Towards Personalized Advertising in Sponsored Search

Ahmad I. Zainal Abidin
Department of Computer Science, University College London
Malet Place, London WC1E 6BT, UK
a.zainalabidin@cs.ucl.ac.uk

Abstract

Web advertising is one of the major sources of income for numerous search engines, news sites and non-commercial publishers. Textual ads, characterized by *Sponsored Search (SS)* and *Content Match (CM)*, make up a significant portion of Web advertising. In *SS*, with limited information about ad contents, given a query, the challenge is to place relevant ads alongside organic search results. Organic search results are ranked based on their relevance to search keyword. However, *SS* results are not necessarily ranked purely based on relevance due to various factors influencing the ads overall ranking such as bid phrase and displayed position. The displayed ads may not relate to a user's information need. In this paper, a study associating ads and users, referred to as personalized advertising is proposed. User profiles are used as external knowledge to establish the relationship between the users and the ads.

Keywords: Web Advertising, Sponsored Search, Information Retrieval, Search Engine

1. INTRODUCTION

The Web has been useful to Internet users around the world for transactional and non-transactional purposes such as information seeking, communication and online shopping. When seeking for information, a query is submitted to a search engine that later displays results relevant to the query. It is estimated that about two billion searches are performed daily by Internet users around the world [1] with various information needs. The trend of using the Web for non-transactional and transactional activities continues to increase, enabling advertisers to reach out more potential customers through Web advertising – one of the major sources of income for search engines, commercial publishers (e.g. news sites), and non-commercial publishers (e.g. blogs). Typically, the ads are distributed through either *Sponsored Search (SS)* or *Content Match (CM)* method. In *SS* (also known as *Paid Search Advertising*), textual ads are placed at the result pages for the query that a user has submitted. An example is as shown in Figure 1.

FIGURE 1: Sponsored advertisements listed at the result page of a query

Currently, most Web advertising consists of textual ads with three elements: an ad *title*; a few lines of *description* relating to the title, promoted products or services; and a URL to the advertisers site called the *landing page*. Major search engines such as Google and Yahoo! generate income through textual ads placed at their search results. Basically, they act as search engines that also provide ad networking services. In the United States, search advertising shared about 45% of the total \$23.4 billion generated by Web advertising for the full year of 2008 with overall Web advertising revenue for that year increased nearly 11% over 2007. The remaining 55% of the total \$23.4 billion was generated by a combination of display-related advertising (33%), classified advertising (14%) and other advertising formats (8%) [2]. Search advertising is therefore in an advantageous position for advertisers to reach out customers and for search engines to generate advertising revenues. Internet users who initially use the engines to seek information are the targeted group to interact with the advertisers' ad campaigns. The format of SS advertising is less annoying compared to unsolicited commercial emails.

The second method of textual ads in Web advertising is *CM* (or *Contextual Advertising*) which places ads within the content of a Web page. In *CM*, an ad network is involved with the goals of optimizing relevant ads and revenues shared between the Web page owner (e.g. blogger) and the ad network. However, *CM* is not the focus of this paper.

Advertisers bid for keywords, known as *bid phrase*, to be associated with their ads. The position where the ads are being displayed is also a factor in determining how much the advertisers are being charged. The three pricing models for textual ads are *pay-per-click (PPC)*, *pay-per-impression (PPI)*, and *pay-per-action (PPA)*. In *PPC* pricing model, advertisers will be charged a certain amount for any clicks on the URLs provided in their ads. For *PPI*, advertisers will be charged every time their ads are displayed, and for *PPA*, advertisers only pay if users visit the advertisers Web page and later perform transactions.

Although the displayed ads relate to the query that a user has submitted, the ads do not necessarily relate to the user as a potential customer. For instance, suppose that users *A* and *B* type "wimbledon 2009" as a query, *A* may be interested in one of the ads shown in Figure 1, but *B* may not. The reason would be that *B* likes travelling in a group so he or she is looking for a tour package not just a ticket. Apparently *A* and *B* have different profiles. Motivated by this problem, a study on personalized advertising in SS is worth carried out.

This paper focuses more on SS advertising and aims at providing an overview of techniques relevant to SS. The organization of the paper is as follows. Section 2 describes an overview of SS advertising, section 3 related works to SS advertising, section 4 future challenges, and section 5 concludes the paper.

2. SPONSORED SEARCH ADVERTISING

Sponsored search advertising mainly concerns about marketing products and services thought to be of interest to a user. Although the user's interests are subjective, a query the user submits to the search engine represents his or her information need. The difference between SS and brand marketing is that the latter promotes specific brands while SS promotes products and services directly based on prior knowledge, i.e. query keywords.

Typically, SS involves three entities namely *advertiser*, *user*, and *search engine* which each has a different goal:

- The *advertiser* provides ads that each consists of a *title*, *description*, and *landing page*. The goal is to promote products and services within a specific campaign period or temporal constraint so that potential customers are reached out as many as possible. For example, Sponsored links in Figure 1 revolves around Wimbledon 2009, in which one of the advertisers sells Wimbledon tickets. So, the ad theme is relevant to the user's query.
- The *user* views ads, and if he or she becomes interested, he or she will visit the advertiser's web page. The user has an information need and if the ads are relevant, the probability that the user will interact with the ads is higher compared with if the ads are not relevant. A study

in [3] confirms that the relevance of ads with user interest is important. In [4], it is evident that the ad and its context have a significant effect even when a conscious response such as click is not present.

