Matching Contextual Ads and Web Page Contents through Computational Advertising: Getting the Best Match

by

Nur Izyan Izzati binti Abd Halim

Dissertation submitted in partial fulfillment of

the requirements for the

Bachelor of Technology (Hons)

(Information & Communication Technology)

MAY 2013

UniversitiTeknologi PETRONAS
Bandar Seri Iskandar,
31750 Tronoh
Perak Darul Ridzuan.
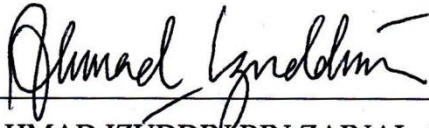
CERTIFICATION OF APPROVAL

Matching Contextual Ads and Web Page Contents through Computational Advertising: Getting the Best Match

by

Nur Izyan Izzati binti Abd Halim

A project dissertation submitted to the
Information Technology Programme
Universiti Teknologi PETRONAS
in partial fulfillment of the requirement for the
BACHELOR OF TECHNOLOGY (Hons)
(INFORMATION & COMMUNICATION TECHNOLOGY)

Approved by,

_____
(MR. AHMAD IZUDDIN BIN ZAINAL ABIDIN)

UNIVERSITI TEKNOLOGI PETRONAS
TRONOH, PERAK
May 2013

ii

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

NUR IZYAN IZZATI BINTI ABD HALIM

ABSTRACT

The technological transformation and automation of digital content delivery has revolutionized the media industry. What is more, the Internet is rapidly turning into an advertising channel. Just in the United States, Internet advertising revenues hit $7.3 billion for the first quarter of 2011, representing a 23 percent increase over the same period in 2010 (iab.net, 2011). Beneficiaries of this investment and growth are search engines such as Google, Yahoo, and MSN. Also, Malaysian advertising landscape is gradually shifting its traditional media forms to the emergent of Internet advertising but still at a budding stage. The latter shows much room for growth, as the industry fuels to content digitization on Web applications.

In this project, the types of Internet advertising that is going to be discussed on are Contextual Ads and Sponsored Search Ads, but the major scope will be on Contextual Advertising. Given that, these types of advertising have the central challenge of finding the "best match" between a given context and a suitable advertisement, through principled way of computational methods. Hence, it is also referred as Computational advertising. Furthermore, there are four main players that exists in the Internet advertising ecosystem that are going to be discussed in this study, which are; Users, Advertisers, Ad Exchange and Publishers.

Hence in order to find ways to counter the centre challenge, this research study will mainly address two objectives, which are to successfully make the best Contextual Ads selections that match to the Web Page contents through the concept of Computational advertising, and to ensure that there is a valuable connection between the Web pages and the Contextual Ads.

Thus, the scope of the study will be mainly on discussing about the theory of Computational advertising itself, besides elaborating on Contextual Ads, matching Contextual Ads and Web pages and also, finding the most feasible way in creating the valuable connection between Contextual Ads and the Web pages. Moreover, at the end of every discussion in every subtopic, some insights on the Internet advertising in Malaysian context are discussed as per related issue.

Consequently, this study employed two main methods to address the research questions rose. Those methods include extensive research and analysis on previous literature works and journals, and also in depth surveys to collect related data and information in real-life situations. Every part of gathered data and findings will then be analyzed accordingly. All discussions, conclusion and future recommendations are presented as per sections. Hence in order to prove the working mechanism of matching Contextual Ads and Web pages by using Computational advertising approach, Web pages together with the ads matching system, will then be developed through FYP-II timeline, as the final product of the study.

Keywords: Advertising, Online advertising, Internet advertising, Computational advertising, Contextual Advertising,

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND NOMENCLATURES

ADX	Ad Exchanges
CPM	Cost-Per Mile
CPC	Cost-Per Click
CPA	Cost-Per Acquisition
CTR	Click-Through Rate
DX	Data Exchanges
DSP	Demand Side Platform
GSP	Generalized Second-Price
IAB	Interactive Advertising Bureau
MCMC	Malaysian Communications and Multimedia Commission
OTC	Over-The Counter
RAD	Rapid Application Development
ROI	Return of Investment
SSP	Supply Side Platform
HCI	Human-Computer Interaction

CHAPTER 1

INTRODUCTION

This chapter will describe the overview of the project that covers the following topics.

- Background of Study
- Problem Statements
- Purpose of Study
- Objectives of Study
- Research Question
- Significance and Feasibility of Study

1.1 Background of Study

1.1.1 Internet advertising

Advertising is a form of communication method for marketing used by businesses, companies and other business organizations to promote their products and services to their audience. According to Burnett (n.d.), "The sole purpose of business is service. The sole purpose of advertising is explaining the service which business renders". This suggests that advertising is a form of communication method. The ultimate aim of advertising is to attract potential customers, by presenting these goods and services in a positive light, so that business entity may increase their sales and revenues. Furthermore, advertising is designed to make an impression on its audience in a way to establish a brand image and create significant products visibility.

