A Survey of Search Advertising

N.M. Mosharaf Kabir Chowdhury
nmmkchow@cs.uwaterloo.ca

December 16, 2007

Abstract

In last decade, yearly revenue from online advertising has soared from a moderate US$907 million to a whooping US$16.9 billion and the driving force behind this huge leap is found to be search advertising. Forecasts suggest that this trend will continue in coming years through the innovation of new types of search-related advertising. This rapidly burgeoning industry involves complex business models and sophisticated technologies. In this paper, we summarize the key concepts of search advertising and identify information retrieval interests in this topic. The aim of our work is to provide a basis to build upon and further consolidate search advertising, so that the current positive trend of growth can be sustained.

1 Introduction

The emergence of the Internet dawned a new age of marketing that opened the possibility of global exposure to a large audience at a very low cost. The prospects were so alluring that during the mid and late 90's, many enterprises were willing to spend large sums of money on online advertising without any apparent concern for their investment return [30]. This situation took a nosedive starting from the first quarter of 2001, when failure of several Web companies caused a dropping in supply of cheap venture capital and considerable reduction in online advertising [29, 30]. This negative trend, however, has been reversed by the end of 2002, due to the increasing adoption of a particular Web advertising format, the search advertising. Interactive Advertising Bureau's (IAB[1]) annual Internet advertising reports show that since 2002, yearly advertising revenues have steadily grown from US$6 billion to US$16.9 billion by the end of 2006 and 40% of that is contributed by search advertising alone [15].

Search advertising has essentially become the driving force behind the monetization of Web services through online marketing. To sustain this unbelievable growth of search advertising and to increase it further, we need to understand the core it is built around. A lot of studies exist focusing on the commercial front of search advertising. But the number of published research works on the theoretical core and various aspects of search advertising is surprisingly low. The aim of our work is to present a comprehensive study of search advertising from information retrieval (IR) point of view and to provide a basis to identify research opportunities that can further improve the quality of search advertising, and sustain the current positive trend of growth.

The remainder of this paper is organized as follows. In Section 2, we propose a taxonomy of online advertising from different perspectives and identify the positioning of search advertising in the taxonomy. In Section 3, we give an overview of the basic concepts of search advertising and discuss its main categories under the light of a generic model. In Section 4, we cover the IR research interests in search advertising and provide a comprehensive overview of the available research works in this area. We conclude in Section 5.

2 Taxonomy of Online Advertising

In this section we try to categorize different types of online advertising techniques that exist on the Internet today. To the best of our knowledge no such classification of different types of ads has appeared before in any other research work.

[1] http://www.iab.net

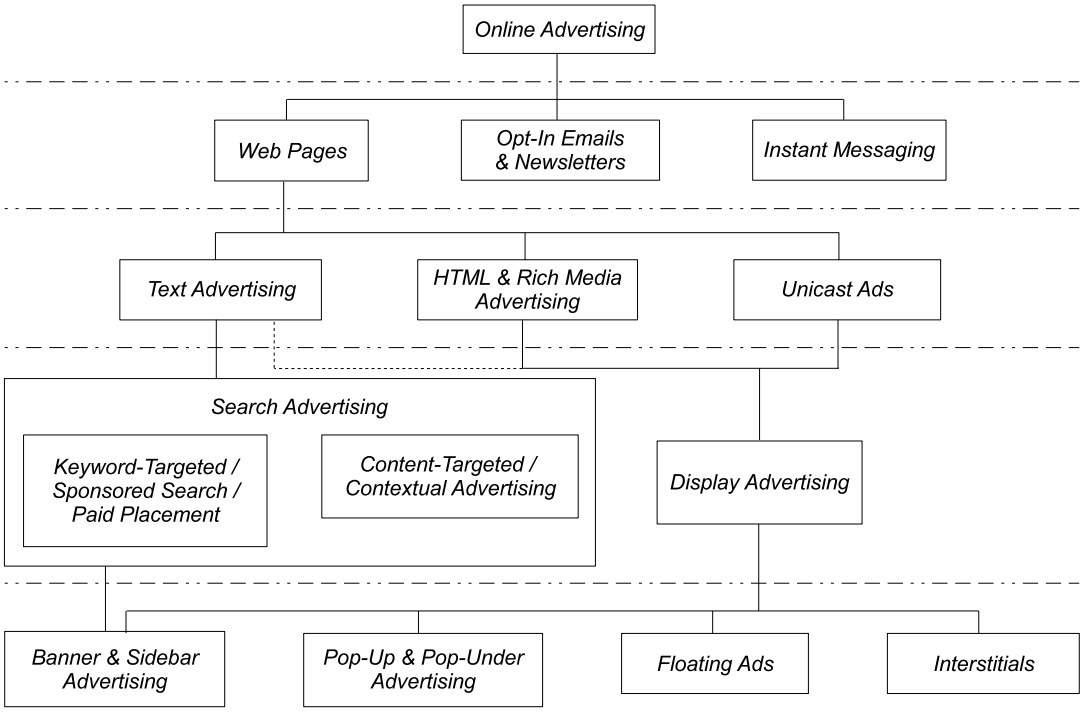

Figure 1: Taxonomy of online advertising

2.1 Classification Based on Medium

The Internet is not limited to Web pages only. There are emails, RSS feeds and advertising has made its way into every single dimension the Web has provided us.

