revenue of an advertisement network is as in Equation 1.1. Here, k is the number of advertisements displayed on a web page p, $price(a_i, i)$ denotes the price per click of an advertisement a_i at position i and $P(click|p, a_i)$ denotes the probability of clicking advertisement a_i on web page p. It can be observed that revenue R can be maximised by selecting the advertisement which maximizes the probability of clicking on it given fixed price of an advertisement at a particular position.

$$R = \Sigma_{i=1..k} P(click|p, a_i) \times price(a_i, i) \tag{1.1}$$

Figure 1.3 Example showing three partitions of a web page in newspaper website. Middle slice shows the launch of a movie while the bottom slice shows the content-based advertisements by Google's Adsense.

In Figure 1.3, the middle slice shows the content of the web page and the bottom slice shows the advertisements picked based on the content by Google's content-based advertising system, Adsense.

1.1.3 Overview of Behavioral Targeting

Behavioral targeting [5] aims to provide advertisements on a web page to particular users targeted based on their search and browsing behaviours. For example, a user browses an apparel website for buying shoes. If the same user visits other websites of his interest, an advertisement related to shoes can be displayed in the corresponding website based on his browsing behaviour. The steps involved in the process of behavioral targeting are as follows (Figure 1.4). In the first step, a user visits a website for his products of interest. The website tracks user's interest by tracking the products visited by him. In the next step, the user visits other popular websites related to news, fashion and apparel. In step-3, he finds the advertisement related to the products visited by him. Finally in step-4, user clicks the advertisement and finally return to the same website in step-1.

BT uses every individual's information like the web pages visited by the user and web searches made by the user. *User segmentation* and *user segmentation ranking* are the two steps involved in *behavioral targeting*. *User segmentation* segments users based on the users' behaviour while *user segmentation ranking* ranks corresponding user segments for a particular advertisement. In this study [5], user be-

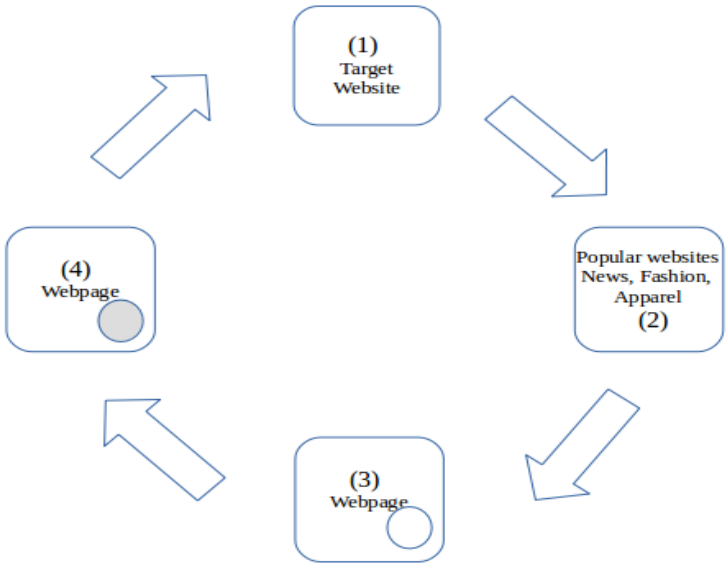

Figure 1.4 Process of Behavioral Targeting

havior is modelled in two ways. User's behavior can be represented based on users' search queries or by users' clicked web pages. Based on users' page views, it can be modelled as follows. Every user is represented as a vector in user-by-URL matrix where each row is a *user* and column is a *url*. Mathematically, it is represented as $U \in R^{g \times l}$, where g is the number of users and l is the number of distinct url's clicked by all these users. User is represented as a row in U with weight of each component in vector is as in Equation 1.2. This is adopted from TF-IDF (term frequency-inverse document frequency) [6] by assuming every user as a document and url's as terms in the document. Similarly, user behavior is also represented based on users' search queries by assuming keywords in a query as the terms in a document.

$$u_{ij} = (\log(\#times\ user\ `i'\ clicked\ URL\ `j') + 1) \times \log(\frac{l}{\#users\ clicked\ URL\ `j'}) \tag{1.2}$$

1.1.4 Overview of Banner Advertising

Banner advertising is a mode of advertising where banners are placed on web pages. A banner advertisement is a box containing graphics with a hyper link associated with it and is redirected when a user clicks on it [7]. The following three major entities are involved in banner advertising.

a) **Advertiser** Advertiser is interested in endorsing his products through banner advertisements. The major aim of the advertiser is to increase the reach of his products to maximum percentage of people visiting web pages.

b) **Publisher** Publisher manages a website by providing advertising space to multiple advertisers. The goal of the publisher is to maximize revenue by efficiently managing advertising space available in the website to satisfy the requests of multiple advertisers.

c) **Visitor** Visitor visit web pages of his interest and explore the banner advertisements in the web pages. These banner advertisements are of more interest to the visitor if they are relevant to the web pages as he visits pages of his interest.

It is necessary for the advertiser to display his advertisement to certain percentage of users visiting the website. For a given website and time period, one could analyse the visitors' behaviour by processing the transactions generated based on the click-through data and identify the sets of relevant web pages to the advertiser that are visited by certain percentage of visitors. This thesis investigates on approaches to find the sets of web pages that can ensure minimum visitors interested in the corresponding advertisement provided by the advertiser. In the next section, we present the issues involved in banner advertising and overview of the proposed approach.

1.2 Overview of the Proposed Approach for Content-specific Coverage Patterns

In this section, we present the issues involved in banner advertising and give an overview of the proposed model of *content-specific coverage patterns*.

1.2.1 Issues Involved in Banner Advertising

The major issues in banner advertising are maximizing revenue of website for a given set of advertisers, efficient auction system to sell advertisement space, scheduling web pages to advertisers for placement of banner advertisements, pricing models of banner advertisements and scheduling multiple banner advertisements on a single banner for maximizing revenue.

The major issues in auction system are efficient allocation of advertisements for maximizing revenue based on the real time bids and the optimal usage of advertising space. Bidding agents tend to acquire advertising space by participating in real time auctions [8, 9]. In addition to this, publisher introduced different pricing models like pay-per-click (PPC), pay-per-impression (PPM) and pay-per-action (PPA) in order to attract advertisers. Different pay-per-action models [10, 11] are introduced in order to prevent click fraud. Scheduling banner advertisements in a given web page with multiple time slots for maximizing revenue is formulated as an integer programming problem [7]. This thesis investigates on approaches for efficient placement of banner advertisements.

In banner advertisement scenario, for a given website advertiser has to place his advertisement on a set of web pages such that it is displayed to minimum percentage of visitors (coverage). Also, he is interested to place his advertisement on relevant web pages. Publisher of the website receives such requests from multiple advertisers with respective coverage and content requirements. Publisher has to manage multiple advertisers by providing set of web pages to every advertiser considering his coverage and content requirements. Given keywords describing every web page and click-through data of a

website, we attempt to solve the problem of providing set of web pages to an advertiser based on the keywords related to advertisement.

1.2.2 Overview of the Proposed Approach

In this section, we present the problem of banner advertisement placement and overview of the proposed model of *content-specific coverage patterns*. In banner advertisement scenario, advertiser expects his advertisement to be displayed to certain percentage of relevant visitors. While the publisher has to manage requests from multiple advertisers for maximizing the revenue of a website. The following two factors are important in placing banner advertisements on web pages of a website. a) Percentage of users visiting a web page. b) Relevance of banner advertisement to the content of web page.

If a web page has more number of visitors, advertisement placed on the web page will be seen by more number of users. In literature [12], a notion of *coverage patterns* is proposed. *Coverage Patterns (CPs)* are set of web pages visited by certain percentage of visitors. Assuming similar visitors behaviour in the future, these patterns extracted from the click-through data are useful by giving multiple options to the advertiser. *Coverage Patterns (CPs)* are extracted using the notions of *coverage support* and *overlap ratio* [13]. *Coverage support* of a pattern indicates the spread of banner advertisement to a fraction of users while *overlap ratio* indicates the amount of repetitive display between users visiting web pages of a pattern. The concept of coverage patterns solved the issue of percentage of visitors while it has not considered the relevance of banner advertisement to the content of web page.

However, coverage patterns generated based only on the percentage of visitors may lead to placing advertisement on irrelevant web pages, which might lead to annoyance for the visitors and also reduces the reach of advertisement because of low click-through rate. This in turn impacts the goals of advertisers and publishers. The click through rate can be improved and the advertisements will be more interesting to the visitor if the advertisement is posted on the web pages whose content is relevant to the advertisement. So, the relevance of banner advertisement to the content of web page also plays an important role in banner advertisement scenario. This can be clearly observed from Example 1. In this thesis, we have proposed the model of *content-specific coverage patterns* to capture these aspects.

Example 1 *Suppose a web page 'a' is related to 'football' and a banner advertisement related to 'hockey' is shown on the web page 'a'. A user visits a web page 'a' may not be interested in 'hockey' which leads to less number of visitors clicking the advertisement.*

If every web page is labelled with the keywords [1] relevant to the content of web page, advertiser may specify the keywords related to the banner along with the percentage of visitors. In general a web page contains content related to multiple topics and different users visit the web page with varied topical interests. The research issue here is to find the fraction of users visiting the web page for a particular

[1] Words 'keyword' and 'topic' are used interchangeably.

topic. This is helpful in placing the advertisement related to a particular topic on a web page, having more number of users visiting for that topic. If a web page is having more visitors for a particular topic and an advertisement related to that topic is placed on it, the probability of users clicking advertisement increases. *Content-specific coverage patterns* proposed in this thesis captures the aspect of finding the fraction of users visiting a web page for a particular topic by analysing the click-through data of the website. This is based on the assumption that the process of users visiting web pages online is a Markov process [14]. Experimental results on real world click-through data shows that by bidding for slots on the web pages of particular content-specific coverage patterns, advertisers will have a good set of visitors. Next, we explain the contributions of the thesis and in the last section we present the outline of the thesis.

1.3 Contributions of Thesis

The concept of *coverage patterns* is not effective as it does not captures the relevance of content in web pages to the advertisement. The existing approaches in the literature related to online advertising have not concentrated on providing options to publisher while assigning to advertisers (Chapter 2). The major contributions of this thesis are as follows.

(I) **Defining the model of content-specific coverage patterns.** We proposed the model of *content-specific coverage patterns* for better placement of banner advertisements based on the number of visitors and relevance of banner advertisement.

(II) **Defining topic coverage, factual topic coverage and algorithms for computing them.** We proposed the notions of *topic coverage* and *factual topic coverage* to capture the fraction of users interested in a particular topic of web pages in pattern. We also proposed algorithms to compute topic and factual topic coverage of a topic with respect to a pattern.

(III) **Algorithm to extract content-specific coverage patterns.** Given click-through data of a website and topic labels of every web page, we proposed an algorithm to extract *content-specific coverage patterns*. This algorithm makes use of fact that process of users visiting the web pages online is a Markov process.

(IV) **Experimental Results.** We conducted experiments on real world dataset for evaluating the usage and efficiency of the model of *content-specific coverage patterns* over the model of *coverage patterns*.

(V) **Literature Survey.** We presented the issues and solutions proposed in literature related to different modes of online advertising.

1.4 Organisation of Thesis

The rest of the thesis is organised as follows.

(I) **Chapter 2: Related Work** In this chapter, we discuss about the work done in literature related to online advertising and show how the proposed approach is different from other approaches in the literature.

(II) **Chapter 3: Overview of Coverage Patterns Model** In this chapter, we present the concept of coverage patterns and the overview of coverage patterns model as proposed in literature.

(III) **Chapter 4: Model of Content-specific Coverage Patterns** In this chapter, we discuss about the proposed approaches for computing the fraction of users interested in particular topic of web pages in a pattern and extracting *content-specific coverage patterns* from click-through data of a website.

(IV) **Chapter 5: Conclusion and Future Work** In this chapter, we discuss about the summary of thesis, conclusion and future prospects in this direction of research related to banner advertising.

Finally, we presented the list of related publications and references at the end.

Chapter 2

Related Work

Online advertising has become predominant and a major factor in driving internet economy. As online advertising helps in faster diffusion of information irrespective of user's location, many advertisers are looking forward to online advertising. The advertisement revenue of a website depends on the number of visitors and the advertisements shown on web pages [15]. There has been a lot of work done in literature related to online advertising. In this chapter, we discuss about the work done in literature related to online advertising and how the proposed model of *content-specific coverage patterns* is different from the other approaches in the literature.

The organisation of this chapter is as follows. In the first section 2.1 we present the overview of work done in literature related to *sponsored search*, *contextual advertising* and *behavioral targeting*. In the next section, we discuss about the related work in the current topic of interest of this thesis. In section 2.2, we discuss about the work done in solving different issues involved in banner advertising. In the section 2.3, we discuss about how the proposed approach in the thesis is different from the above approaches in literature. Finally, we present the summary of this chapter in section 2.4.

2.1 Related Work in Sponsored Search, Contextual Advertising and Behavioral Targeting

2.1.1 Related Work in Sponsored Search

Sponsored search is the mode of advertising where textual advertisements are shown to users when they search for a query. These textual advertisements are selected based on the keywords of the search query. The brief history of *Sponsored Search* is given in [2] and [3]. The process of sponsored search is briefly explained in section 1.1.1.

