A FRAMEWORK FOR A MULTILINGUAL CONTEXTUAL AND BEHAVIORAL ONLINE ADVERTISING NETWORK: A CASE STUDY

Domen Košir[1,2], *Zoran Bosnić*[1], *Igor Kononenko*[1]
[1] Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
[2] Httpool Ltd., Cesta v Gorice 8, 1000 Ljubljana, Slovenia
e-mail: domen.kosir@fri.uni-lj.si

ABSTRACT

An online advertising network connects web content providers and advertisers enabling the providers to monetize their content and the advertisers to reach online consumers. In this paper we study the workings of a successful state-of-the-art online advertising network with branch offices all over the world. The framework is designed to support different targeting strategies and advertising in multiple languages. We describe the implementation of contextual and behavioral targeting, and also discuss different methods to evaluating these strategies. Statistics show that by employing contextual targeting instead of random targeting we can achieve significantly higher CTR values.

1 INTRODUCTION

Online advertising is an increasingly popular form of advertising that uses the internet to deliver advertisments to consumers. An online advertising network acts as intermediary between advertisers and website managers (publishers). The goal of an online advertising network is a better utilization of the available advertising space.

In so-called “block web advertising” websites typically have some parts reserved for advertisments - these are called advertisment blocks. Each advertisment block is usually of a fixed size and can display one or more advertisments at a time.

The publisher can choose to maintain control over the selection of advertisments which will be displayed to the website visitors or he can give control over the advertisment blocks to an advertising network. This way the publisher does not have to deal with the advertisers in order to ensure that the advertisments shown on his website are interesting and diverse.

Modern advertising networks are often specialized for a specific targeting strategy of which most common are:

- contextual targeting [6][2] and
- behavioral targeting [1].

In contrast to random targeting, both mentioned strategies make an effort to find the best possible advertisment for the user who is currently viewing the website. Contextual advertising is based on web content analysis. If the website currently viewed has already been analyzed by the network then advertisments most similar to the website content are displayed. Behavioral targeting on the other hand focuses the user, his habits and previously viewed content. This type of advertising obviously is not possible without some form of user tracking (e.g. via HTTP cookies or IP addresses). The tracking of users itself can be a privacy issue but it will not be addressed here.

This paper is a case study of a successful international state-of-the-art advertising network that supports both contextual and behavioral targeting. The network operates in many countries around the world and is still expanding. So far, it supports more than 10 Indo-European languages.

In Section 2 we describe the network's architecture and its main components. Sections 3 and 4 describe how contextual and behavioral targeting are implemented. In Section 5 we describe different approaches to evaluation of targeting strategies or an advertising network as a whole. WeWe conclude our work in Section 7.

2 FRAMEWORK ARCHITECTURE

As it is shown in Figure 1, the framework consists of three main parts:

- Advertisment server – This is the front-end component which recieves requests from websites and responds by sending back advertisments.
- Web crawler – A background process that browses the web and analyzes the web content.
- Database – It contains information about the ongoing advertising campaigns, analyzed web content and web users. This distributed document-oriented database is exposed as a web service and is highly scalable.

When a user visits a website which is a part of the advertising network:

1. A request is sent to the advertisment server. Each request, among other pieces of information, contains the URL address of the website and user identifiable information.
2. The server then searches the database for information about the website and/or user (depending on the

targeting strategy used). If no information about the website is found the server sends this URL address to the web crawler to be inserted into its priority queue.

3. Using all the available information, the server then chooses the most appropriate advertisments and sends them back to be displayed on the website.

The request times and the origin URL addresses are sent to the web crawler. Its job is to constantly fetch web content, analyse it and update the framework's database.

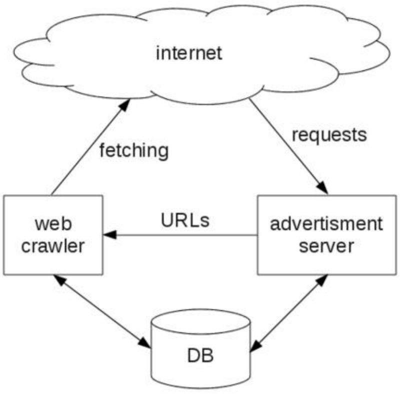

Figure 1: *The architecture of the advertising network.*

2.1 Web Crawler

Any kind of advanced advertisment targeting requires some information about the website content. Network delays and request peaks make fetching of the website content at request-time infeasible. Fetching of website content is done in a separated process using a web crawler. This is an independent software package that visits certain webpages and saves their contents. The framework uses a very straightforward URL-oriented approach using:

- a priority queue used to visit URLs in an orderly fashion,
- a categorization system to process the website content and
- a database to save all the information about each URL.

The priority queue contains all the URL addresses from which requests for advertisments originated. The priority of each URL is affected by:

- Age of the URL – New URL addresses (from the advertising network's point of view) have a higher priority while older have lower priority.
- Frequency of changes in the website's content – More frequently changing websites are fetched more often.
- Number of requests – More popular websites have a higher priority than the unpopular ones.

After the crawler fetches the content of a website the content must be processed (see Figure 2).