2.1.1 Web Pages

Web pages are the most likely place to find ads. These pages range from result pages containing search results pertaining to a particular user-query, to normal Web pages containing publisher provided contents. In recent years, Google has started showing ads even on their Web-mail service. Ads can be placed in different parts of the page and can be based on query words in case of results page, or page content, or just ads without any particular relation to the page (which is decreasing rapidly as the users are more likely to follow an ad that is relevant to something they are looking for, rather than a random one).

2.1.2 Opt-In Emails and Newsletters

Opt-In mailing or newsletters is an advertising technique that is rapidly becoming popular. It consists of sending email messages to a "pre-qualified" list of people i.e. an audience that has expressed an interest in receiving information on a given topic. It is different from spamming in the sense that the email receivers willingly agree to receive the ads and can opt-out any time they want.

2.1.3 Instant Messaging

Advertising in instant messaging softwares based on the content of the messages passed back and forth between contacts got a lot of buzz a few years ago. But up until now, there is no such service in existence. We believe that instant messaging holds much prospect as a possible advertising medium and it is even more interesting from information retrieval point of view.

2.2 Classification Based on Ad Media

Online ads can be further classified by considering the media it uses. Sometimes ads are presented as simple texts; but sometimes it combines flashy images to heavy-weight rich media contents.

2.2.1 Text Advertising

Text ads are probably the most dominant form of online advertising today. They are less annoying in terms of user experience and in most cases received positively. Each of the large search engine companies, e.g. Google, Yahoo!, MSN Search, have their own brands of text ad systems. Text ads first appeared alongside search results in different search engines and eventually made their way into content-based websites. On its way toward evolution, text advertising gained much attention from information retrieval community.

Text ads normally consist of a title, a short description, and a URL address. The content of the title and the description is known as *creative*. The goal of the creative is to provide an attractive summary of the URL to lure users to click on it. Moreover, all three parts act as the basis for selection of an ad to be displayed on a particular Web page.

2.2.2 HTML and Rich Media Ads

HTML ads combine graphics and text with other HTML elements such as pull-down list, check boxes or forms. Rich media ads take it further with the use of multimedia elements such as sound, animation (often using Shockwave or Flash) and Java/Javascript to drive the message home.

2.2.3 Unicast Ads

Unicast ads are a special type of rich media ads. A Unicast ad is basically a TV commercial that runs in a pop-up window. It is animated and it has sound. The ads can last anywhere from 10 to 30 seconds.

2.3 Classification Based on Ad Selection Source

One of the most important decisions in online advertising is how to select ads to be shown on a particular page. Based on the source, we can categorize ads into two major categories.

2.3.1 Search Advertising

Search advertising is the most dominant form of online advertising today. Normally, it comes in the form of text ads in terms of presentation. Interactive Advertising Bureau (IAB) categorizes it into two major divisions:

Keyword-targeted Advertising / Sponsored Search / Paid Placement

Keyword-targeted advertising was introduced by Overture[2] in 1998 [13]. Later Yahoo! acquired Overture and Google, Ask.com and MSN Search implemented their own similar services [8]. It is normally seen at the top or at the right hand side of Web page search results. The ads are selected from the query words provided by the user, hence the name. In general this type of ads are text ads.

Content-targeted Advertising / Contextual Advertising

Keyword-targeted advertising was introduced by Google in 2003 [24]. In this case, ads are selected based on the content of a Web page instead of a particular set of user provided keywords. As a result, it is a much more challenging and interesting problem.

[2]http://www.overture.com

2.3.2 Display Advertising

By display advertising, we refer to all those ads that are neither keyword-targeted nor content-targeted; instead, they are selected based on some predefined agreements between the owners of the Web pages and the advertisers. These are mostly non-textual ads and there is no specific mechanism that might have been used to automatically select them to be displayed on a particular Web page. All sorts of online ads before the emergence of keyword-targeted advertising in 1998 fall into this category and even today they exist, although the popularity has been decreasing since the inception of search advertising.

2.4 Classification Based on Ad Placement

Ads can be categorized into four major categories based on where they are placed on a Web page.

2.4.1 Banner and Sidebar/Skyscraper Ads

On October, 1994 HotWired gave birth to online advertising by introducing graphical banner advertisements to the Web [19]. Since then banner ads have been the most familiar type of Web ads. Even though the effectiveness of banner ads has been questioned, they are still present in most of the Web pages we visit everyday. A banner ad can be described as a graphical bar or button, generally located at the top or bottom of a Web page, containing text or graphics designed to attract a viewer's attention and induce an action.

