Statistical Techniques for Online Personalized Advertising: A Survey

Maad Shatnawi and Nader Mohamed
Faculty of Information Technology, UAEU
P.O. Box 17551, Al Ain, UAE
{shatnawi, nader.m}@uaeu.ac.ae

ABSTRACT
Online advertising is the major source of revenue for most web service providers. Displaying advertisements that match user interests will not only lead to user satisfaction, but it will also maximize the revenues of both advertisers and web publishers. Online advertisement systems use web mining and machine learning techniques to personalize advertisement selection to a particular user based on certain features such as his browsing behavior or demographic data. This paper presents an overview of online advertisement selection and summarizes the main technical challenges and open issues in this field. The paper investigates most of the relevant existing approaches carried out towards this perspective and provides a comparison and classification of these approaches.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Selection process;
H.3.5 [Online Information Services]: Commercial services, Web-based services

General Terms
Algorithms, Measurement, Performance, Experimentation

Keywords
Web advertising, online advertising, contextual advertising, sponsored Search, personalization, matching

1. INTRODUCTION
The World Wide Web continues to grow at an amazing rate and, therefore, its essential role in all parts of our live is increasing in an unbelievable manner. However, online advertising remains the only major source of revenue for most of the service providers on the web such as search engines, social networks, video sharing websites, and blogging sites. The Interactive Advertising Bureau (IAB) shows in their 2010 internet advertising revenue report that internet advertising revenues have grown from $8.09 billion in 2000 to $26.04 billion in 2010 as presented in Figure 1 and 46% of this revenue is achieved through search advertising.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC’12, March 25-29, 2012, Riva del Garda, Italy.
Copyright 2012 ACM 978-1-4503-0857-1/12/03…$10.00.

Some research efforts have been done on surveying Online Personalized Advertising. Jansen and Mullen [20] presented an overview of the processes related to sponsored search from the information searching perspective and discussed the technology behind the sponsored search focusing on the online auction process. Chowdhury [6] proposed a taxonomy of online advertising and covered the Information Retrieval (IR) research in search advertising.

Figure 1. Internet advertising annual revenue trend in billions 2000 through 2010 [18].

In this paper we explore the three categories of internet advertising: sponsored search (SS), contextual matching (CM), and shopping websites (SW) advertising. We provide an extensive comparative study of the existing approaches in these categories. The rest of this paper is organized as follows. The next section provides a brief overview of online advertising. Section 3 addresses the main technical challenges that face personalized advertising. Section 4 lists the commercial online advertising systems that are currently available in the market. In Section 5, which is the focus of our paper, extensive description and comparison are presented for the most current research approaches in the field of online advertisement selection. Open research issues are addressed in Section 6. Concluding remarks are presented in Section 7.

2. ONLINE ADVERTISING OVERVIEW
Online advertising can be performed in different ways such as Opt-In emails and newsletters, instant messaging, and displaying ads on webpages. Opt-In emails and newsletters are performed by sending advertisement messages through e-mail to a list of people who are interested to receive some information on a given topic. There are four forms of webpage advertising: pop-up and pop-under, floating, interstitial, and banner ads. Pop-up advertising opens a small window that pops up over the main browser window when a user enters a site while pop-under opens the ads...
window under the main window. Floating ads appear on a webpage when a user first goes to that page and they float over it for few seconds before settling down at some position. Interstitial ads are displayed during the transition between two pages of a website [6]. The basic concept of banner advertising is the display of a rectangular image of an advertisement on a part of the publisher (host) webpage, and clicking on this advertisement image will direct the user to the advertiser website.

The three main approaches for online advertisements are untargeted, filtered, and personalized [25]. Untargeted advertising is used by early systems and many small scale web publishers by displaying fixed advertisements on their webpages for a certain period of time according to a schedule plan. Although this approach is simple to set up and it does not share any of the privacy concerns of users, the majority of the displayed ads will be most probably irrelevant to user interests which, in turn, will not satisfy advertisers in achieving the desired click-through rate. In filtered, or targeted, advertising, advertisers can specify targeting parameters such as the user operating system, time constraints, and geographic location. The selection mechanism on the Ad server analyzes the request and selects only those advertisements that match the current situation.

