**Project 10: Ad Position Investigation**

**1、Background**

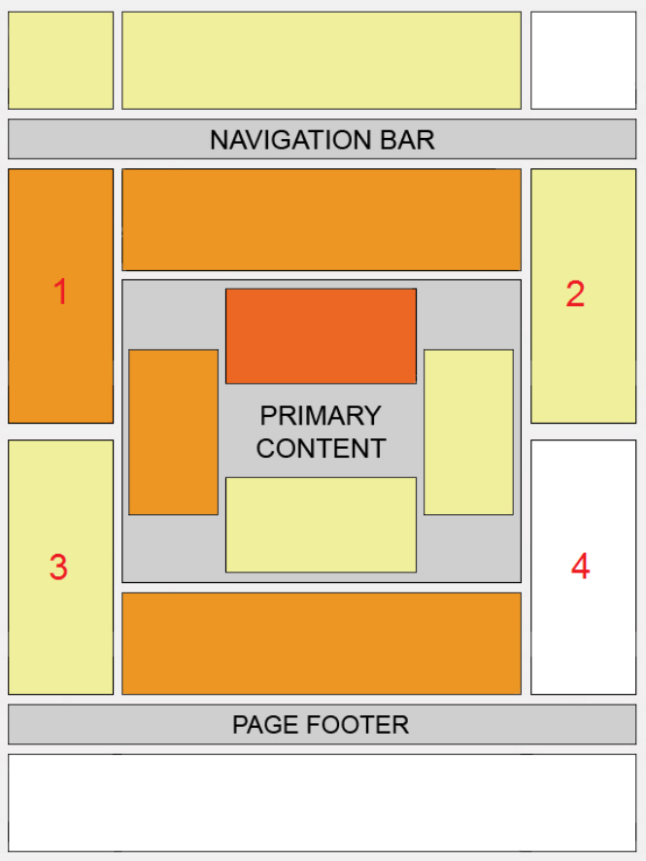
**（1）What is “ad position”?**

Ad position is the position of your ad on a search results page or content-based webpage. It's determined by how your ad ranks against competing ads. Rank is determined by many factors, including:

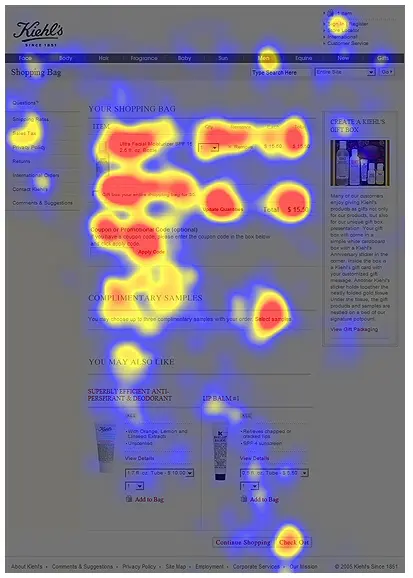
* The amount you bid for your**[keyword](https://help.ads.microsoft.com/apex/index/22/en/14009" \o ")**.
* Your ad's **[relevance](https://help.ads.microsoft.com/apex/index/22/en/14009" \o ")**.
* Your ad's performance, or [click-through rate](https://help.ads.microsoft.com/apex/index/22/en/14009" \o ") (**CTR**), on webpages.

1. When attempting to **monetize a website**–whether through [display advertising](https://monetizepros.com/display-advertising/guide-to-display-advertising/), [affiliate marketing](https://monetizepros.com/affiliate-marketing/guide-to-affiliate-marketing/), product sales, or some combination thereof–one of the most important things you’ll do is decide what goes where. While the type and quality of content that you have will go a long ways toward determining the volume and caliber of your audience, the manner in which you arrange those pieces will have a major impact on how much money you make from that traffic.

One of the basic rules of the Web is that **not all positions are created equal.** After multiple decades of Internet access (and millennia of reading), humans have become creatures of habit when it comes to digesting content. As a result, certain areas of your site will be inherently more popular than others–not because of what they contain but because of where they’re positioned. Here’s the most basic illustration of how we tend to consume the content on web pages, with darker sections illustrating sections that get more attention.



Here’s another helpful illustration to understand how visitors to your site are consuming your content–where their eyes tend to focus and which parts of the page are most likely to be overlooked completely:



So when placing your ad units, be sure to place some in areas of high visibility, namely just above and to the left of your primary content, or even within your primary content.

**2、Motivation**

**What is the goal or purpose of this study**

Not all website layouts are created equal, particularly when it comes to the potential to generate earnings. Decisions that may seem minor can end up having a big impact on how much money you ultimately make from display advertising.

For publishers and bloggers hoping to monetize their traffic primarily through display advertising, the layout of a website becomes a very important strategic decision. In addition to deciding which ad units to use, you’ll need to figure out where each should be positioned in order to maximize revenue without annoying visitors or detracting from the site experience.