A sidebar ad (also known as a skyscraper ad) is similar to a banner ad, but it is vertically oriented on either side of a page rather than horizontally. Because it is vertical, users cannot scroll it off the screen like banner ads. Hence, sidebar ads are believed to be more effective than banner ads.

2.4.2 Pop-Up and Pop-Under Ads

Pop-up ads consist of a small window that "pops up" over the main browser window when a user enters a site (and sometimes when she leaves it, a favorite tactic of adult sites). The pop-up windows can contain anything: simple text, rich graphics, a form to collect information or email addresses, sometimes even little games.

Pop-under ads are similar to pop-up ads except that they place themselves under the content the user is trying to read and hence less intrusive than pop-up ads.

2.4.3 Floating Ads

Floating ads appear on a Web page when a user first goes to the page and they float over the page for 5 to 30 seconds before settling down at some position. While they are on the screen, they obscure the view of the page the user is trying to read, and they often block mouse input as well.

2.4.4 Interstitials

Interstitial ads are shown in the transition between two pages of a site. Whenever a user clicks on a link on page 'A', instead of going to page 'B' she is taken to an intermediate page 'C' containing ads with a link to page 'B', commonly located at the top or at the bottom of that page. Many site visitors find interstitials irritating, and they also increase site loading times.

3 Search Advertising: Basics

In this section, we discuss the basic concepts of search advertising. We present a generic model of search advertising network, identify the main actors and components in the search advertising ecosystem and briefly discuss the relationships between different actors. Then we discuss both keyword-targeted advertising and content-targeted advertising under the light of this model. Finally, we present a conceptual search advertising system that materializes the generic model.

3.1 Search Advertising Model

Search advertising systems fall into a special category of retrieval systems, called Best Bets Systems [1]. In such systems, the contents chosen by a provider for a specific user is in a sense his bet on the fact that the user will find it interesting and providers also bet against each other for user attention. Similarly, the advertisers in search advertising systems provide ads that users might find interesting. the authors in [1] suggest that such retrieval systems can be described in terms of a retrieval model and the retrieval functions they provide.

A retrieval model R consists of an abstract description of the indexing process, the representations used for documents and queries, the matching process between them, and the ranking criteria used to sort the results [1]. R can be formally represented as a tuple $\langle D, Q, match, rank\rangle$, where D is a document collection, Q is a boolean query collection, $match : Q \times D \rightarrow \{0, 1\}$ is a query matching function, and $rank : Q \times D \rightarrow [0, 1]$ is a ranking function.

In a traditional IR system the task is to get the best k ranking documents satisfying query q. Thus, the document retrieval functions can be defined by (1) and (2) [1]:

$$search(q, D) = \{d \in D | match(q, d)\} \tag{1}$$

$$searchTop(k, q, D) = \text{top } k \text{ docs in } sort(search(q, D), rank(q, .)) \tag{2}$$

where $sort(S, rank(q, .))$ sorts the documents in S according to $rank(q, .)$, the ranking function partially applied to query q.

Based on this model, the authors in [6] defined retrieval functions of a search advertising systems. Let $P = Q \times A$ be a set composed of pairs $\langle q, a\rangle$, where $q \in Q$ is a query that defines the criterion used to select consumers interested in products advertised by means of ads $a \in A$. In a search advertising system the task is to get the best k ads in which the selection criteria match the user query or the content of the page browsed by the user. Since a search advertising system can be seen as a reversed version of an IR system, its retrieval function can be defined by (3) and (4) [6]:

$$adSearch(d, P) = \{a \in A | \langle q, a\rangle \in search(d, Q) \wedge \langle q, a\rangle \in P\} \tag{3}$$

$$adSearchTop(k, d, A) = \text{top } k \text{ ads in } sort(adSearch(d, P), rank(., d)) \tag{4}$$

In this model the advertisers define the selection criteria q to indicate the users they want to deliver their ads by using complex queries.

3.2 Search Advertising Network

A search advertising network is composed of four major actors: the publishers, the advertisers, the ad network or the broker and the users [6, 4].

3.2.1 The Publishers

Publishers are the owners of the Web pages on which the advertisements are displayed. They are basically interested in monetizing their content and increasing user loyalty. But it has been found that users do not want to pay for content and are easily annoyed by traditional advertising formats [20]. So a publisher resort to a broker to ensure that it can provide free content while earning ad revenue without hampering good user experience. Since the ads in search advertising are targeted, and hence related to user interest, more users are likely to follow them resulting in more revenue for the publisher. In the mean time, users are also kept satisfied as they can access the content free of cost and they bear a positive attitude toward the publisher and its content [20].

The publishers provide the broker with a description of the content of their pages by using keywords and/or categories. But the current and more effective way is that an automatic system provided by the broker infers the category of the content in an automatic or semi-automatic fashion.

3.2.2 The Advertisers

Unlike publishers, advertisers do not just try to monetize by directly selling their products and services; they also want to promote and build their brand name among quality users, which will eventually help in commercializing their commodities in the long run. From the advertisers' point of view, quality users are those which are interested or could become interested in their products or services.