The major problems involved in *sponsored search* are as follows. Pricing models to determine the cost of advertisements shown to users, matching textual advertisements with the keywords in the search query, platform and a mechanism for advertisers to bid for the keywords, ranking different advertisements based on the metrics computed from bids, predicting the click-through rate (CTR) of advertise-

ments and bidding strategies to be followed by search-engine for generating more revenue are the major issues involved in the process of *sponsored search*. Now, we discuss the work done in literature to solve these problems.

Three different types of pricing models, pay-per-impression, pay-per-click and pay-per-action are proposed and advertisers have to pay based on agreement between him and the search engine provider. *CPM (Cost-Per-Mille)* is introduced as part of pay-per-impression model which means that the cost to display an advertisement thousand times. As there is no risk for search engine, *CPC (Cost-Per-Click)* is proposed as a part of pay-per-click model where the advertiser has to pay based on the number of clicks received by advertisement. As click may not corresponds to a purchase of product, *CPA (Cost-Per-Action)* is proposed under pay-per-action model [10, 11] where advertiser pays based on the number of actions taken based on the advertisement. Several tools are prepared by different companies to enable advertisers to compute *CPA* and set their corresponding *CPC* bids. For example, Google uses *AdWords* [16] tool for advertising.

Two metrics *conversion rate* and *click-through rate* [17] are proposed to rank advertisements for a given search query. Both measures are conditional probabilities. *Click-through rate (CTR)* is the probability of click given an impression which means that the percentage of visitors clicking the advertisement. *Conversion rate* is the probability of action given a click which means that the percentage of visitors who took an action like buying a product. In the literature, work is done on predicting the click-through rate of an advertisement. In [18–23], click-through rate is predicted by building linear models after extracting features from queries and advertisements. *CTR* prediction can also be done by mapping it to a recommendation problem and solving it using collaborative filtering techniques. In [24], matrix factorization techniques used in collaborative filtering are used to predict the *CTR*. In [25], a study is done on different factors of commercial intent of query for analysing click behaviour. Ashkan et al., further modelled it in [26] by adding query bias to it. Recently in [27], contextual factors like depth of an advertisement, diversity of a query and the interaction between advertisements are used for analysing click behaviour in sponsored search.

As advertisers have to bid on keywords, there may be mismatch between the keywords and the search phrases given by a user. In the literature, there were efforts to find relevant keywords based on the keyword co-occurrence. For instance, keywords co-occurrence in search engine query logs is used to find relevant keywords in [16]. In [28], concept hierarchy is used to form concepts containing keywords which are further used to suggest key words. Several auction models [8, 29] which acts as a bidding agents on the behalf of an advertiser have been proposed to manage the advertising campaign.

2.1.2 Related Work in Contextual Advertising

Apart from *Sponsored Search* where advertisements are recommended based on search queries, advertisements are also placed on web pages. In *contextual advertising* [30], advertisements are placed on web pages based on the relevance between the advertisement and web pages. From the study in [31], it is evident that if advertisements are relevant to the interests of user, it improves their experience and the

probability of clicking advertisements. The entities involved in the process of *contextual advertising* are briefly explained in subsection 1.1.2.

The major issues in *contextual advertising* are as follows. Ranking of advertisements based on relevance with the content of web pages, mismatch in the vocabulary of advertisement and web pages, and extracting advertising keywords from web pages are major issues in *contextual advertising*.

In the literature there are different approaches for contextual advertising. In [4], both advertisements and web pages are represented as vectors in vector space model. Advertisement title, body and bid phrases form the basis of advertisement vector. Advertisements are ranked based on the cosine between these vectors. To prevent the impedance mismatch between the vocabulary used in advertisements and web pages, different terms are taken from similar web pages weighted based on similarity between them. A system for contextual advertising based on both syntactic and semantic features is proposed in [30]. Another approach to contextual advertising is to reduce to the problem of *sponsored search* advertising. First we extract the advertising keywords from web pages and matching them with bid phrases of advertisements. This lead to research in the direction of extracting advertising keywords from web pages.

The keyword extraction algorithms have four major steps which are described as follows. a) Content Extraction: In this step, main content is extracted from the given web page in html format. b) Candidate Keywords Selection: This step returns a set of candidate keywords from the contnet of the given web page. c) Feature Extraction: Using some rules, set of features are formed based on these candidate keywords. d) Predictive Model: In this step, the designed predictive model classifies the keywords into relevant and irrelevant categories based on the extracted features from the previous step.

In [32], learning algorithm is proposed based on the features like term frequency of potential words, inverse document frequency, frequency of the term in search query logs and presence of the term in meta data of a web page. A machine learning approach for extracting keywords is proposed in [33]. [33] uses a linear and logistic regression models learnt from human labelled data along with document and text features like term frequency, title, anchor text, phrase length and meta data. A method to extract advertising keywords based on the part of speech (POS) patterns is proposed in [34]. In [34], the POS patterns restrict the number of classes the classifier has to handle and fetches words that are related to any of these POS patterns. Recent work in this direction to extract keywords automatically that represents the web pages based on NLP techniques along with POS and named entities tagging is proposed in [35]. In [35], it also extracts some related keywords which are not present in the web page.

2.1.3 Related Work in Behavioral Targeting

Behavioral Targeting (BT) is the mode of advertising which provide advertisements based on users search and browsing behaviour. The empirical study on *BT* in [5] proved three conclusions based on the click-through data of advertisement over a period of seven days. They are given as follows. 1) Users clicking the same advertisement have similar behaviors on the web. 2) Click through rate of an advertisement can be improved by an average of 670% after proper segmentation of users for behavioral

targeting. 3) Modelling short-term behavior of users is more useful and effective over the long-term behavior of users. The process of *behavioral targeting* is briefly explained in subsection 1.1.3.

The study in [5] shows that users who clicked the same advertisement are 90 times more similar than the users who clicked different advertisement and *CTR* can be improved after *user segmentation*. In [36], work is done on large scale behavioral targeting where it derives linear poisson regression model based on user click-through data and predicts click-through rate (CTR). They have also proposed highly scalable map reduce statistical learning algorithm. This model achieved 20% improvement in *CTR* by using efficient probabilistic model prepared from a large training dataset.

There are many commercial advertising systems by using *behavioral targeting*. For example Yahoo smart ads [37] uses behavioral targeting in addition to geographic and demographical targeting. Adlink [38] uses short term behavior of users for behavioral targeting. Double Click [39] uses browse type and operating system used by the users for segmenting users. The other commercial *behavioral targeting* systems are Blue Lithium [40], Almond Net [41], NebuAd [42], Burst [43] and TACODA [44].

Recent work on *behavioral targeting* for temporal analytics on big data is proposed in [45]. There have been efforts for integrating *behavioral targeting (BT)* into *contextual advertising (CA)* in [46]. In [46], a new notion of relevance between advertisements and web pages is proposed based on the click-through behavior of users. By integrating this into the notion of *contextual advertising*, combined measure for relevance is proposed. The combined model achieved more performance over using either *behavioral targeting* or *contextual advertising*.

2.2 Issues and Related Work in Banner Advertising

Banner advertising is a mode of advertising with banners. A banner is a box containing graphics which links to a product on advertiser's website. It constitutes *publishers*, *advertisers* and *visitors*. The goals of advertisers, publishers and visitors are discussed in subsection 1.1.4. The major issues in *banner advertising* are as follows. Designing efficient auction system for optimal allocation of advertising space, pricing models of banner advertisements, scheduling banner advertisements for maximizing revenue, targeted banner advertising based on user behavior and scheduling web pages of a particular website to advertisers for efficient placement of banner advertisements are major issues in *banner advertising*. Now, we present the work done on every issue in banner advertising.

(I) **Auction System** Designing an efficient auction system for optimal allocation of advertising space by matching multiple advertiser's needs is a very complex issue. Advertisers have two options of buying slots for placing advertisements [8]. a) Guaranteed contract: Publisher ensures to provide pre decided number of impressions in a given time frame to meet the requirements of advertisers. b) Spot market: Advertisers can buy impressions of one page view at time. This means that every time a user loads a web page, auction is held for advertisers to display their advertisements. Publishers such as RightMedia Exchange uses spot markets for allocation [47]. Publisher faces a problem in choosing either guarantee contract or spot auction. Arpita Ghosh et al [8] provides

the solution for this issue. It has two components. One is maximal representative allocation and the other is randomized bidding. Maximal representative allocation specifies the fraction of impressions at every price point allocated to the contract while randomized bidding specifies allocation through auction. Publisher also faces the problems of having more advertisers with few higher bids and higher bids may arrive later after being pre allocated to other advertisers. Jason et al [48] proposed a notion of buy back which can revoke prior allocation decisions with a cost. Florin et al [49] proposed auction mechanism for advertisement slots in future but can cancel the allotment before displaying the advertisement.

(II) **Scheduling of banner advertisements** For an advertisement slot on a given web page, many advertisers compete to place their advertisements for some time interval. Publisher has to decide which advertisements have to be placed in a particular advertisement slot of known parameters in different time intervals. The banner advertisement placement problem is described as follows [7]. Given a set of time slots T, slot size S and a set of advertisements A with each advertisement $i \in A$ characterized by a size $s_i \leq S$ and a weight $w_i \leq |T|$, find a subset of advertisements $A' \subseteq A$ such that each advertisement $i \in A'$ is displayed exactly once in each of w_i slots such that $\sum_{i \in A'} s_i w_i$ is maximized. A feasible subset A' is such that at most one copy of an advertisement can be assigned to a time slot. The pricing scheme can be assumed as proportional to the area of space consumed for a given advertisement. The time slots are assumed to be of same duration. The CPM is generally proportional to the size of banner advertisement, which means that maximizing $\sum_{i \in A'} s_i w_i$ is equivalent to maximize revenue. Adler et al [50] solved offline advertisement placement problem while Amiri et al [7] mapped this problem as integer programming problem and solved it using lagrangean decomposition. In [51], heuristic algorithms are proposed to maximize the revenue for placing multiple banner advertisements in a single web banner.

(III) **Pricing models** Publisher attracts advertisers with multiple pricing schemes to increase the revenue for this website [52]. Online advertisements are being used either for brand awareness or for improving the sales of products. Different pricing models for these purposes are discussed for *sponsored search* in section 2.1.1.

(IV) **Targeted banner advertisements** Targeted advertising is a mode of advertising where advertisements are shown to users based on the their interests. Cookie based, click behavior and navigational patterns based targeted advertising are different modes of targeted advertising. In literature, a framework for targeting *banner advertising* is first proposed in [53]. The basic idea in this work is to have a user profile for every individual and target audience profile for every advertisement and a methodology to match users with target audience profile for a specific advertisement. For example, advertisement related to "books" can be shown to targeted audience who searched for books or purchased books recently. The work in [54] shows that the banner advertising effects the internet purchasing behaviour and targeted banner advertising gives more returns when compared

to non-targeted banner advertising. The detailed process of targeted advertising is explained in subsection 2.1.3.

(V) **Efficient placement of banner advertisements** Advertiser aims to spread his banner advertisement to wider audience for endorsing his products. On the other hand, visitors visit website with various interests. These visitors can be potential customers for advertiser so there is need for efficient placement of banner advertisements to maximize the number of visitors and relevance of banner advertisement to the content of web pages. Using the notion of *coverage*, a new knowledge patterns, *coverage patterns* are proposed in [12] for efficient placement of banner advertisements. In literature, the notion of *coverage* is used in solving set cover problem in set theory [55]. In [56], the notion of *coverage* is used in solving node cover problem in graph theory. In [57], the notion of *coverage* and *overlap* are used in solving topical query decomposition. *Coverage patterns* are set of non-overlapping data items covering certain percentage of transactions in a transactional database. These patterns are found useful in banner advertisement placement. For a given website, using the click-through data of website, one can find the set of web pages that are visited by certain visitors population assuming similar visitors behavior. These sets of web pages helps advertiser by displaying his advertisement to certain visitors population and helps publisher in managing multiple coverage demands of different advertisers. In [12], *coverage patterns* are defined using two measures *coverage support* and *overlap ratio*.

The first attempt to mine *coverage patterns* from transactional databases is done in [13]. A level-wise approach, *CMine* by exploiting the *sorted closure property* [58] of *overlap ratio* is proposed. This approach is similar to *Apriori* algorithm used in *association rule mining* [59]. This approach is basically multi-pass, candidate generation and test approach. This approach requires multiple scans of database and generate a large set of candidate *non-overlap patterns*. Next, an improved *pattern-growth* approach, *coverage pattern projected growth (CPPG)* is proposed based on the notion of *non-overlap projection* [60]. This is based on the idea of *projected databases* [61, 62] used in frequent pattern mining. The general idea is to form *non-overlap projected database* for every *non-overlap pattern* recursively and confines test only to the smaller projected databases. It has been observed that there is a significant improvement in performance of *CPPG* over *CMine*. By using the notion of *minimal coverage patterns* in addition to the notion of *non-overlap patterns*, a new approach *enhanced coverage pattern projected growth (ECPPG)* [63] is proposed. This approach uses the *upward closure property* [64] of *coverage support* which implies that if a pattern satisfies *coverage support* then all its super sets satisfy *coverage support*. A pattern, X is said to be *minimal coverage pattern* if no proper subset of X has coverage support $CS \geq minCS$ and overlap ratio $OR \geq maxOR$. Once *minimal coverage pattern* is found, all the patterns containing this pattern as prefix can be found by adding every item with lower precedence in frequency ordered item list (f-list) and by checking only OR. A brief account on different approaches for mining *coverage patterns* is proposed in [63].