Personalized advertisements are used to tailor advertisements to a particular user based on certain features such as his browsing behavior and demographic data [25]. Personalized advertising approaches can be classified into three categories according to their applications: sponsored search (SS) advertising, contextual match (CM) advertising, and shopping website (SW) advertising. SS places advertisements on result pages of a web search engine based on the user search query [3]. CM displays advertisements within a generic web page according to their relevance to the content of that webpage [6, 26]. SW displays advertisements on the Internet storefronts according to some criteria such as the previous purchases and demographic data of their customers.

Web advertising is an interaction among an advertiser, a host website, and a user. The advertiser provides the supply of ads with a particular temporal and thematic goal. The host website, such as a search engine or a web directory, provides the space for displaying the ads. The user visits the webpage and interacts with the displayed ads. The widely used pricing model for personalized advertising is the pay-per-click (PPC) where the advertiser pays for every click on his advertisement. There are other models such as pay-per-impression (PPI) where the advertiser pays for the number of appearances of the advertisement, and pay-per-action (PPA) where the advertiser pays only if the advertisement leads to a complete transaction. The advertiser usually pays for each sponsored search click according to an auction process where each advertiser places a bid or a price on a certain search phrase query and the ranking of his bid determines his position in the list of ads displayed on the search result page of that query [4, 10].

Displaying relevant advertisements to a particular user leads to satisfying the user, the advertiser, and the web publisher [28]. It satisfies users by displaying advertisements that match their particular interests. Advertisers will be satisfied as they will receive more clicks from users who are more likely interested in their products and services. Web publishers will be satisfied as well since they will obtain better targeted advertisement clicks and, in turn, maximize their advertising revenues.

3. TECHNICAL CHALLENGES

Personalized advertising faces several challenges including feature extraction, prediction technique selection, and the coverage of applications on different types of websites. Personalized advertisement recommendation systems vary in their selection of user features to build their classification models. Many of these systems are based on demographic data while other systems are based on user short-term and long-term browsing history, his current behavior, or his previous purchases. Many personalization techniques rely on user demographic data such as age, gender, location, and annual income. Although this data plays an important role in determining user interests, it has many drawbacks. First of all, user interests frequently change over time due to many reasons. Secondly, many users enter inaccurate information when they register to websites. Thirdly, the use of demographic data violates users’ privacy especially when a firm sells its customer information to other marketing firms.

The main challenge of using the browsing history is the cold start problem which is the lack of current user information at the beginning of his session as a short-term browsing data and the lack of long-term browsing data for new users. Many systems are based on matching search queries with a pre-defined bag of words or matching them with bid phrases. Depending only on the bag of words or bid phrases to match an ad with a highly related query will be insufficient when they use different vocabulary, or when there are spelling mistakes in the query. Secondly, it will be impossible for advertisers to list all relevant queries for their ads. Thirdly, users tend to form their queries with the least possible number of words, and they select the words to get the best web search results rather than ads results. This requires semantic matching rather than simply syntactic matching. Furthermore, distinguishing a shopper from a purely information-seeker is an important challenge for these systems in order to show more ads to the former and less ads to the later [3, 4]. The ads recommendation systems, which are based on matching ads with webpage content, face several challenges such as the frequent updating in the web page content.

Selecting the best machine learning technique for an advertising recommendation predictor is a great challenge. There is a variety of techniques that diverse in accuracy, robustness, complexity, computational cost, data diversity, over-fitting, and dealing with missing attributes and different features. Therefore, compromising these issues should be taken into consideration when picking-up a good technique.

The advertisement selection systems should take care that displaying very relevant ads can be harmful in some situations such as displaying an advertisement of a brand on the website of the brand itself, or displaying an advertisement of its competitor. Therefore, matching systems should be able to filter out such ads.