The advertisers compete among themselves for favorable keywords by bidding in an auction system [9, 10, 13]. They pay to the broker according to the traffic provided to them by the publishers. Usually, the activity of the advertisers are organized around *campaigns*, which are defined by a set of ads with a particular temporal and thematic goal.

3.2.3 The Ad Network/Broker

The ad network or the broker, as the name suggests, is a mediator between advertisers and publishers. The broker is responsible for maintaining the whole search advertising and managing regulatory policies. Furthermore, brokers are responsible for holding auctions by offering tools that the advertisers use to bid on the keywords they think necessary to describe their products and services. And most importantly, the broker implements the technology to successfully match keywords/contents and ads. It also provides feedback and performance data to the publishers and advertisers.

3.2.4 The Users

The last actor in advertising network is the user or the consumer. They are interested in contents provided by publishers to satisfy their information needs and occasionally, they click on ads that seem relevant to their interest. As a result, revenue is generated that keeps the system going. Another important impact of users is that the publishers get genuine stimulus for the production of high quality content. This happens because the publishers income depends on the user satisfaction; if there is no user there will be no online monetization. Thus the publishers strive to provide high quality content to maintain user loyalty.

3.3 Keyword-targeted Advertising / Sponsored Search / Paid Placement

In keyword-targeted advertising, the ads (normally, text ads) that are displayed at the top or at the right hand side of the Web page are related to the content or intent of the user query. Sophisticated algorithms exist to find out the relationships between user queries and ads. And the critical requirement is speed. Since the ads must be based on the user query, the matching must also be fast as well as accurate. This selection procedure has been extensively discussed in IR theory and we will go through the existing methods later in this paper.

To associate a certain keyword K with one of its campaigns, the advertiser bid on the keyword K in an auction type system against its competitors. The more the advertiser bids, the greater are the chances that its ads will be shown in the *paid list*, i.e. the list of ads, associated with that keyword. It should be noted that the advertisers will only pay when the users click on their ads. This system is known as *Pay-Per-Click* as opposed to *Pay-Per-Impression* system where payments are based on how many times an ad is displayed regardless of actual user interest.

By creating a compelling ad, the advertiser expects to attract users to click on the URL and jump to its *landing page*. In the landing page the user will find detailed information related to the product or service and may engage in commercial activities. This event is known as the *conversion* of a click. Conversion rate is the proportion of clicks that result in transactions [6].

If we take a moment to compare keyword-targeted advertising to the generic model of search advertising, we will find that the broker and the publishers have been omitted from our discussion above. Because in this model, in most cases, the broker and the publisher are the same entity (e.g. Google, Yahoo!). But there are smaller search engines like Ask.com that depend on separate broker to get ads and in that case the generic model is exactly followed.

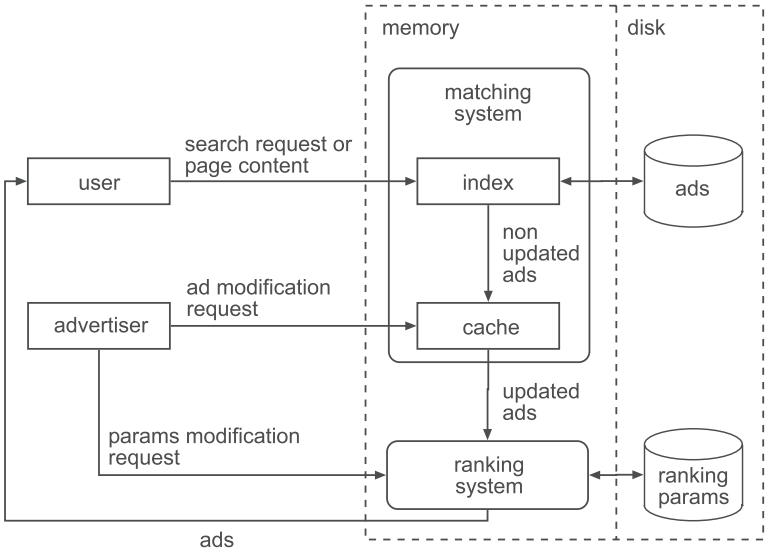

Figure 2: Search advertising system [6]

3.4 Content-targeted Advertising / Contextual Advertising

Content-targeted advertising is similar to its keyword-targeted counterpart except for the fact that there is no user query to select ads; instead ads are selected based on the content of a Web page. The page is known as the *triggering page*. The mapping process that matches pages to ads is far more complex than the mapping process for keyword-targeted advertising. This can be totally automated or semi-automated and the publisher can also directly inform the ad network about the possible categories that its pages might fall in [6].

Similar to keyword-targeted advertising, the advisers still bid for favorable keywords, which are used by the broker later in the process of selecting or mapping ads. And the ads are grouped in paid lists and placed on the triggering page, which is also similar.

So the only part that is different is the mapping or selection process as there is no explicit keyword mentioned by a user. To find appropriate ads many algorithms and techniques exist, which include key phrase extraction methods, genetic algorithms etc. We will defer the discussions of the methods for later.