2.3 How Proposed Approach is Different

Banner advertising deals with banners to be placed on web pages. Every banner is associated with hyperlink to advertiser's website and advertiser provides keywords related to banner while providing advertisement to the publisher. Given click-through data of a website and minimum percentage of visitors to be displayed, the model of *coverage patterns* attempts to generate patterns of web pages that ensures the advertisement to be displayed to given percentage of visitors assuming similar visitors' behavior. The concept of *coverage patterns* is not much helpful because of the following reason. If the banner is not related to atleast one of the *topics* of web pages in pattern, the users visiting web pages may find the advertisements uninteresting which leads to decrease in the number of relevant users of the advertisement. Also, it is important to know the number of users interested in topics of web pages in pattern, as more number of users interested in a particular topic leads to increase in probability of clicking advertisement related to that topic. This is an interesting research issue, if we have only topics related to every web page and click-through data of website. This is different from *behavioral targeting* because users are segmented to create a user profile and advertisements are targeted to users based on the interests of similar users in behavioral targeting. In this thesis, we have proposed approaches to compute the number of users interested in particular topic of web pages in pattern based on assuming the process of visiting web pages online is a Markov process [14]. By integrating these concepts into the model of *coverage patterns*, we have proposed a model of *content-specific coverage patterns* which extracts patterns of web pages which satisfies coverage requirements of advertisers and are also relevant to the banner advertisement. In this thesis, we have also proposed an algorithm, *content-specific coverage pattern projected growth (CSCPPG)* to extract *content-specific coverage patterns* from click-through data given the labels (topics) related to every web page. From the experimental results, it is proved that the proposed approach extracts better set of patterns over the previous model of *coverage patterns*.

2.4 Summary

In this chapter, we have presented the work done in literature related to *online advertising*. We have also shown how our approach is different from the work done in literature. We present the overview of *coverage patterns* model in the next chapter.

Chapter 3

Overview of Coverage Patterns Model

Banner advertising is an important mode of online advertising. The entities involved in banner advertising along with their expectations are briefly explained in subsection 1.1.4. The issues involved in banner advertising and the work done in literature are explained in section 2.2. In this thesis, we proposed the model of *content-specific coverage patterns*. In this chapter, we discuss about the problem of banner advertisement placement and overview of the model of *coverage patterns* as proposed in literature. Later, we discuss about the model of *content-specific coverage patterns* in subsequent chapters. In this chapter, we also present the overview of algorithms used to extract *coverage patterns* from click-through data of a website.

The organisation of this chapter is as follows. First, we present the overview of problem of banner advertisement placement in section 3.1. Next, we present the model of *coverage patterns* as proposed in literature in section 3.2 and overview of algorithms used to extract *coverage patterns* in section 3.3. Finally, we present the summary of this chapter in section 3.4.

3.1 Overview of Banner Advertisement Placement

Advertisers want to display their advertisements to a certain percentage of visitors. They have to bid for advertisement slots on web pages by managing advertising budget. For a given website, advertiser doesn't know the appropriate set of web pages to bid according to this coverage requirements. Publisher receives such requests from multiple advertisers with their corresponding coverage requirements. Visitors visits web pages of his interest and check out advertisements. The click-through data of a visitor captures the navigational path in a web site. The click-through data of all visitors and frequently viewed web pages can be helpful in determining the set of web pages to be assigned to an advertiser. But, it is also important to note that it is annoying to show the same advertisement multiple times to a user in a same session. For example, if we consider the placement of advertisement on frequently viewed web pages of a website, we are displaying the same advertisement multiple times to the same visitor. Overlap of visitors of two web pages captures the set of visitors who are viewing both the web pages in the same session. Displaying same advertisement multiple times to the user is a loss to advertiser as

he has to pay for every impression as per pay-per-impression model. Also, there is a problem of banner burn out which means that showing the same banner advertisement may decrease the click through rate of an advertisement. As it is annoying to show same advertisement in the same session, by minimizing the overlap the redundancy of showing advertisement decreases. So, mining sets of web pages from click through data by minimizing overlap is helpful to publisher in meeting the demands of multiple advertisers.

3.2 Overview of Coverage Patterns Model

The model of coverage patterns is as follows [12, 63]. We consider the banner advertisement scenario and the transactions generated from the click-through data of a website to present the model. However, the model can be extended to any transactional dataset. The click-through data of a website represents the navigational patterns of multiple visitors. 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 and the topic labels of web pages are given in Table 4.1, L.

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 atleast 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 atleast 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 closure property* [58] 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 2 *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:*

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

The usefulness of coverage patterns in real world is shown by extracting a set of *coverage patterns* from *MSNBC* dataset [65]. The *coverage patterns* extracted with $minCS = 0.4$, $maxOR = 0.5$ and $minRF = 0.02$ are shown in Table 3.2. All these patterns extracted in the dataset provide 40% coverage which means that if an advertisement is placed on all the web pages in any pattern, atleast 40% of users visiting website view the advertisement.

Table 3.2 Sample coverage 3-patterns extracted from MSNBC dataset.

S.No	Coverage Pattern	CS	S.No	Coverage Pattern	CS
1	$\{local, misc, frontpage\}$	0.42	4	$\{on\text{-}air, news, misc\}$	0.40
2	$\{news, health, frontpage\}$	0.43	5	$\{tech, weather, on\text{-}air\}$	0.41
3	$\{tech, opinion, frontpage\}$	0.41	6	$\{sports, misc, opinion\}$	0.43

3.3 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 [13]. 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. The overview of *CMine* algorithm is as follows.

3.3.1 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.

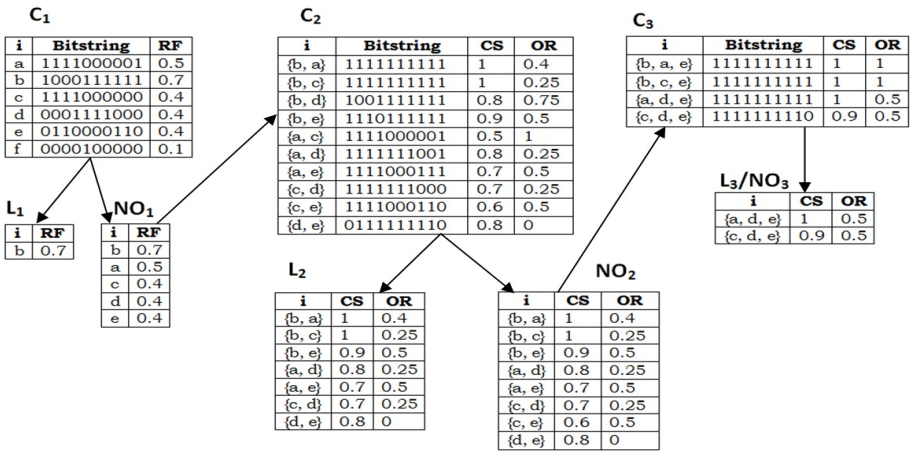

Figure 3.1 Working of CMine algorithm. The term 'I' is an acronym for the item set (or web pages).

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 3.1 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.

Later, a pattern-growth approach, *coverage pattern projected growth (CPPG)* based on the notion of *non-overlap projection* is proposed in [60]. The overview of this approach is briefly explained in section 4.4.1. An enhanced coverage pattern projected growth approach, *ECPPG* is proposed in [63] by using the notion of *upward closure property* of coverage patterns. In the next section, we discuss the missing aspects in the model of *coverage patterns*.

3.4 Summary

In this chapter, we have presented an overview of banner advertisement problem and model of coverage patterns. We also presented the overview of algorithms to extract coverage patterns. In the next chapter, we explain the issues in the model of coverage patterns and further attempts to solve this problem.

Chapter 4

Model of Content-specific Coverage Patterns

In this chapter, we discuss about the model of *content-specific coverage patterns* and a methodology to extract *content-specific coverage patterns*. In the section 4.1, we first present the issues in the model of *coverage patterns*. Next, we explain the motivation behind the concept of *content-specific coverage patterns* and give definitions related to the model in section 4.2. Next, we explain the problem statement and two approaches for finding the number of users interested in particular topic of web pages in a pattern in section 4.3. Next, the overview of *coverage pattern projected growth (CPPG)* [60] and an algorithm to extract *content-specific coverage patterns* are presented in section 4.4. Later, the experimental results are presented in section 4.5 by comparing the present algorithm to the previous *coverage pattern* extraction approach. Finally, we present the summary of this chapter in section 4.6.

4.1 Issues in the model of Coverage Patterns

The model of coverage patterns proposed in [13, 60, 63] provides flexibility to the publisher in providing multiple advertising options to the advertisers. But an advertiser will be more interested to place his advertisement on web pages that relates more to his advertisement. This in turn will increase his probability of attracting more relevant customers for his product. Even though the model of coverage patterns proposed is a good first step, it captures only the coverage of the pattern. If a banner advertisement is not related to at least one of the topics of the web page, users visiting the web page may find it uninteresting which in turn conflicts the interests of advertisers as their aim is to increase the reach of the advertisement to more visitors. In this thesis, we have proposed the model of *content specific coverage patterns* which captures the relevance of the content of the advertisement to that of a *coverage pattern* along with the coverage requirements.

4.2 Motivation

The concept of *coverage patterns* discussed in Chapter 1 is helpful in solving the banner advertisement placement. The model of *coverage patterns* provides flexibility to the publisher in providing

multiple advertising options to the advertisers. But an advertiser will be more interested to place his advertisement on web pages that relates more to his advertisement. This in turn increases the probability of attracting more relevant customers for his products. Therefore, advertiser have to assigned *coverage patterns* that are relevant to the banner advertisement. In order to solve this problem, we need to capture the topics/keywords of web pages in *coverage pattern* so that we can assign the advertisement to the relevant *coverage pattern* which means that the advertisement is to be placed on the corresponding web pages of *coverage pattern*. On the other hand, it is also important to know the number of users interested in a each topic, in the topic set of pattern because more number of users interested in particular topic leads to increase in the probability of clicking advertisement related to that topic. By capturing relevance of the topics of banner advertisement to the coverage pattern along with the coverage requirements, the proposed model of *content-specific coverage patterns* satisfies the requirements of publisher and advertiser. *Content-specific coverage pattern* is a set of web pages visited by certain percentage of users interested in particular content of the web pages. In this chapter, first we propose approaches for computing the fraction of users interested in particular topic of web pages in pattern. Next, by using these approaches and integrating into the model of *coverage patterns*, we propose a model of *content-specific coverage patterns* and an algorithm for extracting these patterns. In this thesis, we used *content-specific* knowledge patterns in order to solve the problem of banner advertisement placement. However, these patterns can be extracted from any transactional dataset provided categories of all items in the dataset.

4.3 Approaches for finding fraction of users

We need to capture the topics related to the web pages in *coverage pattern* to place a relevant advertisement on the corresponding web pages. We assume that every web page is labelled with all the related topics. As discussed in Section 2.1.2, there are approaches in literature for extracting advertising keywords from web pages for *contextual advertising*. We now define the term *TopicSet* to capture the related topics of web pages in the pattern. If the set of all topic labels related to web page w_i is denoted as $Topics(w_i)$, then $TopicSet(X)$ is defined as follows.

Definition 4 *(Topic set of a pattern, $X = \{w_p, \cdots, w_q, w_r\}$, $1 \leq p \leq q \leq r \leq n$). The set of relevant topic labels of all web pages in the pattern is the topic set of a pattern and is denoted by $TopicSet(X)$.*

$$TopicSet(X) = \{Topics(w_p) \cup \cdots \cup Topics(w_q) \cup Topics(w_r)\} \tag{4.1}$$

Example for computing the *TopicSet* of a pattern is given as follows.

Example 3 *From Table 4.1, the topic set of the* coverage pattern $\{a, b\}$ *is* $TopicSet(\{a, b\}) = \{football, cricket\} \cup \{football, rugby\} = \{football, cricket, rugby\}$. *If pattern $\{a, b\}$ is given to the user who is interested in displaying advertisement of topic $hockey$, then less number of users click the advertisement. This is because pattern $\{a, b\}$ is related to $football$, $cricket$ and $rugby$.*

Table 4.1 Topic Labels of web pages, L

Web Pages	a	b	c	d	e	f
Topic Labels	football, cricket	football, rugby	hockey	football	cricket, hockey	cricket, hockey

Next, we need to compute the fraction of users interested in every topic related to *coverage pattern* as the set of topics related to the pattern are known. Later, the model of *content-specific coverage patterns* is proposed by integrating this to the concept of *coverage patterns*. The problem statement for computing the fraction of users is given as follows.

Problem Statement

Given a transactional database D, set of web pages W, topic labels of web pages L and a pattern X, compute the fraction of users interested in every topic in *TopicSet* of pattern.