Our project puts a tremendous amount of information at your fingertips; if you have the time and patience to sort through all the various statistics and features, there can be a huge opportunity to optimize your site setup and immediately boost earnings.

**3、Proposed Methodology**

**（1）Data collection：**

To **obtain the 2000 URLs**, we used a **web crawler** to search for keywords and extract their pages, a key part of which was **using page flips**

Note: At first, we chose to use the requests library to make requests and then crawl, and then found that the code of the crawled down web pages did not have those little ads next to them, after checking the information, we found that maybe these ads were rendered in post with JavaScript at a later stage resulting in. So instead we used the selenium automation library to simulate our behaviour towards browsing and then went in and crawled all the images on the page.

**（2）Image processing：**

When Image processing, it was found that there are two main types of src for images: **url form and base64 direct rendering.** The former is easier and can be downloaded directly to the local area after the image for subsequent ad detection and processing. In the latter case, we need to decode and synthesise the base64 code accordingly. In this process, we also found that the base64 code of some images was perhaps too long for the reason that the base64 code we crawled from the browser was not complete, so we could not recompose the photo.

**Relative position processing:**

The absolute coordinates of the image (x,y) were obtained mainly through the relevant api interface of selenium, and then the relative position of this image was obtained by being at the resolution of the screen. The crawl also encountered some small problems: in the crawl process, we found that occasionally the coordinates of (0,0), and this is certainly impossible, so return to check these pictures, found that these pictures are not directly on the surface of the browser, may give a click on what, this picture will appear in the corresponding position, and therefore lead to the position of (0,0), but these Most of these images are general images, so we choose to skip them directly when processing them.

**Ad detection processing:**

In this section we use a combination of **text extraction tools** and **model detection tools**. Text extraction uses Cnocr to match the text information in the advert with the common information in the advert. For the model, we used the dataset provided by the teaching assistant to train a CNN-based dataset, in which we had encountered data augmentation due to the very different amount of data for each type of advertisement image, and data enhancement for some data processing such as copying + flipping to balance the amount of each image.

In practice, we used **a combination of these two types of tools**, with the text judging whether it is an advertisement, and then the model assisting in determining exactly which kind when the text considers it to have multiple features such as pornography and gambling at the same time

1. **Experiment Results**

**(1)Data Analysis**

①Use the csv module to read tables and construct separate tables for English and Chinese website information

②Combine the two tables and perform data cleaning:

Remove the incorrect data collected (relative size >1,<0)

Calculate the relative size of the images (rel\_x,rel\_y)

③Tables for extracting different types of images, depending on the type of image:

-1 :Non-advertising

1 :Gambling advertisements

2 :Regular advertisements

3 :Pornographic advertisements

④Separate scatter plots of their relative position distribution in the normalised scenario

⑤Histogram of the distribution of the four types of images on the x-axis

Histogram of the distribution of the four types of images on the y-axis

Histogram of the width distribution of the four types of images

Histogram of the distribution of the lengths of the four types of images

⑥The average position of each type of image, expressed as a rectangle, in relation to the normalised web page:

* Blue:non-advertising
* yellow:regular advertisements
* Purple:gambling advertisements
* Cyan:pornographic advertisements

⑦Extract the position and size information from each type of image and check the covariance

**（2）Conclusion**

* non-advertising:Positioned centrally on the lower left
* regular advertisements:Positioned mainly on the left, larger in size
* gambling advertisements:positioned mainly above, large size, more spread out in x-direction
* pornographic advertisements:Positioned mainly in the lower right

The experimental data and results is more in line with the the most basic illustration of how we tend to consume the content on web pages

* **key takeaways**
* **Left Is Right**

This sentence explains why the left side of the page gets so much more attention than the corresponding placements on the right. You read it left to right, which is how our brains are conditioned to consume content thanks to the nature of the English language. Our gaze starts left and gradually shifts right, but our focus wanes as we move across.So here’s the key takeaway: the left side of your site will get much more attention than the right.

* **Height Disadvantage**

In reality, the top of the page is often overlooked. This is a great example of banner blindness; our Internet-frequenting brains have been conditioned to expect an ad at the very top of the page and the content we’re looking for down below. So we don’t dwell long at the top of the page, and instead scroll down almost immediately to get to the good stuff. Take note next time you load a new website; you’ll find yourself scrolling almost as soon as you hit “enter.”

* **Hot Spots**

more sites are starting to feature ads smack dab in the middle of their content, ensuring that it will get increased visibility and clickability. Ads positioned within “hot spots” that traditionally include only content can perform very well because they are more likely to be viewed (and therefore more likely to be clicked) than ads that live in sections of the site where the eye rarely travels.

* **Discussion and Limitations**

**What are the limitations of this work？**

It is difficult to collect URLs with specific pornographic game advertisements, most of the pages are still normal pages.it may be possible to make a pre-judgement in the future and decide whether to keep this page as one of the data sources