3.5 Search Advertising System

A search advertising model is characterized by the dynamism in its ad collections and ranking functions that result in dynamic changes and frequent updates [6]. Authors in [1] proposed a search advertising system that copes with this characteristic by optimizing query search efficiency, performing incremental query update, and implementing dynamic ranking. Efficient query search is essential specially for keyword-targeted advertising where the ads have to be shown on the fly, along with the results of the user query. The ability of incremental updates makes online modification of the ad collection possible. Finally, a dynamic ranking is necessary to model the continuous competition between advertisers updating their ranking parameters. Figure 2 illustrates the system presented in [1].

In this system, ads and their associated selection criteria as well as the parameters related to each advertiser are stored in disk. For each user query, at first, a subsystem performs similarity matching of that query and the stored ads, and selects a set of ads based on the matching function implemented in that system. Then this set is passed to the ranking subsystem, which ranks the ads based on the bidding of the advertisers, who vie for the keywords present in the user query that were used to select the ads in the first place. The advertisers can update there ads as well as bids for keywords and affect the ranking mechanism in real time.

It is clear that the matching and ranking subsystems are the two most important components in the

search advertising system. Hence they are the main focus of IR research in search advertising. In the next section we will discuss present findings related to the matching and ranking process in the context of both keyword-targeted and content-targeted advertising.

4 Search Advertising: IR Issues

In this section, we discuss the IR-specific issues of search advertising that has made this topic so popular among IR researchers in recent years. At first, we explain behind-the-scenes details of matching and ranking subsystems. Then we go through term suggestion mechanisms that help the advertisers to find suitable keywords to represent their products and services. Finally, we talk about adversarial IR issues in search advertising.

4.1 Matching Strategies

Over the years many works in search advertising have found that the most important factor that can make search advertising successful and appealing to the users is the relevance of the ads that are selected to be displayed. The studies in [12] point out that the user perceives any Web content more positively when there is a strong relationship between the content and the advertised products alongside the content. The authors in [10, 28] showed in their studies that ensuring relevance of the ads reinforces positive user attitude to the advertisers and the ad network and also increases click-through rate.

As a result, many sophisticated matching strategies have been proposed and developed to ensure that only relevant ads are passed to the ranking subsystem and no irrelevant ad can make its way to the user. These matching strategies can broadly be categorized into two major categories based on the ad selection source. Some of them are for keyword-targeted advertising and some are for content-targeted advertising. We discuss them in details in the following.

4.1.1 Matching Strategies for Keyword-targeted Advertising

Matching in keyword-targeted advertising is pretty straight-forward in the sense that users directly provide the selection criteria in the form of queries. The system just has to find ads that match with the query. There are two main strategies to perform such matching. One is *exact* matching, where an ad must exactly match with the user query. The other is *approximate* matching where the keywords are matched to the user query partially or completely, regardless of the order. Some matching functions also employ query expansion techniques to include synonyms, related terms and plural forms.

Studies have shown that the nature and size of the keywords also have impact on the likelihood of an ad to be clicked [6]. When the keywords are somewhat generic and very few in number it is more likely that the user is interested in getting an overall idea of a particular group of products or services. As the user gets hold of the matter and start looking for products from a specific advertiser or for a specific product, it can be inferred with very high probability that the user is in buying mode. Matching strategies can take advantage of such subtle hints to make the ads more precise.

By using browser information and IP identification, search advertising networks can determine the geographic location of the user. Using this information, it is possible to display ads that are actually attainable by the user. For example, if someone in Tokyo is looking for restaurants, there is no point in showing her the ads of the restaurants in New York. Same is true for many other services that are highly dependent on geographical positioning. In fact, every single bit of information about the user can help finding the perfect ads for her queries.

4.1.2 Matching Strategies for Content-targeted Advertising

In comparison to keyword-targeted advertising, content-targeted advertising is a fairly new concept, and since there is no user query to rely on, the procedure is far more complex. In this case, the ad network extracts key topics from the full text of Web pages and tries to formally categorize those pages using different techniques, which include employing human judges, reducing this problem to keyword-targeted advertising by extracting keywords from Web pages and the use of genetic algorithms as well as machine learning techniques.

One of the earliest systems for keyword extraction is GenEx system [27]. Even though it was not created with content-targeted advertising in mind, the concept of extracting keywords is still relevant. GenEx is a rule-based key phrase extraction system that is essentially a genetic algorithm trained using 12 document features. The author showed GenEx's superiority to earlier machine learning based key phrase extraction algorithm C4.5 [21], by training C4.5 with the same 12 features. He also showed that GenEx generalizes well across collections: after being trained on a particular topic, GenEx can extract key phrases from Web pages on a different topic quite successfully. Since, the training period of any genetic algorithm is very time consuming, this capability of GenEx makes it a practical solution.