The approaches for computing the fraction of users interested in every topic of pattern are as follows.

4.3.1 Computing Topic Coverage

Now, we define the number of users visited at least one of the web pages related to the topic in the pattern. It means that if a banner advertisement, related to a particular topic is placed on all web pages of the pattern, how many users find the advertisement as interesting. We capture this notion using *topic coverage set* which is denoted as $TPCSet$.

Definition 5 *(Topic Coverage Set of topic tp w.r.t pattern $X = \{w_p, \cdots w_q, w_r\}$). It is defined as the number of users visited atleast one of the web pages in the pattern which are related to the topic. It is computed by the union of set of TIDs of web pages related to tp and is denoted as $TPCSet(tp)$ w.r.t X.*

$$TPCSet(tp)\ w.r.t\ X = U\ (T^{w_i})\ (\forall\ w_i \in X\ \wedge\ tp \in TopicSet(w_i)) \tag{4.2}$$

Example 4 *From Example 3, $TopicSet(\{a, b\})$ = $\{football, cricket, rugby\}$. Now, $TPCSet(football)$ = $T^a \cup T^b$ = $\{T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8, T_9, T_{10}\}$, $TPCSet(cricket)$ = T^a = $\{T_1, T_2, T_3, T_4, T_{10}\}$ and $TPCSet(rugby)$ = T^b = $\{T_1, T_5, T_6, T_7, T_8, T_9, T_{10}\}$.*

From Example 4, $TPCSet(football)$ with respect to the pattern $\{a, b\}$ denotes that 10 users are interested in the advertisement related to $football$ if the advertisement related to $football$ is placed on the web pages 'a' and 'b'. As *topic coverage set* is an absolute measure, we capture the fraction of users interested in a particular topic of a pattern through *topic coverage support*.

Definition 6 *(Topic Coverage Support topic tp w.r.t pattern $X = \{w_p, \cdots w_q, w_r\}$). It is defined as the fraction of users visited atleast one of the web pages which are related to the topic. It is the ratio of size of $TPCSet(tp)$ to the size of D and is denoted as $TPCS(tp)$ w.r.t X.*

$$TPCS(tp) = \frac{|TPCSet(tp)\ w.r.t\ X|}{|D|}$$

Example 5 *From Example 3 and Example 4, $TopicSet(\{a, b\})$ = $\{football, cricket, rugby\}$. Now, $TPCS(football)$ = $\frac{|T^a \cup T^b|}{|D|}$ = $\frac{10}{10}$ = 1, $TPCS(cricket)$ = $\frac{|T^a|}{|D|}$ = $\frac{5}{10}$ = 0.5 and $TPCSet(rugby)$ = T^b = $\{T_1, T_5, T_6, T_7, T_8, T_9, T_{10}\}$.*

From Example 5, $TPCS(cricket)$ denotes that 50% of users visiting the website finds the advertisement related to $cricket$ as interesting if the advertisement is placed on the web pages 'a' and 'b'.

4.3.1.1 Algorithm

The algorithm for computing the *topic coverage set* of a topic, tp with respect to pattern X is given in Figure 4.1. Here, D denotes the click-through dataset, L denotes topic label database of web pages, tp refers to topic, X refers to the pattern, TID refers to transaction id, $tpcset$ refers to *topic coverage set* and $tpsynset$ refers to the set of synonyms related to the topic tp. In the first step, Line 1 we initialize $tpcset$ as empty. Every web page is tagged with related topic labels and for computing *topic coverage set*, we need to find if a topic is related to web page. A topic tp is related to web page w if tp is similar to any of the topics in web page w_i. Next, we traverse over all web pages in the pattern X at Line 3 and traverse all the topics in topic set of web page w_i at Line 4 to find if any topic in topic set is similar to the given topic tp and add all the $TID's$ containing w_i (Line 7-9). Computing the similarity between topics is a challenging task, if the topics are semantically similar but the keywords do not match exactly. To solve this problem, expand each topic to a set of synonyms (synsets) using word net [66] at Line 2 and Line 5. Path similarity between synsets computes a score between 0 to 1 which denoting the similarity of word senses in a relationship taxonomy. $similarity(synset_1, synset_2)$ is the maximum of all similarities between every possible context of $synset_1$ and $synset_2$. Two synsets are considered similar if the similarity score is greater than specified minimum similarity threshold $minSim$. In the Example 4, the *topic coverage set* is easier to compute as the topics related to the web pages are similar to the input topic.

If an advertiser specifies the topics of banner advertisement, a pattern is interesting if it has high CS, low OR and high $TPCS$ for all the topics specified by advertiser. In this approach, if a user visits a web page related to multiple topics, it is assumed that the user is interested in all the topics of the web page.

Relevance of input topics of banner advertisement as given by advertiser to the *topic set* of the pattern is computed by cosine similarity. Both input topics and topic set of pattern are expanded using

Figure 4.1 Topic Coverage Set Extraction Algorithm

Input: Transactional database D, topic label database L, topic tp, Pattern X
Output: Topic coverage set of tp
Method:

```
1: Set tpcset to empty
2: find tpsynset for tp using wordnet
3: for w_i in X do
4:    for tp_i in TopicSet(w_i) do
5:       find tpsynset_i for tp_i using wordnet
6:       if similarity(tpsynset, tpsynset_i) > minSim then
7:          for TID in T^{w_i} do
8:             add TID to tpcset
9:          end for
10:      end if
11:   end for
12: end for
```

wordnet because exact matching of keywords may not be possible. If the input topic vector is denoted as *InTopicVector* and topic set vector denoted as *TopicSetVector*, the relevance score of pattern X is computed as follows.

$$Relevance(X) = Cosine(InTopicVector, TopicSetVector) \tag{4.3}$$

Example 6 *Web pages 'a' and 'b' with topic labels are shown in Figure 4.2. From Example 3, TopicSet({a,b}) = {football, cricket, rugby}. Let {football, cricket} be the input topics of advertisement. By computing the cosine similarity between the vectors* $(1, 1, 1)$ *and* $(1, 1, 0)$*, relevance of* X *is as follows.*

$$Relevance(X) = Cosine((1, 1, 1), (1, 1, 0)) = \frac{2}{\sqrt{6}}$$

A web page contains content related to multiple topics and different users visit a web page with different topical interests. The visitors of a web page may not be interested in all the topics of the web page. This shows that actual number of visitors interested in a particular topic of a web page is not same as the total number of visitors of web page. In the next subsection, we compute the actual number of visitors interested in specific topics of pattern.

4.3.2 Computing Factual Topic Coverage

In this section, first we give basic idea and next an algorithm to compute the actual number of visitors interested in a specific topic of the topic set of pattern.

Figure 4.2 Web pages having content related to multiple topics

4.3.2.1 Basic Idea

Figure 4.3 Virtual web pages constituting the web pages having content related to multiple topics

If a web page is virtually divided into multiple portions, related to specific topic, the actual number of people interested in every topic of page can be used in placing the advertisement related to specific topics. For example we can see the virtual page divisions of web pages 'a' and 'b' in Figure 4.3. In this approach, if a pattern is assigned to the advertiser, advertisement is placed on the relevant portions of the web page. Hence, every web page related to the given topics is assumed to be a *virtual page* having all of its content related to that topic. If a web page has multiple relevant portions, advertisement can be placed on any relevant portion. Now, *virtual topic set* is formed by union of topic sets of all virtual pages in the pattern. The relevance of given topics with respect to the pattern is the cosine similarity of given topics to the virtual topic set of the pattern. Here, *InTopicVector* and *VTopicSetVector* denotes vector of input topics as specified by advertiser and topic set vector of pattern X. If the virtual topic set of pattern and input topics are expanded using word net, the relevance score of pattern is computed as follows.

$$Relevance(X) = Cosine(InTopicVector, VTopicSetVector) \tag{4.4}$$

Example 7 $TopicSet(\{a, b\})$ = {*football, cricket, rugby*}. *If advertisement is related to the topics* $\{football, cricket\}$*, virtual topic set of* $\{a\}$ *is* $\{football, cricket\}$ *and advertisement can be placed in any portion of 'a'. virtual topic set of* $\{b\}$ *is* {*football*} *as only 'football' portion of 'b' is related to the input topics. Therefore,* $VTopicSet(\{a, b\}) = \{football, cricket\} \cup \{football\}$.

$$Relevance(\{a, b\}) = Cosine((1, 1), (1, 1)) = 1$$

From Example 6 and Example 7, we can observe that relevance of pattern {a, b} is increased. Hence, it is useful to compute the actual number of visitors interested in a particular topic of pattern. The basic idea of finding the actual number of users interested in particular topic of page is as follows. A transaction in a click-through data is a sequence of web pages. An example transaction is shown in Figure 4.4. Let the sequence of web pages visited by a user u_1 be (w_p, w_q, w_r). User visits w_q only after visiting w_p. If u_1 is interested in some topic in w_p, he will explore for that topic in w_q. This implies that the user u_1 is actually visiting for some topic in w_q, if that topic is in w_p. For the web page w_p occurring at the beginning of the transaction, it is assumed that user is interested in all the topics of the web page w_p. For a web page w_i tagged with set of topics, the number of users actually interested in a topic of the web page is the number of transactions having web page related to the topic and penultimate to w_i.

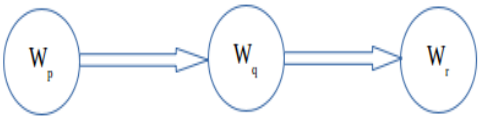

Figure 4.4 Transaction showing sequence of web pages

4.3.2.2 Definitions

We define the notion of how many users visiting at least one web page in the pattern and are actually interested in a topic of the web page. It means that if an advertisement related to particular topic is placed on all web pages of the pattern, how many users actually find the advertisement as interesting. We capture this notion using *factual topic coverage set* which is denoted as $FTPCSet$.

Definition 7 *(Factual Topic Coverage Set of topic* tp *w.r.t pattern* $X = \{w_p, \cdots w_q, w_r\}$*). It is defined as the actual number of users visited atleast one of web pages in the pattern which are related to the topic. It is the set of TID's containing atleast one web page at the beginning of the transaction or web page having topic similar to* tp *penultimate to atleast one web page of pattern and is denoted as*

$FTPCSet(tp)$. *Here,* w_k *refers to the the web page occurring before any of the web page in the pattern.*

$$FTPCSet(tp)\ w.r.t\ X = U\ (T^{w_i})\ \forall\ w_i \in X\ and\ tp \in w_i, w_k \tag{4.5}$$

Example 8 *From Table 3.1 and Table 4.1, TopicSet(*$\{a\}$*) =* $\{$*football, cricket*$\}$*. Number of users visiting 'a' is* 5 *and the TIDs are* $\{T_1, T_2, T_3, T_4, T_{10}\}$*. Users visiting 'a' at the beginning of the transaction may be interested in all the topics of the 'a'. As 'b' is accessed before 'a' in* T_1 *and* T_{10} *and 'b' is related to 'football', the set of transactions actually accessing 'a' for 'football' is* $FTPCSet(football)\ w.r.t\ \{a\} = \{T_1, T_2, T_3, T_4, T_{10}\}$*. As 'b' is not related to 'cricket', the set of transactions actually accessing 'a' for 'cricket' is* $FTPCSet(cricket)\ w.r.t\ \{a\} = \{T_2, T_3, T_4\}$.

From Example 8, 'a' is related to $football$ and $cricket$. Here, $FTPCSet(football)$ with respect to pattern $\{a\}$ denotes that 5 users are actually interested in the advertisement related to $football$ if the advertisement related to $football$ is placed on web page 'a'. $FTPCSet(cricket)$ denotes that 3 users are actually interested in the advertisement related to $cricket$. This implies that of all 5 users visiting 'a', 3 users are actually interested in both $football$ and $cricket$, but the other 2 users are interested only in $football$. As *factual topic coverage set* is an absolute measure, we define the fraction of users interested in a particular topic of a pattern through *factual topic coverage support*.

Definition 8 *(Factual Topic Coverage Support of topic* tp *w.r.t pattern* $X = \{w_p, \cdots w_q, w_r\}$*). It is defined as the fraction of actual number of users visited atleast one of web pages in the pattern which are related to the topic. It is ratio of size of* $FTPCSet(tp)$ *to the size of* D *and is denoted as* $FTPCS(tp)$.

$$FTPCS(tp)\ w.r.t\ X = \frac{|FTPCSet(tp)\ w.r.t\ X|}{|D|} \tag{4.6}$$

Example 9 *From Example 8, Table 4.1, 3.1, TopicSet(*$\{a\}$*) =* $\{$*football, cricket*$\}$*. Now,* $FTPCS(football) = \frac{|\{T_1,T_2,T_3,T_4,T_{10}\}|}{|D|} = \frac{5}{10} = 0.5$ *and* $FTPCS(cricket) = \frac{|\{T_2,T_3,T_4\}|}{|D|} = \frac{3}{10} = 0.3$

From Example 9, $FTPCS(cricket)$ denotes 30% of users visiting the website actually finds the advertisement related to $cricket$ as interesting if the advertisement is placed on the web page 'a'.