- The *search engine* selects ads based on the user's query and places them on the result page. The goal is to select relevant ads so that users might be interested in the ads. As the search engine earns advertising revenue, maximizing click-through rate is therefore desirable. In this case, the pricing model is assumed to be *PPC*, in which when a user clicks an ad, the advertiser is being charged.

User profiles are important so that ad campaigns will be more personalized across different users. Given different users submitting the same search keyword relatively at the same time, all users should view different ads depending on their profiles. Basically, an advertiser interacts with a search engine ad networking service to add or update its ads. The engine stores the ads in its storage. A user interacts with the search engine that records the user's Web browsing and search activities. These activities are useful in formulating the user's profile that is used as an additional knowledge to relate different users with different targeted ads.

The click revenue R that a search engine can estimate for a given query q and a set of ads $A = \{a_1, a_2, \dots, a_n\}$ is expressed as

$$R = \sum_{i=1}^n P(\text{click}|q, a_i) \cdot \text{Price}(a_i, i) \quad (1)$$

where $P(\text{click}|q, a_i)$ is the click probability given the query q , the set of ads A , the total number of ads displayed n and $\text{Price}(a_i, i)$ is the ad a_i click price at position i .

The main challenge with *SS* advertising is that a query is short, typically 2-5 terms, and ad contents are limited. Initially, when a user submits a query, he or she anticipates the best search results not the best ads associated with his or her query terms. Due to the brevity of the query terms and the ad contents, the best scenario would be an *exact match* between the ad bid phrases and the query terms.

One of the possible approaches to rank query q with regard to the ad a_i is by computing their cosine similarity:

$$\text{sim}(q, a_i) = \frac{\vec{q} \cdot \vec{a_i}}{|\vec{q}| \times |\vec{a_i}|} = \frac{\sum_{i=1}^n w_{iq} \cdot w_{ij}}{\sqrt{\sum_{i=1}^n w_{iq}^2} \sqrt{\sum_{i=1}^n w_{ij}^2}} \quad (2)$$

Although many bid phrases may be associated with an ad, it is impractical for advertisers to provide an exhaustive list of bid phrases for their ad campaigns. Therefore, a *broad match* is an alternative to the *exact match* where a query is related to but does not have to match the phrases. In *broad match* advertising, the criterion is relaxed by not relying solely on the phrases. For example, the ad title, description, and URL for the landing page are typically used to be matched against the query. Additionally, information related to the user or the advertiser may also be considered.

3. RELATED WORK

Broder et al. [5] use a global score threshold in their work to predict whether the entire set of ads is relevant to be displayed. For any query, ads with scores higher than the set threshold are perceived to be more relevant than those with lower scores and ads with scores lower than the threshold should not be returned. So, different thresholds will result in different suggested number of ads to be returned. Clearly, there is a trade-off between the effectiveness and the coverage of ad results. The drawback of this method is that it is difficult to set a threshold that leads to a set of optimal ads because the result coverage also depends on the ad corpus. The follow up work uses machine learning technique to classify whether an entire set of ads is relevant [5]. The input

TABLE 1: Categorization of methods for ad selection

Approach	Representative Works
Machine learning	<p>Predicting whether the entire set of ads is relevant to be displayed [5]</p> <p>Using click data for training and evaluation, determining which framework is more suitable, and determining useful features for existing models [9]</p>
Threshold	Global threshold to determine whether to return ads [8]
Query expansion, augmentation and substitution	<p>Rewriting of tail queries [6]</p> <p>Using knowledge in search result to create augmented query [8]</p> <p>Optimizing ad relevance and revenue using query substitution [11]</p>
Click-through and click feedback	Estimating the click-through rate [12]

is a query and a set of ads, while the output is either “yes” if the entire set should be displayed or “no” otherwise. In their classification model, they use Support Vector Machine.

In SS, query rewriting, done mostly offline using various sources of external knowledge, is a common technique used to perform *broad match*. Repeating queries from the head to torso of the query volume will benefit from this approach. Online query rewriting for SS was proposed in [6]. In their work, they use pre-processed queries to develop inverted index of the expanded query vectors and run incoming queries against it to enrich the original tail query representation. In their work to rewrite queries for SS, Jones et al. [7] used search engine query logs to gather user session information and later produce alternative queries for ad selection. Candidate substitutions are first generated by examining pairs of successive queries that user issue in the same session. Subsequently, the candidates are examined to find common transformations. A machine learning ranking is used to determine the most relevant rewrites that match against ads. In [8], the authors use the search results (*top-n* pages) so that an additional knowledge about the query submitted by the user can be gathered. The additional knowledge is then used to augment the query. The augmented query is later evaluated against the ad corpus to retrieve relevant ads. In [9], the authors use machine learning approach to predict the likelihood of an ad is to be clicked. Knowing which ads likely to be clicked is very important to both the advertisers and the search engines.