At the same time period of GenEx, KEA key phrase extraction algorithm was developed using naive Bayes machine learning scheme [11]. Even though Bayesian network is a very simple procedure to learn a function, the authors showed that it performs on a par with GenEx. The authors demonstrated that using only three features ($TF \times IDF$, *distance* of the first occurrence of the key phrase from the beginning of the document and key phrase *frequency*) KEA's performance can be significantly boosted if it is trained on documents belonging to the same domain as those from which key phrases are to be extracted. And, of course, it is much faster than any genetic algorithm based learning method. Later, KEA's performance was further improved by adding a new set of features for measuring coherence [26]. The authors in [17] added *semantic ratio* feature to the standard KEA algorithm to increase its effectiveness by 50% . Semantic ratio is simply the frequency of a candidate key phrase in a particular document divided by its frequency in all the documents that the base document has hyperlinks to.

The use of linguistic knowledge for keyword extraction was first studied in [14]. It showed that instead of relying on just statistical features, a better result can be obtained by using part-of-speech information of candidate key phrases. Unlike KEA, it used only abstracts of documents instead of the whole document to extract key phrases.

The authors in [25] implemented a four stage keyword extraction system that outperforms KEA and GenEx. They used 12 features to train a logistic regression model. The features include linguistic, statistical and historical information. They showed that query logs of search engines can boost the performance of keyword extraction more than any other feature.

In contrast to the keyword extraction based mechanisms, there are several algorithms that directly match Web pages to ads. One such algorithm to directly match ads to Web pages based on genetic programming can be found in [18]. The authors in [22] describe an Impedance Coupling technique which expands the text of the Web page to reduce *vocabulary impedance* with regard to an advertisement, increasing average precision by 50%. At the core, they used Bayesian network to train their algorithm using differnt features of the triggering page. In terms of performance this system is a tad slower than keyword extraction mechanisms, but the authors ignore the issue arguing that in content-targeted advertising ads can be associated to Web pages off-line and there is no requirement for real time computation.

Over the years, the impact of semantic information in content-targeted advertising has been overlooked by the researchers. The authors in [4] uses semantic information by creating large hierarchical taxonomy of common commercial topics and then classifying pages and ads to nodes in that taxonomy using Rocchio's framework [23]. Finally, they match pages and ads by measuring the distance between the nodes they were classified to. The hierarchical taxonomy provides an easy way to generalize in case there is no ad closely relevant to a Web page.

Finally, it is important to notice that there are situations when even relevant association between ads and page contents could lead to improper, often harmful, advertising. For example, it is not a good idea to place ads of a brand on a Web page that criticizes the brand itself or uses it to promote its competitor by showing the brand inferior. Matching systems must be able to filter out such pages. However, manual editorial control is often employed to handle such extreme cases.

4.2 Ranking Mechanism

Even though the matching subsystem finds out the related ads to a user query or a particular Web page, the ranking subsystem puts the ads into real life business context by ordering them promoting monetization for the publishers and the broker without hurting the interest of the advertisers and the users. A good ranking mechanism must meet the interests of all the actors in the search advertising network in a fair and transparent way.

The main function of the ranking process is to order the already selected ads in such a way so that the advertisers who pay the most get their ads placed at the top of the paid lists. In the mean time, it must also consider whether the ads are related to the user interest at all. Since there are only a few spots in the paid list, if an ad is not relevant enough to interest any user, then the publishers and the broker will lose revenue as well as the users will have negative impression about the Web site and the advertisers will not get the customers they are looking for. So the ranking algorithms must balance between the requirement of placing the top bidders higher than others, while increasing click-through rate from users.

Authors in [3, 9, 10] present detailed studies of such ranking algorithms for keyword-targeted advertising. Based on the bids (v) by the advertisers for a particular keyword K and the click-through rate (α) they present four alternative ranking mechanisms: rank by willingness to pay (v), rank by the product $v \times \alpha$, rank by click-through rate (α), and posted-price mechanism, where the broker sets a reserve price for each slot in the paid list for a period of time. But the predominant choices for ranking in search advertising networks are the following two.

4.2.1 v Ranking: Highest Players at the Top

This mechanism allocates slots in the paid list based on the advertisers' willingness to pay i.e. their bids (v). For a particular keyword, each advertiser bids and top k bidders get their ads placed on the paid list for that keyword. The payments are made according to a *first price* auction system, that is, the advertiser pay what they bid. This ranking strategy focuses on finding a slot for an ad without considering the impact of order within the paid list that in turn influences its click-through rate. This is a simplified form of the ranking approach used by the first keyword-targeted advertising provider Overture.

4.2.2 $v \times \alpha$ Ranking: Relevance and Bid Price Jointly Determine Rank

In this mechanism, the ads with the highest bids are selected and ranked according to the product of their bids and their expected click-through rate ($v \times \alpha$). The actual payments are made according to a variant of a *second price* auction system, where the winner pays the bid just below its own bid. In the case of a tie, it pays an amount that just exceeds the bid of its immediate opponent according to the ranking.