Finally, there are other challenges, such as boredom prevention and respecting the advertisement policies, which should be considered during the building up of an advertisement matching model. Boredom prevention determines a periodical schedule of advertisements for the user who will prevent the frequent display of the same advertisement even if this advertisement receives the best ranking in the recommendation model. The list of advertisements obtained from the matching process should be filtered according to the advertising policy constraints which were agreed between the advertiser and the publisher such as time of the day, day of the week, and location of the advertisement in the webpage [21].
4. COMMERCIAL SYSTEMS

The advertising broker acts as an intermediate body between publishers and advertisers who bid for specific keywords using an auction system. Alternatively, advertisers can directly access thousands of possible web sites to automatically advertise on. There are several commercial programs that work as advertising brokers and manage the online advertisement selection. Publishers can insert few lines of Javascript code on their web pages which will then automatically serve relevant ads using a search algorithm. The algorithm matches search keywords or webpage content to relevant advertisements. It also considers other factors such as the geographical location of users and their languages [8].

Google AdSense [14] is the first major contextual advertising program which allows website publishers to display relevant Google ads on their webpages and earn money. It's also a way for those publishers to provide the Google search service on their websites and to share the revenue of displaying Google ads on the search results pages.

As Google AdSense is a service provided to website publishers, Google AdWords [13] is a service provided to advertisers which allows them to create ads and choose keywords. When people search on Google using one of these keywords, the advertisement appears next to the search results according to a ranking mechanism which considers both relevance and bid phrases. Google AdWords allows advertisers to target online customers in a particular region which can be as narrow as twenty five miles diameter. AdWords charges advertisers based on the PPC pricing model.

Other big players such as Yahoo! and MSN are now implementing this advertising technology. The main existing online advertising providers are as follows [8]:

- Google AdSense: https://www.google.com/adsense
- Google AdWords: https://www.adwords.google.com
- AdBrite: http://www.adbrite.com
- AdGenta: http://www.adgenta.com
- AdKnowledge Miva: http://www.miva.com
- BidVertificate: http://www.bidvertificate.com
- CBprosense: http://www.cbprosense.net
- Chitika: http://chitika.com
- Clicksor: http://www.clicksor.com
- Infolinks: http://www.infolinks.com
- Kanoodle BrightAds: https://www.kanoodle.com/about/advertise.html
- Kontera: http://www.kontera.com
- Marchex Adhere: http://www.marchex.com
- Microsoft Advertising adCenter supported by Bing and Yahoo! Search engines: https://adcenter.microsoft.com
- Vibrant: http://www.vibrantmedia.com
- Yahoo! Advertiser and Agency Solutions: http://searchmarketing.yahoo.com

5. SELECTION APPROACHES

Personalized advertising systems can be classified into three main categories according to their applications; sponsored search (SS) advertising, contextual match (CM) advertising, and shopping website (SW) advertising. This section presents the state-of-the-art approaches for each of these categories.

5.1 Sponsored Search Advertising

SS advertising is placing advertisements on the result pages of a web search engine based on the user search query [3]. This type of advertising was introduced by Overture in 1998 [6, 16]. Major web search engines such as Google, Yahoo, and Microsoft support this type of advertisements and play as ads search engines in addition to their main role as web search engines [4]. SS advertising is the main revenue source for web search engines. This section presents and evaluates some of the state-of-the-art sponsored search approaches.

Approach 1: Search Advertising Using Web Relevance Feedback. Broder, et al., [2] presented an improvement to advertisement matching in search engine websites. The main idea of this work is to use web search results as new features which will be integrated with the search query keywords which, in turn, will be used in selecting the suitable advertisements that will be shown alongside the search results. When a user enters his query, it will be first sent to the web search engine, and then, the returned top-scoring pages of the search engine results will be used to gather additional knowledge about the query, and therefore, create an augmented query which is evaluated against the index of ads. To retrieve relevant ads, the entire content of the ad is used, rather than only its bid phrase. This work uses text classification with respect to a taxonomy, made by human editors, in order to identify commonalities between relevant but different vocabularies, and builds a document centroid-based classifier that maps an input fragment of text into a number of relevant query classes. Furthermore, this approach uses a tool, called Altavista's Prisma refinement tool, for phrase extraction. This tool analyzes the fragment of text to identify named entities and other stable phrases. It also analyzes the phrases found in all the crawled pages and retains the most significant ones based on their statistical properties.