Given a set of relevant documents and a set of irrelevant documents, the work on query expansion by Rocchio [10] shows how the best query is crafted. With regard to SS, documents refer to ads. Rocchio’s expanded query consists of a linear combination of the original query, and the vectors for relevant and irrelevant documents. The vectors for relevant documents are added to the original query, while the vectors for irrelevant documents are subtracted from the original query. Radlinski et al. in [11] propose a query substitution approach to optimize relevance and revenue in SS. They analyze frequent queries in pre-processing phase that constructs lookup table which will map queries into bid phrase. User original queries are transformed to substituted queries that have higher matching chances with bid phrase. Richardson, Dominowska and Ragno in [12] propose a method to estimate the click-through rate for new ads. They develop a model using logistic

regression. Their dataset consists of information on landing page, bid term, title, body, display URL, clicks, and views for each ad. The dataset is used to train and test their regression model.

Recent research works on SS advertising are classified in Table 1.

4. FUTURE CHALLENGES IN SPONSORED SEARCH

The problem in Web advertising will remain chiefly on selecting relevant ads for a given query especially the ads volume is expected to increase. However, the research and applications in Web advertising are dominated by search engine companies due to advertisement datasets and ads bidding networks are available to these companies. Search engines generate the ads relevance and ranking by their proprietary algorithms which lack public evaluations or involvement of researchers in academia.

When a user submits a query, the expectation is to have the best search results based on their relevance to the query. SS results are influenced by various factors: the relevance between the user's query and the advertiser's bid phrase, the bid amount for the phrase, and the position chosen to display their ads. Each of these factors has its own weight in determining the ads overall ranking. Therefore, studying the trade-off between ads relevance and advertising revenue is essential. An ad positioned in the top list may be due to the dominant weight assigned to its bid phrase. The click-through data are worth analyzed to study the number of viewed ads that are later clicked regardless of their displayed positions.

While research works on query expansion and query substitution have been done to augment or substitute original query to improve ads relevance, the displayed ads may not necessarily be relevant to users' interests. Therefore, given the same query, two different users with different Web browsing and search interests should be displayed with different ads. Therefore, learning consumer behaviors that represent interests will be important in personalized advertising. Click-through logs that have the data on users' web browsing and search interests can be used to model behavioral profiles. However, monitoring users' web access raises a privacy concern. If their Web access activities are being monitored for marketing purposes, users should be made aware of the monitoring usage. Perhaps, users should be rewarded for allowing their web access activities to be monitored. A study on the technical, legal and ethical implications of behavioral profiling needs to be addressed.

Most researches in Web advertising focus on issues related more to search engines. To the best of my knowledge, there has not been much research on advertiser-oriented Web advertising in the literatures. Generally, advertisers have the budgets to run their campaigns. It is important to optimize advertisers' satisfactions and to minimize their advertising costs. The metrics to measure the advertisers' satisfaction need to be developed.

5. CONCLUSION

This paper provides an overview of Web textual advertising especially on sponsored search, the recent research works in sponsored search and its potential future directions. The author views that personalized advertising in sponsored search is essential because different users will view ads relevant to their interests and information needs. It is known that an advertising revenue is a driving force in sponsored search. So, the displayed ads in sponsored search may not necessarily be the most relevant ads to a user's query or interest. Therefore, improving ads relevance that relate to users as potential customers may be more rewarding to the advertisers than displaying ads that no users relate to.

REFERENCES

- [1] 61 billion searches conducted worldwide in August. October 2007. Available from <http://www.comscore.com/press/release.asp?press=1802>

- [2] IAB Internet advertising revenue report: 2008 full-year results, March 2009. Available from http://www.iab.net/media/file/IAB_PwC_2008_full_year.pdf
- [3] C. Wang, P. Zhang, R. Choi, and M. D'Eredita. Understanding consumers attitude toward advertising. In *8th Americas Conference on Information Systems*, 2002
- [4] C. Y. Yoo. Preattentive processing of Web advertising. *PhD Thesis*, University of Texas at Austin, 2005.
- [5] A. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras. To swing or not to swing: learning when (not) to advertise. In *CIKM08*, page 1003-1012, 2008
- [6] A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, D. Metzler, L. Riedel, and J. Yuan. Online expansion of rare queries for sponsored search. In *WWW09*, 2009
- [7] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *WWW06*, page 387-396, 2006
- [8] A. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel. Search advertising using web relevance feedback. In *CIKM08*. Page 1013-1022, 2008
- [9] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *WWW08*, page 227-236, 2008
- [10] J. J. Rocchio. (1971). *Relevance feedback in information retrieval*. Prentice Hall: Englewood Cliffs, New Jersey
- [11] F. Radlinski, A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, and L. Riedel. Optimizing relevance and revenue in ad search: a query substitution approach. In *SIGIR08*, page 403-410, 2008
- [12] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW07*, page 521-529, 2007