This ranking mechanism was introduced by Google and it is found to perform the best among all four alternatives [9, 10]. Because of its use of click-through rate, it can support dynamic ranking based on user feedback. The more users click on a particular ad, the more likely it is that the ad is relevant. The users here act as a large panel of human editors. Normally, ranking of the ads change linearly with click-through rate. But the authors in [10] proposed a weighted change mechanism to put more weight when a user clicks on a lower ranked ad. This addition makes convergence to optimal ranking faster and is more stable.

4.3 Search Term Suggestion

Besides the major two components, matching and ranking, there are some other issues that require IR attention. Search term suggestion is one of those. Users can express their interest in many ways due to the richness of natural languages. It is not always possible to think of all the relevant keywords for a particular product or service by the advertiser. Search term suggestion systems help the advertisers to select terms that might be of interest to them based on their already expressed interest in more common and well-known keywords.

The search volume per unit time of search terms exhibits a long-tailed distribution [2]. While a small number of search terms account for a large volume of searches, a large number of terms with extremely low search volumes, when aggregated, can account for a significant percentage of total number of searches. Since those infrequent search terms individually account for very low search volume, demands for those are not as high as for the highly frequent keywords; in short, infrequent keywords are cheaper to bid on. When an advertiser has a goal of capturing a fixed volume of search traffic, it is profitable for the advertiser to take advantage of this situation by spending less money to bid on many low frequent terms rather than competing for the more expensive high frequency search terms.

Sources of data that might be used to suggest keywords can be [2]: advertiser database, search click logs, search session logs, search resultant similarity, search resultant content similarity etc. The widely used method for recommendation systems is collaborative filtering framework. The authors in [12] presented a

singular value decomposition (SVD) based approach to suggest keywords/terms. They used latent semantic indexing (LSI), an SVD-based method, to find term-term similarity where terms are represented as vectors in a space of all advertisers with nonzero entries corresponding to advertisers bidding on that keyword. The similarity between keywords is calculated as the cosine of the angle between their corresponding vectors. The resultant system provided a smoothly controlled level of "generality" of suggested terms. the authors in [2] presented a logistic regression method trained using search logs that performs statistically similar to a standard collaborative filtering method for search term suggestions.

It is also possible to suggest other keywords by identifying the cluster to which a particular keyword belongs. The authors in [5] evaluated two clustering methods to determine: (1) groups of keywords that belong to the same marketplace, and (2) sub-markets of advertisers with common bidding behavior. They represented the advertisers and the keywords they bid on as a bipartite graph, where there is an edge between an advertiser node and a keyword node only if the advertiser bids on that keyword. The first concept was that the advertisers with common interests will bid on the same subset of keywords forming strongly connected subgraphs between themselves and the subset of keywords they bid on. A flow-based graph partitioning method was employed to find such subgraphs. For the second method, they created a non-bipartite graph containing keywords from the original graph with edges weighted proportional to the amount of overlapping between their set of bidders. Then they performed clustering using a agglomerative method. Their results showed that the first approach is better for small number of large clusters, while agglomerative clustering works better for large number of small, highly specific clusters. A hybrid of both methods might have been be a better solution.

4.4 Adversarial Aspects of Search Advertising

Even though the business model of search advertising has proven to be very effective for all the participators, it is not free from misuses. Since the revenue here is directly related to the user traffic and the more there are clicks on ads the more the advertisers will pay, a unique form of adversarial IR phenomenon known as *click fraud* has become popular among many dishonest publishers who simulate fraudulent clicks to increase their revenue [7, 16]. The problem is serious because if the broker cannot protect its advertisers from such misconducts, the advertisers could eventually lose confidence in the ad network, which could harm business for all the associated parties.

To eliminate this problem, researchers have been trying to characterize fraudulent clicks. The problem of successfully detecting such clicks with high precision is very difficult because it has to be made sure that the advertisers do not have to pay for fake traffic and the broker do not miss revenue from discharging valid clicks. It might seem that intervention of human judges is the best way out. But the author in [7] suggested use of unlabeled data to train such detection mechanisms; because it is, in fact, impossible to label millions of clicks for different keyword-ad combinations. Other proposals to solve the click fraud problem include using Pay-Per-Action paradigm instead of Pay-Per-Click, blocking blacklisted IPs and aggressive monitoring etc. In general, this front of search advertising research is still in its infancy and needs a lot of work to reach a satisfactory and reliable state.

5 Conclusion

In this paper, we reviewed the key concepts related to search advertising. We presented a full taxonomy of online advertising for the first time, and identified the positioning of search advertising in that taxonomy. After that, we presented a generic model of search advertising and discussed the main actors that constitute the search advertising network. Based on the discussion of the generic model, we presented the main categories of search advertising: keyword-targeted advertising and content-targeted advertising. Then, we presented a conceptual model of search advertising system identifying the main components: the matching subsystem and the ranking subsystem.

At that point, we dived into advanced IR-related issues of search advertising. We reviewed the available matching strategies for both the keyword-targeted and the content-targeted advertising, and pointed out some research opportunities in those directions. Later, we elaborated on the ranking mechanisms available in the literature. We also presented research results from search term suggestion systems and discussed the growing field of adversarial aspects of search advertising.