The first advantage of this approach is that using web search results has many benefits in query augmentation because modern search engines correct spelling mistakes and use additional knowledge, such as past query statistics and click-through data, to return relevant results even for very short queries. The second advantage is that it presents a personalized advertisement system that does not rely on user demographic data with its several limitations such as the violation of user privacy and the nonstop changes in user interests. The main disadvantage of this approach is that it provides an enhancement to advertisement personalization on search engines only and is not applicable to online shopping portals as well as other types of websites.

Approach 2: Adaptation Between Organic Users and Shoppers. The authors of [28] described an approach to adapt the sponsored search functions to an individual user's preferences between pure web search results (organic) and advertisements. This approach trained a statistical prediction model with the user-specific features based on view and click events. These features include long-term, within the last twenty eight days, and short-term, within the last twenty four hours, historical user behavior. The prediction model uses stochastic gradient-descent boosted trees, due to its robustness and good experience in web search modeling, to predict the user's relative preferences between organic results and ads, and to adapt the number of ads shown and
their location (top, bottom, or right side) on the search results page. This model decides on which ads to show and in which order. This depends on an estimated probability of a click on an ad along with its bid. The experimental results, both off-line and on live-traffic, show that the model considerably increases the accuracy of click prediction rate on the historical search log data.

The main advantages of this approach is that it distinguishes between a purely information-seeking user and a shopper which leads to more satisfaction for both of them by showing fewer ads to the former and more ads to the latter. It also satisfies search engine providers, by gaining more and better targeted ad clicks, and advertisers, by receiving more clicks from users who are more likely interested in their products and services. In addition to that, this approach does not use the demographic information of users in which it may improve the prediction rate, but it may violate the user privacy. The main limitation of this approach is that it is only applicable to search service providers and not applicable to other types of websites such as news and on-line shopping portals.

**Approach 3: Maximum Entropy Classification Model.** Chen and Cantu-Paz [5] proposed a personalized click prediction model that improves the accuracy of click prediction in sponsored search. The model starts from an existing non-personalized system that uses user-independent features and improves this system by introducing some personalized features. The user-independent features include textual and semantic similarities between the query and the ads, the historical performance of ads, and the time of the day or day of the week. The user-specific features are the click feedback and the user demographic information such as gender and age. These personalized features are based on the observations of search and click behaviors of a large number of users of a commercial search engine. The new user-dependent features are combined with the user-independent features in a maximum entropy classification model to predict the clickability score of each ad and then display the winning ad for this specific user. The proposed model is evaluated, both off-line and on live-traffic, based on logs from the Yahoo! sponsored search traffic logs. The experimental results show that adding user-specific features improves the accuracy of click prediction rate.

This personalizing click prediction approach benefits both the users and the advertisers. The users will be presented by most relevant advertisements and the advertisers will receive clicks from users who are more engaged with their ads. Although the use of demographic information contributes in improving the ads selection criteria, there are many users who consider this usage is against their privacy rights. In addition to that, there are many users who provide incorrect demographic information which, in turn, will mislead the prediction model. Another limitation of this work is that it is not applicable to different types of commercial websites other than search engines.

**Approach 4: Web-Scale Bayesian Click-Through Rate Prediction.** Graepel et al. [15] presented a Bayesian online learning algorithm called adPredictor used for binary Click-Through-Rate (CTR) prediction in Bing Sponsored Search advertising. The algorithm is based on a generalized linear model with a probit (cumulative Gaussian) link function and a factorizing Gaussian belief distribution on the feature weights. It calculates the approximate posterior using message passing with automatic feature-wise learning rate adaptation. The described algorithm won the Bing/adCenter competition for the most accurate and scalable CTR predictor. As a result it was chosen to replace the previous Bing CTR prediction algorithm.

The input features can be grouped into three categories: Ad features, Query features, and Context features. Ad features include bid phrases, ad title, ad text, landing page URL, landing page content, a hierarchy of advertiser, account, campaign, and ad group. Ad Query features include search keywords, possible algorithmic query expansion, cleaning and stemming. Context features include geographic location, time, user data, and search history.