4.3.2.3 Algorithm

The algorithm for computing *factual topic coverage set* of a topic tp with respect to pattern X is given in Figure 4.5. Here, D denotes the click-through dataset, L denotes topic label database of web pages, tp refers to topic, X refers to the pattern, TID refers to transaction id, $ftpcset$ refers to *factual topic coverage set* and $ftpsynset$ refers to the set of synonyms related to the topic tp. In the first step, Line 1 we initialize $ftpcset$ as empty. Next, we traverse over all the web pages in the pattern X at Line

Figure 4.5 Factual Topic Coverage Set Extraction Algorithm

Input: Transactional database D, topic label database L, topic tp, Pattern X
Output: Factual topic coverage set
Method:

```
1:  Set tpcset to empty
2:  for w_i in X do
3:     for TID in T^{w_i} do
4:        Find w_k occurring before w_i in TID
5:        if w_i is at the beginning of TID then
6:           add TID to tpcset
7:        else if ∃w_k then
8:           find tpsynset for tp using wordnet
9:           for tp_i in TopicSet(w_k) do
10:             find tpsynset_i for tp_i using wordnet
11:             if similarity(tpsynset, tpsynset_i) > minSim then
12:                add TID to tpcset
13:             end if
14:          end for
15:       end if
16:    end for
17: end for
```

2 and traverse all the transactions $TIDs$ containing corresponding web page w_i at Line 3. Next, find the web page w_k visited before w_i for every transaction TID at Line 4 and add TID to $ftpcset$ if w_i visited at the starting of every transaction at Line 5-6. Next, check if the given topic tp is related to any of topics in topic set of w_k at Line 9-14 and add corresponding transaction TID to $ftpcset$ at Line 12. Similarity between topics is computed based on the similarity discussed in Section 4.3.1.1.

If an advertiser specifies the set of topics of banner advertisement, the pattern is interesting if it has high CS, low OR and high $FTPCS$ for all topics specified by the advertiser. This helps in showing advertisement to the users actually interested in the topics of advertisement.

4.4 Mining Content-specific Coverage Patterns

Now, we define the concept of *content-specific coverage pattern* and problem statement.

Definition 9 *(Content-specific coverage pattern). A pattern X is said to be a content-specific coverage pattern if $CS(X) \geq minCS$, $OR(X) \leq maxOR$, $RF(w_i) \geq minRF$, $\forall w_i \in X$, all w_i are related*

to any of the topics tp_i specified by the advertiser and for $X, FTPCS(tp_i) > minFTPCS_i$. Here, $minFTPCS_i$ denotes the minimum factual topic coverage support *from the topic tp_i.*

Example 10 *Consider the transactional database, Table 3.1 and topic label databse, Table 4.1 with $minRF = 0.4$, $minCS = 0.7$ and $maxOR = 0.5$. Consider an advertiser is interested in coverage patterns related to topics 'football' and 'cricket' with minimum topic coverage, $minFTPCS$ as 0.42 and 0.28 respectively. Consider the pattern {b, a}. From Example 2, $RF(b) = 0.7$, $RF(a) = 0.5$, $CS(\{b, a\}) = 1$ and $OR(\{b, a\}) = 0.2$. $FTPCS(football)$ and $FTPCS(cricket)$ with respect to {b, a} is 1 and 0.3 respectively. The pattern {b, a} is* content-specific coverage pattern *because $RF(a) \geq minRF$, $RF(b) \geq minRF$, $CS(\{a, b\}) \geq minCS$, $OR(\{a, b\}) \leq maxOR$, $FTPCS(football) \geq minFTPCS(football)$ and $FTPCS(cricket) \geq minFTPCS(cricket)$. This patterns is written as follows.*

$$\{b, a\} \quad [CS = 1,\ OR = 0.1,\ FTPCS(football)\ =\ 1,\ FTPCS(cricket)\ =\ 0.3]$$

Problem Statement

Given a transactional database D, set of web pages W, topic labels of web pages L, minimum relative frequency (minRF), maximum overlap ratio (maxOR), minimum coverage support (minCS) and the corresponding percentage of coverage support from the topic labels of advertisement, extract the complete set of *content-specific coverage patterns*.

In the next section, we present the overview of *coverage pattern projected growth (CPPG)* used to extract *coverage patterns* [60].

4.4.1 Overview of CPPG

The model of CPPG is as follows. *f-list* is the list of all items in the database in the decreasing order of frequencies. For a pattern $X = \{i_1, i_2, \cdots, i_n\}$, a pattern $Y = \{i'_1, i'_2, \cdots, i'_m\}$ is called prefix of X if and only if $i'_j = i_j,\ \forall j \leq m \leq n$. A pattern $Z = \{i_{m+1}, i_{m+2}, \cdots, i_n\}$ is the postfix of pattern X with respect to pattern Y. Consider the pattern $X = \{w_p, \cdots, w_q, w_r\}$, where $1 \leq p \leq r \leq n$ and $|T^{w_p}| \geq \cdots \geq |T^{w_q}| \geq T^{w_r}$. A transaction T, in a database D is said to be a non-overlap transaction with respect to a pattern X if and only if T doesn't have any of the item belonging to X. For example, $T = \{a, b, c\}$ is a non-overlap transaction with respect to pattern $X = \{d, e\}$. A non-overlap projection of a transaction T, with respect to a pattern X, is non-empty if T doesn't have any item belonging to X. It doesn't have any item w_i occurring before w_r in the f-list and the items in the non-overlap projection are ordered with respect to the f-list. For the transaction, $T = \{b, d, f\}$ in Table 3.1, non-overlap projection of T with respect to $X = \{a, c\}$ is $Y = \{d, f\}$. Non-overlap projected database of

a pattern X with respect to a database D, contains the non-overlap projections of all transactions in D with respect to the pattern X. For example, non-overlap projected database of $X = \{a\}$ with respect to D is shown in Table 4.2.

Table 4.2 Projected database with respect to X={a}

TID	Pages	TID	Pages	TID	Pages	TID	Pages	TID	Pages
5	d, f	6	d	7	d	8	e	9	e

4.4.1.1 Algorithm

The algorithm in [60] is as follows. Consider $minRF = 0.4$, $minCS = 0.7$ and $maxOR = 0.5$. First, scan the transactional database D and find length-1 frequent patterns. They are {a:5}, {b:7}, {c:4}, {d:4}, {e:4}. All these patterns are also *non-overlap patterns*. Construct *f-list* using frequencies of items, f-list = (b, a, c, d, e). Next, transactional database is arranged in the order of f-list by removing items that don't satisfy *minRF* as shown in Table 4.3. Now, the set of all *non-overlap patterns* can be divided into 5 subsets having a, b, c, d and e as prefixes. For each item in the f-list construct its corresponding non-overlap projected database on D. If the item satisfies *coverage support*, report it as *coverage pattern*. Now, recursively mine the projected databases as follows to find the *non-overlap patterns* by computing support of each item in the projected database. For every *non-overlap pattern* found, check for its CS and report it as *coverage pattern*. Now, recursively projected the *non-overlap pattern* on its corresponding projected database and continue mining process until the projected database is empty. Now, we show the mining process by considering the prefix 'a'. The mining process of extracting *coverage patterns* with other prefixes can be done similarly. The *non-overlap projected database* of 'a' is shown in the second column of second row in Table 4.3. By scanning this projected database, 2-length non-overlap patterns can be formed. They are $\{a, d\} : \{0.25, 0.8\}$ and $\{a, e\} : \{0.5, 0.7\}$. Here, first and second component represents OR and CS respectively. Now, by recursively considering $\{a, e\}$ as prefix, the projected database is empty and by considering $\{a, d\}$ as prefix, the projected database is $\{\{e\}, \{e\}\}$. The only 3-length non-overlap pattern formed from this projected database is $\{a, d, e\} : \{0.5, 0.1\}$. All *non-overlap patterns* and *coverage patterns* having prefix as 'a' are shown in second row of Table 4.3. The *non-overlap projected databases*, *non-overlap patterns* and *coverage patterns* extracted in this process for the transactional database in Table 3.1 are shown in Table 4.4.

Table 4.3 Database arranged in the order of f-list

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

Table 4.4 Projected databases, non-overlap patterns and *coverage patterns* generated in CPPG

Prefix	Non-overlap projected database	Non-overlap patterns	Coverage Patterns
b	{{a,c,e},{a,c,e},{a,c,d}}	{{b},{b,a},{b,c},{b,e}}	{{b},{b,a},{b,c},{b,e}}
a	{{d},{d},{d},{e},{e}}	{{a},{a,d},{a,e},{a,d,e}}	{{a,d},{a,e},{a,d,e}}
c	{{d},{d},{d},{e},{e}}	{{c},{c,d},{c,e},{c,d,e}}	{{c,d}, {c,d,e}}
d	{{e}, {e}, {e}, {e}}	{{d}, {d,e}}	{{d,e}}
e	empty	{{e}}	empty

4.4.2 Content-specific Coverage Pattern Projected Growth (CSCPPG)

4.4.2.1 Basic Idea

CPPG generates *coverage patterns* by using the notion of *non-overlap pattern projection*. First, *non-overlap projected database* is generated for every frequent item. This partitions both data set and set of non-overlap patterns and confines testing to smaller projected databases. These projected databases are scanned for non-overlap patterns and in turn projected recursively till the projected databases are empty. The non-overlap patterns formed in this process are checked for *coverage support* and *coverage patterns* are generated. A basic approach for extracting *content-specific coverage patterns* would be to extract all *coverage patterns* and compute *topic coverage support* for every topic related to topic set in every coverage pattern generated. Finally, output the patterns which satisfy *topic coverage* constraints along with coverage constraints. This is computationally expensive because the search space is large due to the number of *coverage patterns* are significantly larger when compared to the number of *content-specific coverage patterns*

The process in *CSCPPG* starts with the collection of web pages having atleast one topic related to the topics given by advertiser. For every frequent web page (1-length non-overlap pattern) in this collection, the process of *CPPG* is continued. All the non-overlap patterns formed in this process are checked for *FTPCS* of the topics in pattern given by advertiser and are reported as *content-specific coverage patterns*. The computation of $FTPCS$ of k-length *non-overlap* patterns can be done by reusing the $FTPCS$ of (k-1)-length patterns. The search space is decreased in this process because we are considering web pages having at least one of the topics as specified by advertiser.

4.4.2.2 Algorithm

We explain the Algorithm 4.6 as follows. Scan D, find frequent items and construct f-list (Line-1). Scan L, construct inverted index and find all the items related to atleast one of the input topic labels. Remove all other items from f-list (Line-2). Next, D is arranged in the order of f-list and remove items that are not in f-list (Line-3). For each item in f-list, check for its CS and $FTPCS$ for all input topics by using Algorithm 4.5. If CS is no less than $minCS$ and $FTPCS$ is no less than corresponding $minFTPCS$, report it as the *content-specific coverage pattern* (Lines 5-7). Next, construct its non-overlap projected database on D by calling $ConstructNOP$ method. Now, recursively mine the non-

overlap projected database by calling $CSCPPGRec$ as follows. Find the support count of every item in projected database and for each non-overlap pattern find the CS and $FTPCS$ for every input topic label by using Algorithm 4.5. If the pattern satisfies CS and $FTPCS$ for all input topics, it is reported as *content-specific coverage pattern*. Now, recursively project the database on every *non-overlap pattern* until the projected database is empty.

For the databases in Table 3.1, 4.1 with $minRF = 0.4$, $minCS = 0.7$ and $maxOR = 0.5$, if an advertiser is interested in *coverage patterns* related to the topics {football, cricket} and minimum percentage of coverage from total $minCS$ for topics is $\{60\%, 40\%\}$. Hence, $minFTPCS$ for 'football' and 'cricket' are 0.42 and 0.28 respectively. The extraction of complete set of *content-specific coverage patterns* is explained as follows. Scan D and find the frequencies of items. They are {a:5}, {b:7}, {c:4}, {d:4}, {e:4}, {f:1}. Given, $minRF = 0.4$ frequent items are {b,a,c,d,e}. Hence, *f-list* is (b, a, c, d, e). Scan topic label database, L and find the set of items that contain atleast one topic from {football, cricket}. They are {a, b, d, e, f} are remove items from *f-list* which are not in this set. The f-list formed after removing items not related to the topics is (b, a, d, e). Now, arrange items in the database, D in the order of f-list after removing items not in f-list. The new database formed is given in Table 4.5. Next, we partition the complete set of non-overlap patterns into four subsets each having items in f-list as prefixes.