Even before a decade of its inception, search advertising has placed itself at the top of the list of online monetization sources [15]. But, surprisingly enough, there is very little academic research available. The primary goal of this study was to provide a comprehensive view of the search advertising area, as a whole, and to stimulate further research in this intriguing area.

References

[1] Giuseppe Attardi, Andrea Esuli, and Maria Simi. Best Bets: thousands of queries in search of a client. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters (WWW Alt.'04)*, pages 422–423, New York, NY, USA, 2004. ACM.

[2] Kevin Bartz, Vijay Murthi, and Shaji Sebastian. Logistic regression and collaborative filtering for sponsored search term recommendation. In *Proceedings of the Second Workshop on Sponsored Search Auctions, 2006*, 2006.

[3] Hemant K. Bhargava and Juan Feng. Paid placement strategies for internet search engines. In *Proceedings of the 11th International Conference on World Wide Web (WWW'02)*, pages 117–123, New York, NY, USA, 2002. ACM.

[4] 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.

[5] J. Carrasco, D. Fain, K. Lang, and L. Zhukov. Clustering of bipartite advertiser-keyword graph. In *Proceedings of International Conference on Data Mining*, 2003.

[6] Marco Cristo, Berthier Ribeiro-Neto, Paulo B. Golgher, and Edleno Silva de Moura. Search advertising. *Soft Computing in Web Information Retrieval*, 197:259–285, 2006.

[7] Elena Eneva. Detecting invalid clicks in online paid search listings: A problem description for the use of unlabeled data. In *Workshop on the Continuum from Labeled to Unlabeled Data, Twentieth International Conference on Machine Learning*. AAAI Press, 2003.

[8] Daniel C. Fain and Jan O. Pedersen. Sponsored search: A brief history. In *Proceedings of the Second Workshop on Sponsored Search Auctions, 2006*, 2006.

[9] Juan Feng, Hemant K. Bhargava, and David Pennock. Comparison of allocation rules for paid placement advertising in search engines. In *Proceedings of the 5th international conference on Electronic commerce (ICEC'03)*, pages 294–299, New York, NY, USA, 2003. ACM.

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

[11] Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. Domain-specific keyphrase extraction. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99)*, pages 668–673, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[12] David Gleich and Leonid Zhukov. SVD based term suggestion and ranking system. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 391–394, Washington, DC, USA, 2004. IEEE Computer Society.

[13] David Green. Search engine marketing: Why it benefits us all. *Business Information Review*, 20(4):195–202, December 2003.

[14] Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[15] IAB and PricewaterhouseCoopers. IAB internet advertising revenue report, May 2006. Available at, http://www.iab.net/media/file/resources_adrevenue_pdf_IAB_PwC_2006_Final.pdf.

[16] Bernard J. Jansen. Adversarial information retrieval aspects of sponsored search. In *AIRWeb'2006*, pages 33–36, 2006.

[17] D. Kelleher and S. Luz. Automatic hypertext keyphrase detection. In *IJCAI'2005*, 2005.

[18] Anísio Lacerda, Marco Cristo, Marcos André Gonçalves, Weiguo Fan, Nivio Ziviani, and Berthier Ribeiro-Neto. Learning to advertise. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'06)*, pages 549–556, New York, NY, USA, 2006. ACM.

[19] EC^2 @ University of Southern California. Internet advertising history, 2001. Available at, http://www.ec2.edu/dccenter/archives/ia/history.html.

[20] Jeffrey Parsons, Katherine Gallagher, and K. Dale Foster. Messages in the medium: An experimental investigation of web advertising effectiveness and attitudes toward web content. In *Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS'00) - Volume 6*, page 6050, Washington, DC, USA, 2000. IEEE Computer Society.

[21] J. R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[22] 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.

[23] J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. PrenticeHall, 1971.

[24] Reuters News Service. Key dates in the history of Google, 2004. Available at, http://www.forbes.com/business/businesstech/newswire/2004/04/29/rtr1353500.html.

[25] 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, New York, NY, USA, 2006. ACM.

[26] P. Turney. Coherent keyphrase extraction via web mining. In *IJCAI'2003*, pages 434–439, August 2003.

[27] Peter D. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336, 2000.

[28] Chingning Wang, Ping Zhang, Risook Choi, and Michael D'Eredita. Understanding consumers attitude towards advertising. In *Proceedings of the Eighth Americas Conference on Information Systems*, pages 1143–1148, 2002.

[29] Melius Weideman. Ethical issues on content distribution to digital consumers via paid placement as opposed to website visibility in search engine results. In *The Seventh ETHICOMP International Conference on the Social and Ethical Impacts of Information and Communication Technologies*, pages 904–915. Troubador Publishing Ltd, April 2004.

[30] Melius Weideman and Timothy Haig-Smith. An investigation into search engines as a form of targeted advert delivery. In *Proceedings of the 2002 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology (SAICSIT'02)*, pages 258–258. South African Institute for Computer Scientists and Information Technologists, 2002.