The use of Bayesian models in ad selection has many advantages. It is simple, effective in text categorization, easy to train, and adaptive with missing attributes.

Table 1 presents these approaches including the features that are used in matching and classification and the machine learning technique and/or the tools applied.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Extracted Features</th>
<th>Technique/Tools</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broder, et al. (2008)</td>
<td>• Search query keywords • Web search results</td>
<td>• Document centroid-based classifier • Altavista's Prisma refinement tool</td>
<td></td>
</tr>
<tr>
<td>Schroedl et al. (2010)</td>
<td>• Long-term and short-term historical user behavior</td>
<td>• Gradient-descent boosted trees</td>
<td></td>
</tr>
<tr>
<td>Chen and Cantu-Paz (2010)</td>
<td>• Textual and semantic similarities between query and ads. • Historical performance of ads • Time of the day or day of the week. • Click feedback • User demographic information</td>
<td>• Maximum entropy classification model</td>
<td></td>
</tr>
<tr>
<td>Graepel, et al. (2010)</td>
<td>• Ad features include bid phrases, ad title, ad text, landing page URL, landing page content, and a hierarchy of advertiser • Ad Query features include search keywords, possible query expansion • Context features include geographic location, time, user data and search history</td>
<td>• Online Bayesian Probit Regression algorithm (adPredictor)</td>
<td>Bing CTR prediction algorithm</td>
</tr>
</tbody>
</table>
5.2 Contextual Match Advertising

Contextual match advertising is the display of advertisements within a generic web page according to their relevance to the content of that webpage. This type of advertisement selection was introduced by Google in 2003 [6, 26]. Since users spend more time on web content pages more than search engine result pages, CM has greater potential for publishers and advertisers. However, CM is a more challenging task than SS because advertisers in SS decide the keywords that are matched against user searching keywords, while matching in CM is performed based on the page content which significantly complicates the matching task [7]. This section presents and evaluates some of the existing approaches in contextual matching.

Approach 1: Finding Advertising Keywords Using Machine Learning and Natural Language Processing Techniques. Yih et al. [31] investigated different aspects of keyword extraction. The proposed system consists of four stages: preprocessing, keyword selection, classification, and post-processing. The preprocessor analyses the HTML document and returns sentences and noun phrases using natural language processing tools including a part-of-speech (POS) tagger and a chunker. It also extracts hypertext information and Meta information in the header. The system employs “Beautiful Soup”, which is a Python library for HTML parsing. The candidate selector speeds up the processing time by considering fewer keyword candidates. The system extracts forty linguistic, statistical and historical features from the candidate phrases including the existence of nouns, proper nouns, capitalized words, the appearance in hyperlinks, title, URL, the Meta section of the HTML header, term frequency, document frequency, and the location of the term in the context. In addition to that, the frequency of the phrase in the query log of the MSN search engine was found a helpful feature in determining if that word or phrase is relevant to that page. The classification stage employs a logistic regression model for learning process. The post-processing stage calculates the probabilities of relevance of whole phrases from the probabilities of relevance of individual words of each phrase. The experimental results show a large improvement over KEA [12] and GenEx [29]. The paper refers this improvement to the successful choice of features specially the query logs of MSN Search. The main advantage of this approach is that it does not simply extract features from the main body of the HTML page, but it also includes other important features such as title, URL, and META. Another advantage of this work is the employment of both machine learning and natural language processing techniques. In addition to that, this approach can be applied to different types of websites. This approach does not need to use all users’ demographic data, it only needs the geographic location which can be known from users IP addresses.

Approach 2: Impedance Coupling. Ribeiro-Neto et al. [27] introduced an Impedance Coupling technique which expands the text of the webpage with new terms to reduce vocabulary variance with relevant advertisements. This approach facilitates the task of matching ads with webpages and increases the matching accuracy by 50%. A Bayesian network model is implemented to train the algorithm using different features of the triggering page. The model matches the triggering page with other similar pages on the web and then uses the text of these pages to expand the vocabulary of the triggering page.