1. HTML code cleanup – The fetched web content are usually HTML documents. Since the quality of HTML code varies greatly it must be cleaned up before further processing.
2. Text extraction – The HTML document is parsed and text is extracted. We focus mainly on bigger parts of text with few HTML tags in order to reduce the noise caused by menus, sidebars, footers, etc.
3. Language identification – This step is optional but may be needed for the categorization process. A very robust approach to language identification is with the usage of n-grams [3][4]. N-grams are fast, consume very little space and work reliably in spite of textual errors.
4. Text categorization - The categorization process maps each document to one or more categories based on its content. For the purpose of providing high-level overview of the system, in the following we will consider the text categorization process as a black box function that takes text as input and produces a list of categories with corresponding biases.

After the web document analysis, the contents and categorization results are saved in the network's database to be used by the advertisment server.

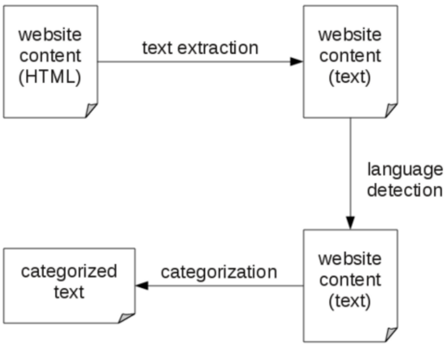

Figure 2: *The process of website content analysis.*

2.2 Advertisment Server

The server's job is to recieve the requests from the internet, select advertisments and send them back for display. Each request comes from a specific advertisment block on the website. The request contains the following information:

- the URL address of the website,
- the size of the advertisment block,
- the publisher's preferences regarding the targeting strategies used to select advertisments,
- the publisher's preferences regarding the types (e.g. textual, animated graphics, Flash banners) and colors of advertisments to be displayed on his website,
- the viewer's language preferences,
- the viewer's IP address and possibly other user identifiable information.

The process of selecting the appropriate advertisments can be a little tricky. For example, there is no obvious rule specifying in which language the displayed advertisments should be. If a German-speaking user is viewing an English website then the language of the web content probably does not match the user's language preferences.

We can either decide to match the language of advertisments to the website's language and show him English advertisments or we can consider the user's preferences and show him advertisments in German language.

This is of course only an issue when we have an international network with advertisers and publishers from different countries.

3 CONTEXTUAL TARGETING

The goal of contextual targeting is to display advertisments that match the website content. If we assume that our web crawler has already fecthed and analyzed the content, currently viewed by the user, we can select the advertisments in the following manner:

1. Like the web content all the advertisments should also be categorized. The advertisments can be categorized based on their content (using advertisment title and text or image analysis) or they can be categorized by manually assigning them to specific categories. In either case, it is important that the same set of categories is used as in the web content analysis.
2. Depending on the size of the block we select a number of advertisments. The selection of contextual advertisments is illustrated in Figure 3. A categorized web document and all the advertisments are placed in a n-dimensional vector space where each dimension represents a single category. The web document and the advertisments are vectors with the initial point in the origin and the terminal point defined by the categorization results. We use cosine similarity [4][7] to compare the vector representing the web document to the vectors representing the advertisments. We select the advertisments closest to the web document.

Cosine similarity is used in order to reduce the impact of fluctuations of categorization results which may occur because of:

- varying amount of text in web documents,
- varying number of meaningful words in a web documents (e.g. spam sites typically contain many keywords while other websites do not) and
- the difference in size of a web document and an advertisment.

4 BEHAVIORAL TARGETING

Behavioral targeting is based on the analysis of the user's behavior. This is a very broad definition and there are numerous strategies that fall under it.

A very common approach in web user profiling is building of two user models [5]. One model represents the user's short-term interests and is derived from the users activities in the past few days. The other model is derived from user's older web browsing history and represents his long-term interests (see Figure 4). By building two models to describe the user's interests we can separate his true interests from the interests influenced by events like birthdays, holidays, car purchase, interesting news, etc.

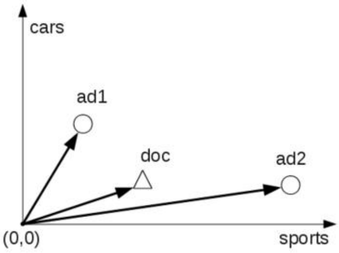

Figure 3: *In this example of selecting advertisments using contextual targeting we limit ourselves to only two categories: cars and sports. We are trying to select the most appropriate advertisment (1-NN) for the website "doc". For the purpose of contextual advertising we are more interested in the ratio of biases in the categorization results than in their absolute values. That is why we use the cosine similarity which calculates the cosine of the angle between two vectors. By using the cosine similarity we get "ad2" as the best match for "doc".*

There are several other attributes that influence the user's behavior and can be taken into account when modeling the user:

- The time of day – We can expect that many users will show different interests during various activities like working, studying, playing games, etc. These activities are often performed at the same time of day.
- The day of the week – The first thing we want to do is to distinguish between the workdays and the weekend. A more comprehensive analysis of the user's behavior is possible by looking at his day-to-day activities.
- The user's IP address – Given that we are able to reliably identify a user without using his IP address (e.g. using web cookies), the address can still provide some information about the user's interests. A user with a laptop computer will always be identified by the same web cookie but by using his IP address we can distinguish among his interests related to his workplace, home, favourite internet cafe, etc.