Now, we explain the extraction of non-overlap patterns with prefix 'b'. CS of pattern {b} is 0.7. Pattern {b} is a CP, but to be a *content-specific coverage pattern* it has to satisfy minimum factual topic coverage support for 'football' and 'cricket'. As 'b' is related to topics 'football' and 'rugby', $FTPCS(cricket) = 0$. From Algorithm 4.5, $FTPCSet(football)$ is calculated as follows. As 'b' is occurring at the beginning of the transaction in TIDs {5, 6, 7, 8, 9}, they are included in $FTPCSet$ of 'football'. In TIDs {1, 10} web page 'a' which is also related to 'football' is accessed before 'b'. This implies $FTPCSet(`football') = \{1, 5, 6, 7, 8, 9, 10\}$. Pattern {b} is not a *content-specific coverage pattern* as it not related to 'cricket'. Now, D is projected with respect to {b} to form the non-overlap projected database shown in Table 4.6. The 2-length non-overlap patterns are {b, a} and {b, e}. Factual topic coverage set of a topic with respect to pattern {b, a} is the union of factual topic coverage sets with respect to items 'a' and 'b'. From Algorithm 4.5, $FTPCS$ of 'football' and 'cricket' are calculated. {b, a}:[1, 0.4] [football:1, cricket:0.5] and {b, e}:[0.9, 0.5] [football:0.7, cricket:0]. pattern:[CS, OR] [topic: FTPCS] denotes the pattern with coverage support, overlap ratio and factual topic coverages of topics. {b, a} is *content-specific coverage pattern* as $FTPCS(`football') > 0.42$ and $FTPCS(`cricket') > 0.28$. Next the database in Table 4.6 is projected with respect to {b, a} and {b, e}, which are empty. Similarly, the projected databases are formed for other items in *f-list*. *Content-specific coverage patterns* extracted in this process are different if *topic coverage* is used in place of *factual topic coverage*. The *content-specific coverage patterns* formed in this process using *TPCS* and *FTPCS* are shown in Table 4.7.

Figure 4.6 Content-Specific Coverage Pattern Projected Growth Method

Input: A transactional database D, topics labels of web pages L, $minRF$, $maxOR$, $minCS$, topic labels of advertisement *t-list* and corresponding $minTPCS$.
Output: Set of all content-specific coverage patterns
Method:

1: Scan D, find all frequent web pages and construct f-list.
2: Scan L, form inverted index, find all web pages related to topics in t-list and retain these in f-list.
3: Arrange D to contain only web pages in f-list and arrange in f-list order.
4: **for** x in f-list **do**
5: **if** $FTPCS(tp_i)|_x \geq minTPCS_i, \forall tp_i \in$ t-list **then**
6: Add x to *content-specific coverage pattern* list
7: **end if**
8: Construct non-overlap projected database of 'x',
$NOP(D)|_x = ConstructNOP(D, x,$ f-list$)$
9: Call $CSCPPGRec(x, l, NOP(D)|_x,$ f-list$)$
10: **end for**

Subroutine: $CSCPPGRec(\alpha, l, NOP(D)|_\alpha),$ f-list)
Parameters: α: non-overlap pattern; l: length of α; $NOP(D)|_\alpha$: non-overlap projected database of α with respect to D
Method:

1: Scan $NOP(D)|_\alpha$, find the set of non-overlap items 'i'.
2: **for** each 'i' **do**
3: Append 'i' to α to form non-overlap pattern α'.
4: **if** $FTPCS(tp_i)|_x \geq minTPCS_i, \forall tp_i \in$ t-list **then**
5: Add α' to *content-specific coverage pattern* list
6: **end if**
7: Construct non-overlap projected database of α',
$NOP(D)|_{\alpha'} = ConstructNOP(D, \alpha',$ f-list$)$
8: Call $CSCPPGRec(\alpha', l, NOP(D)|_x,$ f-list$)$
9: **end for**

Subroutine: $ConstructNOP(D, x, list)$
Parameters: D: transactional database, x: pattern, $list$: f-list
Method:

1: **for** t in D **do**
2: Find non-overlap transaction of t with respect to x.
3: Add to new database D'.
4: **end for**
5: Output D'.

Table 4.5 Database arranged in the order of f-list

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

Table 4.6 Projected database of Table 4.5 with respect to X={b}

TID	**Pages**	**TID**	**Pages**	**TID**	**Pages**
2	a, e	3	a, e	4	a, d

4.4.3 Discussion

In Section 4.4.2, algorithm is proposed based on the pattern-growth approach proposed in [60]. An enhanced approach is proposed in [63] which can also be integrated with *topic coverage* to extract *content-specific coverage patterns*. A pattern is said to be more relevant to advertisement if the *topic set* of the pattern has more topics of advertisement. From the concept of relevance discussed in Equation 4.3 and 4.4, a pattern is said to be totally relevant to advertisement if the *topic set* of pattern has all the topics of advertisement. If an advertiser is interested in displaying the advertisement related to 'football' and 'cricket', from Table 4.4 and 4.8, we can observe that the *coverage patterns* {b}, {b, c} and {c, d} are irrelevant to advertiser. Even though the patterns {d, e}, {c, d, e} are relevant, they are not having enough *topic coverage* to match advertiser's needs. we can also observe that {b, e} is formed in *CSCPPG* using $TPCS$ but not in *CSCPPG* using $FTPCS$ because $FTPCS(\{b, e\}) = 0$. It means that actual number of users visiting 'b' and 'e' interested in 'cricket' is zero. From Table 4.7, we can observe that the set of all *content-specific coverage patterns* are relevant to advertiser. From Table 4.8, each row indicates a coverage pattern and the last two columns indicate whether it is content-specific or not. Here, 'CS' indicates coverage support, 'Topics' indicates the keywords given by the advertiser, 'TPCS' indicates the topic coverage support and 'FTPCS' indicates the factual topic coverage support of topics 'football' and 'cricket'. 'CSCP (TPCS)' and 'CSCP(FTPCS)' indicates content specific coverage patterns. We can observe the *coverage patterns* which are not *content-specific coverage patterns*.

4.5 Experimental Results

In this section, we present the experimental results for extracting *content-specific coverage patterns* and analyse the patterns generated by comparing with the patterns generated in the model of the *coverage patterns*. We have also compared the two approaches proposed *CSCPPG (TPCS)* and *CSCPPG (FTPCS)* by comparing the average relevance of a random sample of patterns to the topics specified by the advertiser.

Table 4.7 Projected databases, non-overlap and content specific coverage patterns generated in CSCPPG

Prefix	Non-overlap projected database	Non-overlap patterns	Content-specific CPs (FTPCS)	Content-specific CPs (TPCS)
b	{{a,e},{a,e},{a,d}}	{{b},{b,a},{b,e}}	{{b,a}}	{{b,a},{b,e}}
a	{{d},{d},{d},{e},{e}}	{{a},{a,d},{a,e},{a,d,e}}	{{a,d},{a,e},{a,d,e}}	{{a,d},{a,e},{a,d,e}}
d	{{e},{e},{e},{e}}	{{d}, {d,e}}	empty	empty
e	empty	{{e}}	empty	empty

4.5.1 Description of Dataset

We have used *delicious dataset* [67], [68] for the extraction of *content-specific coverage patterns*. This dataset contains bookmarking and tagging information from a set of 2000 users of delicious bookmarking system. The user click-through information in this dataset is recorded as <UserID, BookmarkID, TagID, Timestamp>. It means that users tag web pages in the form of 'bookmarks' with corresponding 'tags'. First, we segment each user's tag events into sessions based on generally used rule [69] that if the time interval between simultaneous click events exceeds 30 minutes, they belong to different sessions. Now, we get a sequence of click events. This can be assumed as the sequence of accesses to web pages where each web page is tagged with related tags. This dataset has $69,227$ distinct bookmarks, $53,389$ tags and $15,970$ transactions. The method for computing *factual topic coverage* and *CSCPPG* algorithm are written in *Python* and executed in *Linux* on $2.27GHz$ machine with $2GB$ memory.

4.5.2 Generation of Coverage Patterns and Content-specific Coverage Patterns

The total number of *content-specific coverage pattern*s extracted in *CSCPPG* for an advertiser interested in the topic 'design' with minimum percentage of coverage 20% from total coverage (minCS) using $TPCS$ and $FTPCS$ is shown as *CSCPPG ('design') (TPCS)* and *CSCPPG ('design') (FTPCS)* respectively. Similarly, for two topics 'design' and 'tools' with 20% each from total coverage (minCS), they are represented as *CSCPPG ('design','tools') (TPCS)* and *CSCPPG ('design','tools') (FTPCS)* respectively.

4.5.2.1 Varying Coverage Support

The Figure 4.7 and Figure 4.8 shows the number of coverage patterns on (y-axis) extracted for delicious dataset for different values of $minCS$ (x-axis) while $minRF$ and $maxOR$ are fixed at 0.001 and 0.001 respectively. Figure 4.7 shows the coverage patterns extracted for advertiser specified topic 'design' while Figure 4.8 shows the the coverage patterns extracted for advertiser specified topics 'de-

Table 4.8 Figure showing all coverage patterns in comparison with content-specific coverage patterns with TPCS and FTPCS for topics 'football' and 'cricket'

CP	CS	Topics	TPCS	FTPCS	CSCP(TPCS)	CSCP(FTPCS)
{b}	0.7	football	0.7	0.7	✗	✗
		cricket	0	0		
{b, a}	1	football	1	1	✓	✓
		cricket	0.5	0.3		
{b, c}	1	football	0.7	0.7	✗	✗
		cricket	0	0		
{b, e}	0.9	football	0.7	0.7	✓	✗
		cricket	0.4	0		
{a, d}	0.8	football	0.8	0.8	✓	✓
		cricket	0.5	0.4		
{a, e}	0.7	football	0.5	0.5	✓	✓
		cricket	0.7	0.4		
{a, d, e}	1	football	0.8	0.8	✓	✓
		cricket	0.7	0.4		
{c, d}	0.7	football	0.4	0.3	✗	✗
		cricket	0	0		
{c, d, e}	0.9	football	0.4	0.3	✗	✗
		cricket	0.4	0		
{d, e}	0	football	0.4	0.3	✗	✗
		cricket	0.4	0		

sign' and 'tools'. However, the similar behavior is observed for any topic related to the web pages in the transactional database as given by the advertiser.

In Figure 4.7 and Figure 4.8 the number of *coverage patterns* and *content-specific coverage patterns* extracted decreases with the increase in $minCS$. This is because when the $minCS$ is low, patterns with smaller length satisfies $minCS$ but when the $minCS$ is increased, smaller patterns do not satisfy $minCS$.

From Figure 4.7, the number of *content-specific coverage patterns* generated in *CSCPPG ('design') (FTPCS)* and *CSCPPG ('design') (TPCS)* are significantly smaller than the number of *coverage patterns* extracted in *CPPG* because only some patterns are related to specific topics and satisfies *topic coverage* needs of the advertiser. This difference gradually decreases as $minCS$ increases because *minimum topic coverage support* is fixed at 20% of $minCS$ which leads to increase in the percentage of *content-specific coverage patterns* in the set of *coverage patterns*. Also, the number of *content-specific coverage patterns* generated in *CSCPPG (FTPCS)* is similar to the the patterns generated in *CSCPPG (TPCS)* because the actual number of users interested in single topic 'design' is similar to the users visiting the web pages.

From Figure 4.8, the number of *content-specific coverage patterns* generated in *CSCPPG* for multiple topics 'design' and 'tools' as specified by the advertiser is also significantly smaller than the number of *coverage patterns* generated in *CPPG*. The difference between the number of *coverage patterns* gen-

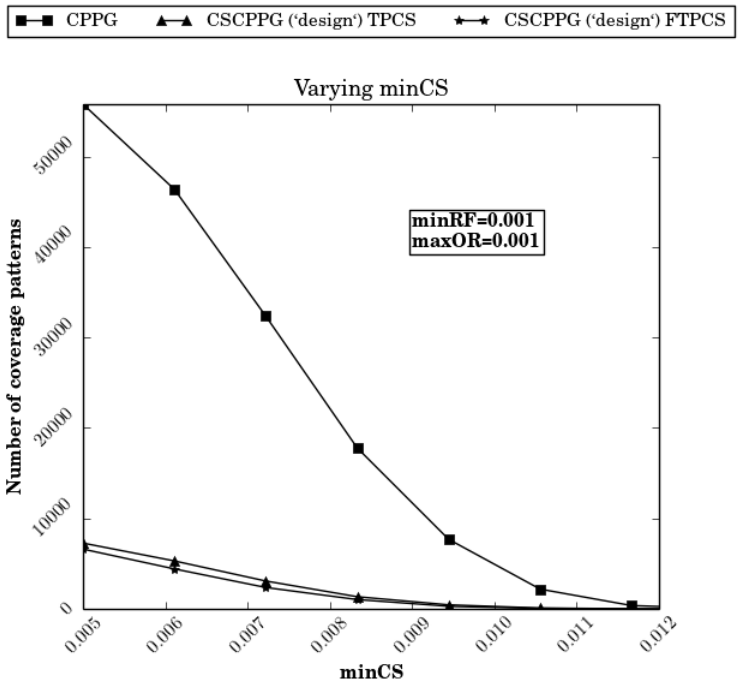

Figure 4.7 Number of content-specific coverage patterns and the total number of coverage patterns (y-axis) generated by varying $minCS$ (x-axis)

erated in *CPPG*, *CSCPPG (TPCS)* and *CSCPPG (FTPCS)* gradually decreases because *minimum topic coverage* is fixed for two topics 'design' and 'tools' as 20%. The number of *content-specific coverage patterns* generated in *CSCPPG (FTPCS)* is lower to the patterns generated in *CSCPPG (TPCS)* because the actual number of users interested in multiple topics is lower than the users visiting web pages. This difference is more significant in Figure 4.8 than in Figure 4.7.