Approach 3: Genetic Programming. [23] proposed a framework based on Genetic Programming (GP) to associate ads with the contents of web pages. GP is a machine learning technique, inspired by biological evolution, used to approximate non-linear functions and to solve a wide range of complex optimization problems. By using a collection of real advertisements and web pages from a Brazilian newspaper, the experimental results show that GP is able to learn ranking functions that are very efficient in placing relevant ads on web pages. Moreover, GP was able to learn functions that successfully avoid the placement of irrelevant ads by calculating thresholds based on the page content where the ads should be placed.

The main advantage of this work is the employment of GP into the advertisement selection system for the first time. In addition to that, this approach can be applied to different types of websites. It does not use demographic data of users. This approach can be extended to not only extract keywords from the main text of the webpage, but also to include other important features such as title, URL, and the META section. In addition, it can deal with webpages that have no text content.

Approach 4: Semantic Matching. Broder, et al [3] introduced a way of combining semantic matching with syntactic matching of advertisements to web pages. The semantic stage classifies pages and ads to nodes in a taxonomy of common commercial topics, and then it matches pages and ads by measuring the distance between the nodes they were classified to. Several classifiers and taxonomies were evaluated in this work and it was found that Rocchio’s classifier gave the best results using a taxonomy of about six thousands nodes. In case there is no ad closely relevant to a webpage, the hierarchical taxonomy provides a way to gradually generalize by going to the parent node. The classification relies on the full content of the page rather than its individual phrases. The final matching score is a combination of both the semantic and syntactic sub-scores.

The main advantage of this work is the use of semantic matching and taking the full content of webpages into account. In addition to that, this approach is applicable to all different types of transactional and non-transactional websites. It also satisfies webmasters, by gaining more and better targeted ad clicks, and advertisers, by receiving more clicks from users who are more likely interested in their products. In addition to that, this approach does not use the demographic information of users which, although it may improve the prediction rate, it may violate the user privacy.

Approach 5: AdROSA. Advertising Remote Open Site Agents (AdROSA) system [21] is based on web content-mining and web usage-mining techniques to extract knowledge from both publishers and advertisers web pages, historical user sessions including information about advertisements clicked during these sessions, and current behaviors of online users including presented and clicked advertisements during their active sessions. The system combines several factors in one framework. These factors are the content of advertisers’ web sites, click-through probability, advertising policy, and the boredom prevention mechanism that determines a periodical scheduling of advertisements for the user. The system matches advertiser portals with the recent user interest which is expressed by his navigational behavior. Based on the nearest neighbor technique, AdROSA dynamically assigns items that have been evaluated positively by similar users. The list of advertisements obtained
from the matching process is filtered using advertising policy features such as time of the day, location, and page layout constraints. This work is implemented through a multi-agent architecture where each agent is responsible for a specific task and interchanges knowledge with other agents. The system was demonstrated and evaluated on a Polish portal, poland.com. AdROSA presents a personalized advertisement system depending on individual user browsing behavior rather than his personal data input or demographic information. Therefore, the system is far away from the limitations caused by relying on user demographic information such as the inaccuracy of user personal data input, the violation of user privacy, the continuous changes in user interests, and the restricted applications to websites that requires user membership. AdROSA is applicable to open public web sites including news service providers, search engines, and e-commerce gateways. Furthermore, this system operates in a dynamic way that reflects current user behavior and monitors updates in webpage contents.

There are two main limitations of this approach: the lack of current user information at the beginning of his session and the lack of visiting patterns for newly added advertisements to the ads database.

**Approach 6: Ant Colony Optimization.** White et al. [30] introduced the use of an ant-based algorithm to improve online advertisement matching with webpage content. The model uses historical data from user click-through patterns and the Ant Colony Optimization (ACO) [9] which is a computational technique inspired by the way of how ant colony finds food.

Although the model was evaluated through simulation and the results show significant improvements in advertisement selection, it should be implemented and tested on a real web server to evaluate its effectiveness.