These attributes are somewhat problematic – we can only use them in the user behavior analysis after we have tracked the user for an extended period of time.

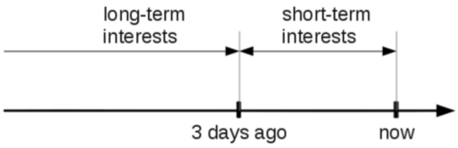

Figure 4: *The division between the user's short-term and long-term interests. The boundary at "3 days ago" is just illustrative.*

5 EVALUATING TARGETING STRATEGIES

Advertising campaigns are usually planned toward a specific goal (e.g. trademark promotion, new product promotion, increasing sales). Depending on the goal a pricing model is chosen. The most common pricing models in the online advertising are:

- PPI (Pay Per Impression) - The advertiser pays for each appearance of his advertisement on a website.
- PPC (Pay Per Click) - An online advertising pricing model, where the advertiser pays for each click on their advertisement. The click usually redirects the user to the advertiser's website.
- PPA (Pay Per Action) - The advertiser pays to the publisher for each specified action (e.g. a purchase or a form submission) that results from his advertisment.

One way to measure the success of a targeting strategy is to sum up all the profits made using this strategy. This value is undoubtedly very informative but it is not an objective measure as it is influenced by the varying advertisment prices. Because we do not want the prices to influence our evaluations, we only monitor the users' engagement. We do this by looking at all the clicks and other actions users performed.

Click-through rate (CTR) is a standard way to measure the success of an online advertising campaign or an individual advertisment. It is calculated as the ratio between the number of clicks on the advertisment and the number of times it was shown to the users. It tells us the probability a user will click on the advertisment when he sees it.

We have to be careful to remember whether we are dividing the number of clicks with impressions or displays when calculating the CTR. Most advertising networks do not distinguish between these two events. An *impression* happens when the server delivers an advertisment to the webpage and HTML code of the advertisment is rendered by the user's computer. A *display* happens when the advertisment is actually shown on the user's monitor for a minimum of 2 seconds. If the webpage is larger than the user's monitor and if the advertisment block is not located near the top of the page where the user can see it, then a display does not happen until the user scrolls down to it. In other words: a display cannot occur without an impression. This is an important distinction, because in our experience the number of displays is typically 20-30% smaller that the number of impressions. This, of course, affects the value of CTR.

6 LOG ANALYSIS

We analyzed the advertising networks server logs for the first 6 months of 2011. In the analysis we limited ourselves to the Slovenian subnetwork – this is the primary and one of the largest subnetworks under Httpool’s management. We took all advertising campaigns that were in circulation during the period of 6 months and calculated average CTR values per advertisement type, distinguishing campaigns with random targeting from those with contextual targeting. Behavioral targering is is not yet used in production.

We found that CTR values tend to be bigger in campaigns that use contextual targeting. We saw a 28% increase in CTR for textual advertisements, a +12% increase for shop advertisements (text with a picture) and a 74% increase for rich media advertisements (Flash, animated GIF).

7 CONCLUSION AND FUTURE WORK

In this paper we described a framework for an online advertising network capable of analyzing web content and displaying advertisment using contextual and behavioral targeting.

We found that CTR values tend to be higher when we are using contextual targeting instead of random targeting.

Behavioral targeting is not yet used in production. We expect that it will prove to be an effective advertising tool.

So far, only Indo-European languages are supported by the framework. As the advertising network grows, we expect that other languages will have to be supported as well. Languages like Mandarin and Arabic will undoubtedly be a challenge.

Acknowledgment

Operation part financed by the European Union, European Social Fund. Operation implemented in the framework of the Operational Programme for Human Resources Development for the Period 2007-2013, Priority axis 1: Promoting entrepreneurship and adaptability, Main type of activity 1.1.: Experts and researchers for competitive enterprises.

References

[1] A. Addis, G. Armano and E. Vargiu. Profiling users to perform contextual advertising. In *Proc. WOA 2009.*

[2] A. Broder, M. Fontoura, V. Josifovski and L. Riedel. A semantic approach to contextual advertising. In *Proc. SIGIR 2007*, pp. 559-566, 2007.

[3] W. B. Cavnar and J. M. Trenkle. N-gram based text categorization. *In Proc. SDAIR-94*, pp. 161-175, 1994.

[4] Z. Elberrichi and B. Aljohar. N-grams in Texts Categorization. *Scientific Journal of King Faisal University (Basic and Applied Sciences),* 8(2):1428H, 2007.

[5] S.E. Middleton, N.R. Shadbolt and D.C. De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS),* 22(1):54-88, 2004.

[6] B. Ribeiro-Neto, M. Cristo, P. B. Golgher and E. S. de Moura. Impedance coupling in content-targeted advertising. In *Proc. SIGIR 2005*, pp. 496-503, 2005

[7] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.