4.5.2.2 Varying Overlap Ratio

The Figure 4.9 and Figure 4.10 shows the number of coverage patterns on (y-axis) extracted for delicious dataset for different values of $maxOR$ (x-axis) while $minRF$ and $minCS$ are fixed at 0.001 and 0.008 respectively. Figure 4.9 shows *coverage patterns* extracted by varying $maxOR$ for single topic 'design' as specified by the advertiser while Figure 4.10 shows the *coverage patterns* generated for two topics 'design' and 'tools'. However, the similar behavior is observed for any number of topics specified by the advertiser.

From Figure 4.9 and Figure 4.10, the number of coverage patterns and *content-specific coverage patterns* generated generally increases with the increase in $maxOR$. This is because when the $maxOR$ is high, more number of patterns can be grouped to form non-overlap patterns. This can be observed when $maxOR$ is greater than 0.04. Here, the total number of coverage patterns and *content-specific cover-*

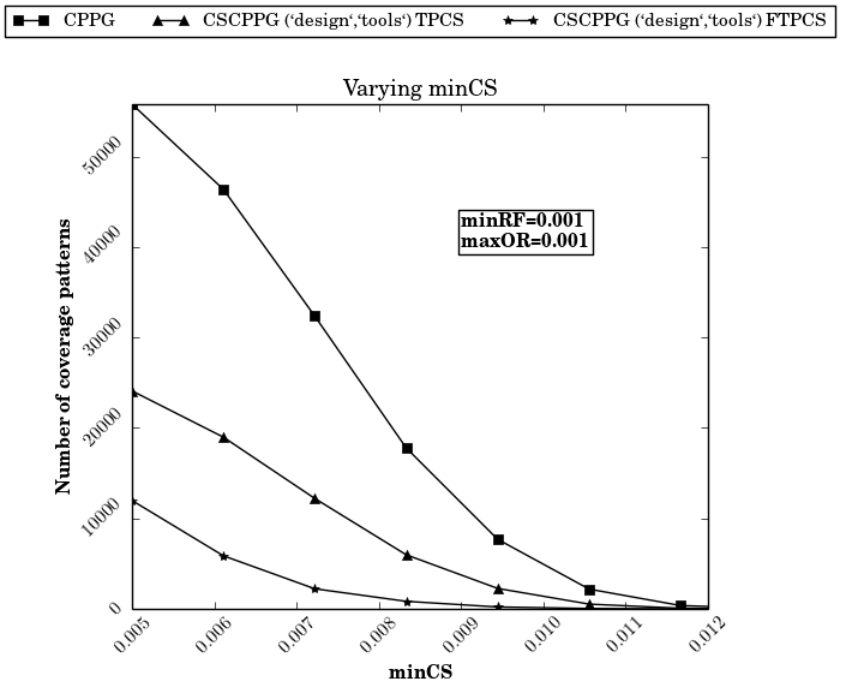

Figure 4.8 Number of content-specific coverage patterns and the total number of coverage patterns (y-axis) generated for multiple topics by varying $minCS$ (x-axis)

age patterns are constant till $maxOR = 0.04$ because more non-overlap patterns formed by grouping patterns is not satisfying $minCS$.

From Figure 4.9, the number of *content-specific coverage patterns* generated in *CSCPPG ('design') (FTPCS)* and *CSCPPG ('design') (TPCS)* are significantly smaller than the number of *coverage patterns* extracted in *CPPG* because only some patterns are related to specific topics and satisfies *topic coverage* needs of the advertiser. This difference remained constant till $maxOR = 0.04$ because $minCS$ is constant and *minimum topic coverage* of topic 'design' is also constant at 20% of $minCS$. When no new non-overlap patterns formed satisfies these coverage constraints, no new *coverage* and *content-specific coverage* patterns can be formed. The difference increased after $minCS \geq 0.04$ because more *non-overlap patterns* will be termed as *coverage patterns* than the *content-specific coverage patterns* as these are restricted by fixed *topic coverage* constraints. Also, the number of *content-specific coverage patterns* in *CSCPPG (TPCS)* and *CSCPPG (FTPCS)* for topic 'design' is similar because the actual number of users interested in the topic 'design' is same as the number of users visiting web pages.

From Figure 4.10, the number of *content-specific coverage patterns* generated in *CSCPPG* for multiple topics 'design' and 'tools' as specified by the advertiser is also significantly smaller than the number of *coverage patterns* generated in *CPPG*. The difference between the number of *coverage patterns* generated in *CPPG*, *CSCPPG (TPCS)* and *CSCPPG (FTPCS)* for multiple topics 'design' and 'tools' is remained constant while varying $maxOR$ till $minOR = 0.04$. This is because $minCS$ is constant

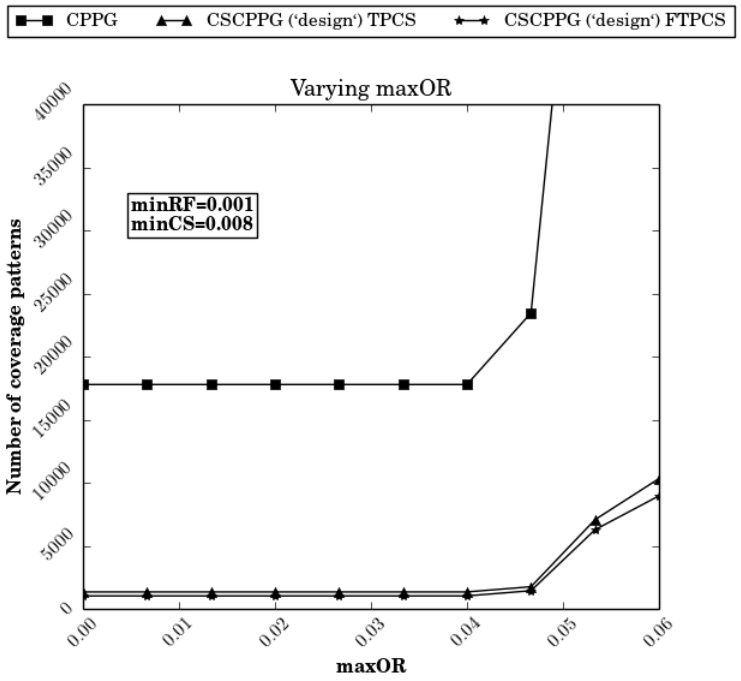

Figure 4.9 Number of content-specific coverage patterns and the total number of coverage patterns (y-axis) generated by varying $maxOR$ (x-axis)

and *minimum topic coverage* of two topics 'design' and 'tools' is also constant at 20% of $minCS$ due to no new non-overlap patterns generated satisfies these topic coverage requirements. Also, the number of *content-specific coverage patterns* in *CSCPPG (FTPCS)* is lower than the patterns generated in *CSCPPG (FTPCS)* for topics 'design' and 'tools' because the actual number of users interested in the topics 'design' and 'tools' is lower than the number of users visiting web pages. This difference is more significant in Figure 4.10 than in Figure 4.9

4.5.3 Comparison of Relevance of Patterns

Average relevance for a random sample of 10 patterns is computed for $CPPG$ and *CSCPPG (TPCS)* approaches using Equation 4.3. Equation 4.4 is used for computing relevance of patterns in *CSCPPG (FTPCS)* approach. Here, $minRF = 0.001$, $maxOR = 0.001$ and $minCS = 0.005$. The average relevance is in order of 10^{-2} because most of the patterns have atleast one web page related to 'web' and every pattern has more than 100 tags. From Table 4.9, we can observe that average relevance of patterns to input topics in *CSCPPG (TPCS)* is more than in *CPPG* because in *CSCPPG* every web page of pattern is atleast related to one of given topics. From Equation 4.4, average relevance of patterns in *CSCPPG (FTPCS)* is 1 because every web page in the pattern generated have atleast one topic from the set of input topics. The similar behaviour was observed for any input set of topics.

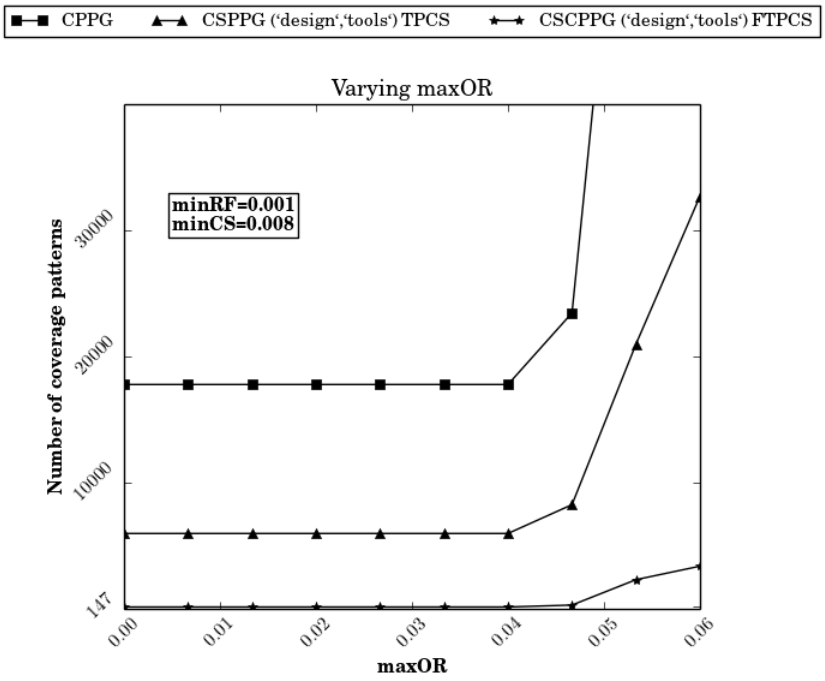

Figure 4.10 Number of content-specific coverage patterns and the total number of coverage patterns (y-axis) generated for multiple topics by varying $maxOR$ (x-axis)

Table 4.9 Average relevance for a sample in CPPG, CSCPPG (TPCS) and CSCPPG (FTPCS)

Topics	CPPG	CSCPPG (TPCS)	CSCPPG (FTPCS)
web	0.02	0.08	1

4.6 Summary

In this chapter, first we presented the importance of *content-specific coverage patterns* in banner advertisement placement as the present model of *coverage patterns* is not capturing the relevance of advertisement to the content of web pages. We also presented the importance of computing the fraction of users interested in particular topics of web pages to integrate into the model of *content-specific coverage patterns*. We have defined terms *topic coverage* and *factual topic coverage* and proposed approaches to compute for a particular topic with respect to a particular pattern. Factual topic coverage is computed based assuming the user visiting behavior as markov process. We showed that the *factual topic coverage* is more accurate measure over the *topic coverage*.

Next, we defined the concept of *content-specific coverage pattern* and presented pattern growth approach for mining these patterns by integrating the concepts of *topic coverage* into the model of *content-specific coverage patterns*. Later, we have shown the present model of *content-specific coverage patterns* gives only patterns that are relevant to the topics of interest of the advertiser. Experiments done

on the *delicious* dataset showed that lot of irrelevant coverage patterns generated in model of *coverage patterns* are not generated in this model. It has been observed that there has been a significant decrease in the number of *coverage patterns* by generating only relevant *coverage patterns*. Finally, we present the summary of the thesis and future work in the next chapter.

Chapter 5

Conclusion and Future Work

In this chapter, first we present the summary of thesis followed by conclusions made in the thesis. Finally, we present the scope of future work in this direction of research.

We present the summary of the thesis as follows. Online advertising is an important mode of advertising. Banner advertising plays an important role in online advertising. The knowledge of *coverage patterns* is helpful in banner advertisement placement by improving the options offered by publisher to the advertiser. *Coverage patterns* are extracted from the click-through data of a website for a certain period of time by assuming similar visitors behavior. However, the *coverage patterns* generated are not much helpful to the advertiser because if the advertisement is not related to the content of the web pages, the probability of clicking advertisement decreases. This may lead to annoyance of visitors which leads to decrease in click-through rate of advertisement. In this thesis, we have conceptualised *topic set* which captures the topics related to all web pages in the pattern. In literature, there are many approaches proposed to extract topics from the web pages. However, we use only the topics generated for the web page instead of focusing on approaches to extract these keywords. Web pages containing topics is not much helpful for banner advertisement placement with out having the statistics of number of users visited corresponding web pages for every related topic.

An attempt is made in this thesis to capture the number of users interested in a particular topic of web pages in the pattern by assuming the process of users visiting web pages is a markov process. It is known that more number of users interested in a particular topic on a web page leads to increase in the probability of clicking an advertisement related to that topic. It is important to note that we are trying to capture the relevance of advertisement to the web page on the basis of number of users visiting the web page for the corresponding topic by having only topic labels of web pages instead of total content of the web pages. We have conceptualised the terms *topic coverage* and *factual topic coverage* to discover the number of users interested in particular topic of web pages in the pattern. Later, by integrating these concepts we have proposed the model of *content-specific coverage patterns* and proposed an approach, *CSCPPG* to extract these patterns.