**Approach 7: Clickable Terms.** Hatch et al. [17] introduced a “clickable terms” approach to contextual advertising which matches a website directly with a set of ad-side terms regardless of the page content. This approach can be summarized as follows. The log-likelihood ratios (LLRs) are derived from a Bayesian framework for measuring click probabilities of ads on a given site. These LLRs are then used to measure the relative clickability of a given ad-side term on a given site which, in turn, is used to form a new set of features for training a maximum entropy (ME) click model. The work investigates various techniques for feature normalization which shows a significant improvement on the test results. The experiments performed on Yahoo!’s contextual advertising system shows significant improvements in click-through-rate compared to a model that uses only lexical match features.

This approach has several advantages. It overcomes the limitations of using the lexical match as a measure of relevance of an ad to a webpage. Lexical matching fails to match ads to pages that have limited or no commercially relevant words such as a webpage that contains only pictures without text. Certain websites attract users who share particular interests which can be inferred from the identity of these websites regardless of their textual content. Secondly, this approach does not simply consider prior information from the page and ad only.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Extracted Features</th>
<th>Technique/Tools</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yih et al. (2006)</td>
<td>• Sentences and noun phrases</td>
<td>• Natural language processing tools</td>
<td>MSN search engine</td>
</tr>
<tr>
<td></td>
<td>• Hyperlinks, title, URL, Meta section</td>
<td>• Beautiful Soup Parser/ Python</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Term frequency, document frequency, and location of the term in the context</td>
<td>• Logistic regression</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Frequency of the phrase in the query log</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ribeiro-Neto, et al. (2005)</td>
<td>• Ad title, text, hyperlink, and keywords</td>
<td>• Bayesian network</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Webpage content</td>
<td>• K-nearest neighbors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Ad keyword expansion</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Webpage terms expansion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lacerda, et al. (2006)</td>
<td>• Ad keywords, title, description, and landing page content</td>
<td>• Genetic Programming</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Statistical features (number of appearances, etc…)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broder, et al. (2007)</td>
<td>• Full content of the page</td>
<td>• Taxonomy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Syntactic matching</td>
<td>• Rocchio’s classifier</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Semantic matching</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kazienko and Adamski (2007)</td>
<td>• Historical user sessions</td>
<td>• Nearest neighbor</td>
<td>Cold start problem</td>
</tr>
<tr>
<td></td>
<td>• Current user behaviors</td>
<td>• Multi-agent architecture</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Time of the day</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Location</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Page layout</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White, et al. (2010)</td>
<td>• Uses historical data from user click-through patterns</td>
<td>• Ant Colony Optimization</td>
<td></td>
</tr>
<tr>
<td>Hatch et al. (2010)</td>
<td>• lexical match: words and phrases</td>
<td>• Log-likelihood ratios</td>
<td>Yahoo!’s contextual</td>
</tr>
<tr>
<td></td>
<td>• clickable ad-side terms</td>
<td>• Bayesian framework</td>
<td>advertising system</td>
</tr>
<tr>
<td></td>
<td>• Click feedback</td>
<td>• Maximum entropy</td>
<td></td>
</tr>
</tbody>
</table>
It also considers the click feedback to estimate the click probability of the ad. Thirdly, the use of Bayesian models in ad selection has many advantages. It is simple, effective in text categorization, easy to train, and adaptive with missing attributes. Finally, this approach can be applied to a variety of website types including commercial and non-commercial sites.

Table 2 presents these approaches including the features that are used in matching and classification and the machine learning technique and/or the tools applied.

### 5.3 Shopping Website Advertising

Transactional or shopping website advertising is displaying advertisements on online storefronts according to some criteria such as the previous purchases or the demographic information of their customers. This section presents and evaluates a couple of approaches in this type of advertising.

**Approach 1: Decision-Tree Induction Techniques.** Kim et al. [22] proposed a rule-extraction induction method for personalized recommendation on internet storefronts using decision tree induction techniques that match customer demographic information to product categories. The extraction of useful marketing rules begins by defining a hierarchy tree of product categories to extract marketing rules at various abstraction levels of product categories. The rules can be extracted for specific product items and then stored in a marketing rule-base and to be used for real time personalized ads selection when customers visit the internet store. Advertisements are selected based on both customer profiles and extracted marketing rules. The main limitation of this approach is that it relies on demographic data which has several disadvantages such as the inaccuracy of user personal data input, the violation of user privacy, and the continuous changes in user interests. Another limitation of this approach is that it is not applicable to websites other than e-commerce sites such as news service providers and search engines.