The proposed approaches for computing *topic coverage* and *factual topic coverage* are useful in extracting *content-specific coverage patterns*. Also, the proposed approaches content-specific coverage

pattern projected growth, *CSCPPG* using both *topic coverage support* and *factual topic coverage support* extracts relevant set of patterns efficiently over the pattens generated in *CPPG*, coverage pattern projected growth method. We have demonstrated the above fact by conducting experiments on *delicious* dataset. We have also demonstrated that by bidding for slots on the web pages of content-specific coverage patterns, advertisers will have relevant set of expected number of visitors.

In the proposed approach, the actual number of visitors interested in particular content of a web page are found by examining the content of the web pages within the immediate neighbourhood of this page in click-through data. As a part of future work, we are investigating on developing a feature which idealises the neighbourhood for maximizing the relevance of topic with the content of web pages in the pattern. The model of *coverage patterns* and *content-specific coverage patterns* provides multiple options to advertiser for placing their advertisement, but if we have multiple advertisers with different topic coverage requirements, there would be problem of efficient allocation of *coverage patterns* by maximizing revenue to the publisher. We need to solve an optimization problem for scheduling multiple *coverage patterns* to different advertisers by satisfying their interests along with maximizing revenue of the publisher. We are planning to investigate on using both *frequent* and *coverage* pattern knowledge in banner advertisement placement. We are also planning to investigate on applications of *content-specific coverage patterns* in the field of bio-informatics for extracting potential knowledge patterns which can be further used in drug discovery process.

Chapter 6

Publications

6.1 Related Publications

(I) Trinath, A. V., P. Gowtham Srinivas, and P. Krishna Reddy. "Content specific coverage patterns for banner advertisement placement." In Data Science and Advanced Analytics (DSAA), 2014 International Conference on, pp. 263-269. IEEE, 2014.

6.2 Other Publications

(I) Srinivas, P. Gowtham, P. Krishna Reddy, A. V. Trinath, S. Bhargav, and R. Uday Kiran. "Mining coverage patterns from transactional databases." Journal of Intelligent Information Systems (2014): 1-17.

(II) Reddy, P. Krishna, A. V. Trinath, M. Kumaraswamy, B. Bhaskar Reddy, K. Nagarani, D. Raji Reddy, G. Sreenivas et al. "Development of eAgromet prototype to improve the performance of integrated agromet advisory service." In Databases in Networked Information Systems, pp. 168-188. Springer International Publishing, 2014.

Bibliography

[1] Iab internet advertising revenue report. 2013. URL $http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_FY_2013.pdf$.

[2] Daniel C. Fain and Jan O. Pedersen. Sponsored search: A brief history. *Bulletin of the American Society for Information Science and Technology*, 32(2):12–13, 2006. ISSN 1550-8366. doi: 10.1002/bult.1720320206. URL http://dx.doi.org/10.1002/bult.1720320206.

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

[4] Berthier Ribeiro-Neto, Marco Cristo, Paulo B. Golgher, and Edleno Silva de Moura. Impedance coupling in content-targeted advertising. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 496–503, New York, NY, USA, 2005. ACM. ISBN 1-59593-034-5. doi: 10.1145/1076034.1076119. URL http://doi.acm.org/10.1145/1076034.1076119.

[5] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 261–270, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526745. URL http://doi.acm.org/10.1145/1526709.1526745.

[6] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, August 1988.

[7] Ali Amiri and Syam Menon. Efficient scheduling of internet banner advertisements. *ACM Trans. Internet Technol.*, 3(4):334–346, November 2003. ISSN 1533-5399. doi: 10.1145/945846.945848. URL http://doi.acm.org/10.1145/945846.945848.

[8] Arpita Ghosh, Preston Mcafee, Kishore Papineni, and Sergei Vassilvitskii. Bidding for representative allocations for display advertising. In *In Fifth Workshop on Internet and Netowork Economics*, pages 208–219, 2009.

[9] Alex Rogers, Esther David, Terry R. Payne, and Nicholas R. Jennings. An advanced bidding agent for advertisement selection on public displays. In *Proceedings of the 6th International Joint*

Conference on Autonomous Agents and Multiagent Systems, AAMAS '07, pages 51:1–51:8. ACM, 2007. ISBN 978-81-904262-7-5.

[10] Mohammad Mahdian and Kerem Tomak. Pay-per-action model for online advertising. In *Proceedings of the 1st International Workshop on Data Mining and Audience Intelligence for Advertising*, ADKDD '07, pages 1–6. ACM, 2007.

[11] Hamid Nazerzadeh, Amin Saberi, and Rakesh Vohra. Dynamic cost-per-action mechanisms and applications to online advertising. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 179–188, New York, NY, USA, 2008. ACM.

[12] Bhargav Sripada, Krishna Reddy Polepalli, and Uday Kiran Rage. Coverage patterns for efficient banner advertisement placement. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 131–132, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0637-9. doi: 10.1145/1963192.1963259. URL `http://doi.acm.org/10.1145/1963192.1963259`.

[13] P. Gowtham Srinivas, P. Krishna Reddy, Bhargav Sripada, R. Uday Kiran, and D. Satheesh Kumar. Discovering coverage patterns for banner advertisement placement. In *PAKDD (2)*, pages 133–144, 2012.

[14] Sheldon M. Ross. *Introduction to Probability Models, Ninth Edition*. Academic Press, Inc., 2006.

[15] Subodha Kumar, Varghese S Jacob, and Chelliah Sriskandarajah. Scheduling advertisements on a web page to maximize revenue. *European journal of operational research*, 173(3):1067–1089, 2006.

[16] Google adwords tool. 2014. URL `https://adwords.google.com/`.

[17] Rex Briggs and Nigel Hollis. Advertising on the web: Is there response before click-through? *Journal of Advertising research*, 37(2):33–45, 1997.

[18] Deepak Agarwal, Andrei Zary Broder, Deepayan Chakrabarti, Dejan Diklic, Vanja Josifovski, and Mayssam Sayyadian. Estimating rates of rare events at multiple resolutions. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–25. ACM, 2007.

[19] Haibin Cheng and Erick Cantú-Paz. Personalized click prediction in sponsored search. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 351–360. ACM, 2010.

[20] Kushal S Dave and Vasudeva Varma. Learning the click-through rate for rare/new ads from similar ads. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 897–898. ACM, 2010.

[21] Krzysztof Dembczynski, Wojciech Kotlowski, and Dawid Weiss. Predicting ads click-through rate with decision rules. In *Workshop on targeting and ranking in online advertising*, volume 2008, 2008.

[22] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530. ACM, 2007.

[23] Thore Graepel, Joaquin Q Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 13–20, 2010.

[24] Aditya Krishna Menon, Krishna-Prasad Chitrapura, Sachin Garg, Deepak Agarwal, and Nagaraj Kota. Response prediction using collaborative filtering with hierarchies and side-information. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 141–149, New York, NY, USA, 2011. ACM.

[25] Azin Ashkan and Charles LA Clarke. Characterizing commercial intent. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 67–76. ACM, 2009.

[26] Azin Ashkan and Charles LA Clarke. Modeling browsing behavior for click analysis in sponsored search. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2015–2019. ACM, 2012.

[27] Dawei Yin, Shike Mei, Bin Cao, Jian-Tao Sun, and Brian D Davison. Exploiting contextual factors for click modeling in sponsored search. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 113–122. ACM, 2014.

[28] Yifan Chen, Gui-Rong Xue, and Yong Yu. Advertising keyword suggestion based on concept hierarchy. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 251–260, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-927-2. doi: 10.1145/1341531.1341564. URL `http://doi.acm.org/10.1145/1341531.1341564`.

[29] Matthew Cary, Aparna Das, Ben Edelman, Ioannis Giotis, Kurtis Heimerl, Anna R Karlin, Claire Mathieu, and Michael Schwarz. Greedy bidding strategies for keyword auctions. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 262–271. ACM, 2007.

[30] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. A semantic approach to contextual advertising. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 559–566, New York, NY, USA, 2007. ACM.

[31] Chingning Wang, Ping Zhang, Risook Choi, and Michael D'Eredita. Understanding consumers attitude toward advertising. In *In: Eighth Americas Conference on Information Systems. (2002) 11431148*, pages 1143–1148, 2002.

[32] Wen-tau Yih, Joshua Goodman, and Vitor R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 213–222. ACM, 2006.

[33] Xiaoyuan Wu and Alvaro Bolivar. Keyword extraction for contextual advertisement. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 1195–1196. ACM, 2008.

[34] Kushal S. Dave and Vasudeva Varma. Pattern based keyword extraction for contextual advertising. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1885–1888. ACM, 2010.

[35] Pengqi Liu, Javad Azimi, and Ruofei Zhang. Automatic keywords generation for contextual advertising. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 345–346. International World Wide Web Conferences Steering Committee, 2014.

[36] Ye Chen, Dmitry Pavlov, and John F. Canny. Large-scale behavioral targeting. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 209–218. ACM, 2009.

[37] Yahoo! smart ads. 2014. URL `http://advertising.yahoo.com/marketing/smartads/`.

[38] Ad link. 2014. URL `http://www.google.com/adsense/`.

[39] Double click. 2014. URL `http://www.doubleclick.com/products/dfa/index.aspx`.

[40] Blue lithium. 2014. URL `http://www.bluelithium.com/`.

[41] Almond net. 2014. URL `http://www.almondnet.com/`.

[42] Nebu ad. 2014. URL `http://www.nebuad.com`.

[43] Burst. 2014. URL `http://www.burstmedia.com/`.

[44] Tacoda. 2014. URL `http://www.tacoda.com/`.

[45] Badrish Chandramouli, Jonathan Goldstein, and Songyun Duan. Temporal analytics on big data for web advertising. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*, ICDE '12, pages 90–101. IEEE Computer Society, 2012. ISBN 978-0-7695-4747-3.

[46] Ting Li, Ning Liu, Jun Yan, Gang Wang, Fengshan Bai, and Zheng Chen. A markov chain model for integrating behavioral targeting into contextual advertising. In *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, ADKDD '09, pages 1–9. ACM, 2009.

[47] Rightmedia exchange. 2014. URL `http://richmedia.com/`.

[48] Moshe Babaioff, Jason Hartline, and Robert Kleinberg. Selling banner ads: Online algorithms with buyback. In *The Fourth Workshop on Ad Auctions (SSA'08)*, 2008.

[49] R Contantin, J Feldman, S Muthukrishnan, and M Pal. Online ad slotting with cancellations. In *Fourth workshop on Ad Auctions; symposium on discrete algorithms (SODA)*, 2009.

[50] Micah Adler, Phillip B. Gibbons, and Yossi Matias. Scheduling space-sharing for internet advertising. *Journal of Scheduling*, 1998.

[51] Victor Boskamp, Alex Knoops, Flavius Frasincar, and Adriana Gabor. Maximizing revenue with allocation of multiple advertisements on a web banner. *Computers & Operations Research*, 38 (10):1412–1424, 2011.

[52] Moshe Babaioff, Jason Hartline, and Robert Kleinberg. Selling ad campaigns: Online algorithms with cancellations. In *ACM Conference on Electronic Commerce (EC'09)*, July 2009.

[53] Katherine Gallagher and Jeffrey Parsons. Framework for targeting banner advertising on the internet. *2014 47th Hawaii International Conference on System Sciences*, 4:265, 1997.

[54] Puneet Manchanda, Jean-Pierre Dubé, Khim Yong Goh, and Pradeep K Chintagunta. The effect of banner advertising on internet purchasing. *Journal of Marketing Research*, 43(1):98–108, 2006.

[55] Vasek Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 4(3):233–235, 1979.

[56] Michael R Garey, David S. Johnson, and Larry Stockmeyer. Some simplified np-complete graph problems. *Theoretical computer science*, 1(3):237–267, 1976.

[57] Francesco Bonchi, Carlos Castillo, Debora Donato, and Aristides Gionis. Topical query decomposition. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 52–60. ACM, 2008.

[58] Bing Liu, Wynne Hsu, and Yiming Ma. Mining association rules with multiple minimum supports. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 337–341. ACM, 1999.

[59] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499. Morgan Kaufmann Publishers Inc., 1994.

[60] P.Gowtham Srinivas, P.Krishna Reddy, and A.V. Trinath. Cppg: Efficient mining of coverage patterns using projected pattern growth technique. In *Trends and Applications in Knowledge Discovery and Data Mining*, volume 7867 of *Lecture Notes in Computer Science*, pages 319–329. Springer Berlin Heidelberg, 2013.

[61] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent itemsets. *Journal of Parallel and Distributed Computing*, 61:350–371, 2000.

[62] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 1–12. ACM, 2000.

[63] P. Gowtham Srinivas, P. Krishna Reddy, A. Venkata Trinath, S. Bhargav, and R. Uday Kiran. Mining coverage patterns from transactional databases. *Journal of Intelligent Information Systems*, pages 1–17, 2014. ISSN 0925-9902. doi: 10.1007/s10844-014-0318-3. URL `http://dx.doi.org/10.1007/s10844-014-0318-3`.

[64] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, SIGMOD '97, pages 265–276. ACM, 1997.

[65] Andrew Frank, Arthur Asuncion, et al. Uci machine learning repository. 2010.

[66] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[67] http://www.delicious.com. 2014.

[68] Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011, New York, NY, USA, 2011. ACM.

[69] Ryen W. White, Mikhail Bilenko, and Silviu Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 159–166. ACM, 2007.