**Approach 2: Application of Half-Life Theory and Fuzzy Theory.** Lai and Hwang [24] described a consumer preference analysis system based on the half-life theory and the fuzzy theory which uses the consumer browsing habits and purchase amounts to predict his preference levels in relation to products and then proposes a selection model for product advertisements. The model matches member personal preferences with advertisement products. The system uses recommendation rules to create a delivery scheduling system for determining the order of advertisement delivery and enhancing the efficiency and commercial fairness of advertisement delivery. In addition to that, the system produces a list of members having similar shopping preferences. This list is provided to these members, when they log in the next time, as a suggested list so each member determines whether or not to exchange among the listed members. The practical part of this work is the construction of a membership-based shopping website that displays web advertisement based on the proposed methodology. This website is equipped with a chat room and a friend-making function to promote exchange between online groups.

The main advantages of this approach can be summarized as follows; first, it increases the marketing effect of advertising websites based on the consumer browsing habits and previous purchases. Secondly, it takes into consideration the fact that consumer preferences change over time. Thirdly, this approach enhances the effects and fairness of advertisement delivery so that the user will not be bored due to high repetition of the same ads and, in the same time, it will make sure that there are no ads that are delivered at an excessively low frequency or not delivered at all. Finally, the methodology is examined in practical operation by the constructed shopping website. The main limitation of this approach is that it is not applicable to websites other than e-commerce sites such as news service providers and search engines.

Table 3 presents these approaches and includes the features that are used in matching and classification and the machine learning technique and/or the tools applied.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Extracted Features</th>
<th>Technique/Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim, et al. (2001)</td>
<td>Customer demographic information</td>
<td>Decision tree induction techniques</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extraction of marketing rules</td>
</tr>
<tr>
<td></td>
<td>Purchase amounts</td>
<td>Half-life theory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fuzzy theory</td>
</tr>
</tbody>
</table>

### 6. OPEN ISSUES

Although the research in advertisement personalization is very promising, there are several open issues that require adequate attention from researchers. One of the major issues in PPC, the most commonly used payment model, is the click fraud problem where many dishonest publishers simulate fraudulent clicks to increase their revenue [11, 19]. This problem can lead the advertisers to lose their confidence in the advertising network, which could harm business for all the associated parties. One solution to address this issue is the intervention of human judges. The author in [11] suggested the use of unlabeled data to train such detection mechanisms as it is impossible to label millions of clicks for different keyword-ad combinations. There are other solutions to this phenomenon such as the use of the PPA model instead of PPC, blocking blacklisted IP’s, and the aggressive monitoring. In general, the research in this field is still in its infancy and needs a lot of work to reach an acceptable and reliable situation [6].

Privacy and security are important issues in the field of information retrieval in general and advertisement personalization in particular. Advertising selection approaches infer the user interests from his personal data, either demographic or browsing behavior. However, the issue of privacy has not been properly addressed yet [1].

As can be seen, most of the research in advertising selection is applied to English in particular and Latin languages in general. There is a lot of work that should be performed for applying advertising selection to other languages such as Arabic and Chinese. Furthermore, the cross-language retrieval should be addressed, where the search query or the webpage content is in one language while the advertisement keywords are in another language.

One of the challenges that face the use of browsing history is the cold start issue which is the lack of current user short-term browsing data at the beginning of his session, and his long-term browsing data in case of a new user. One of the approaches to
address this issue is the collaborative filtering by clustering users who have similar features into a number of groups, and then assign the features of the closest group to a new user. However, collaborative filtering requires some previous user information to determine his similarity with other users.

7. CONCLUSION

This survey is focused on internet advertisement personalization. The main challenges that face online advertisement personalization were presented. We investigated most of the relevant existing approaches carried out towards this perspective and provided a comparison and classification of these approaches. Online advertisement personalization approaches are mainly classified into three categories: sponsored search, contextual matching, and shopping website advertising. Open research issues were discussed. Some of these issues are privacy and security, other languages support, and the cold start issue.

8. REFERENCES