Practically, any medium can be used for advertising. Conventionally, advertisers take leverage on radio, cable television, newspapers, prints, magazines and billboards to promote their products and services. These mediums are among the popular and major channels that place ads. However, the since Internet and the World Wide Web (WWW) have gone aboard on a bright advancement, hence this phenomenon facilitates users to seek information online. The Internet, the largest global network of networked computers, was founded in the spirit of free access information. Therefore users are able to put across their information requests on the net, browse specific Websites and carry out e-commerce and online transactions.

Consequently, this has inspires major search engines to continue improving their retrieval services and user's browsing experience by providing relevant results. Thus, as a result, the Internet has emerged as an important medium for advertising, chosen by advertisers to broaden their approaches in reaching their potential customers among Web users. In simple words, advertisers turn unexplored territory into an opportunity with the vision that a portion of those who are online and on the lookout for specific products or services may click the ads. Precisely, Internet advertising is the union of traditional advertising and direct response marketing. A significant part of this market consists of ads delivery distributed through three main advertising channels: sponsored search, Contextual Advertising and display ads.

It appears that currently Internet advertising is complementing the traditional methods of advertising, but with the rapid progression of technology, it is believed that soon, it will change the current trends and starts to dominate the existing media as the preferred medium for placing ads. This is mainly because Internet advertising holds four distinct advantages over traditional advertising, which are; more specific and target based, the ability of measuring response by tracking system, high deliverability and flexibility, and valuable interactivity between advertisers and consumers.

Figure 1 shows the U.S. Advertising Revenue Among The Largest Media Companies (2006 -2011). Therefore it is proved that Internet advertising (online advertising) is continuing to take the dominant place in the advertising industry, only four per cent left behind the television medium in 2011.
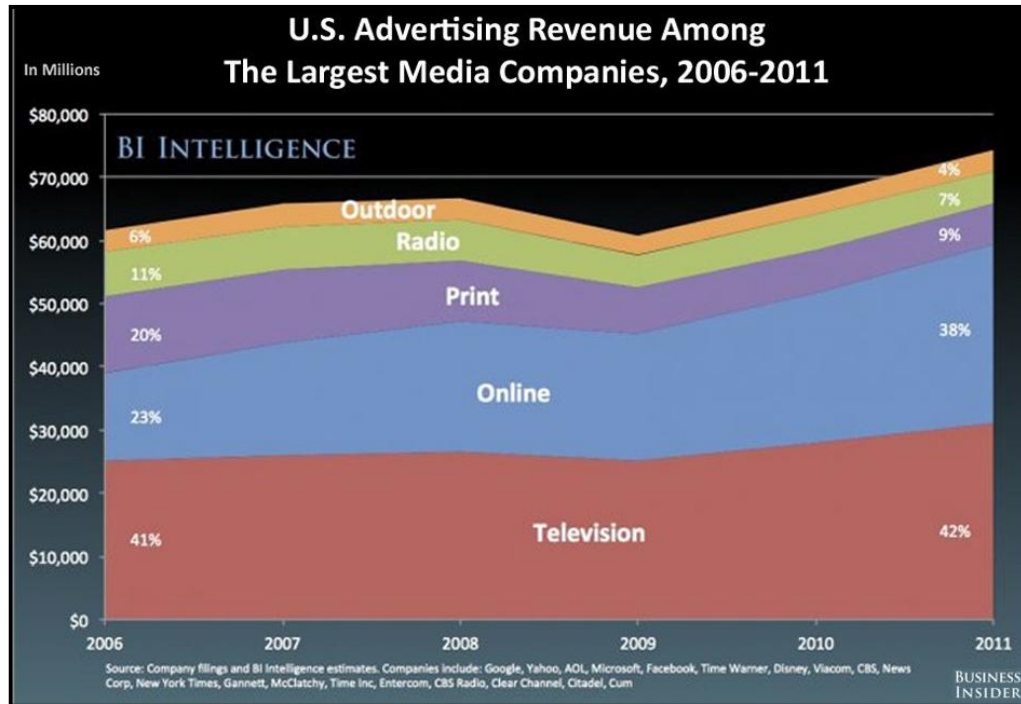
Figure 1: U.S. Advertising Revenue Among The Largest Media Companies (2006 -2011)

As mentioned by Hua, Mei, and Li (2008), Internet advertising has embarked on a dramatic evolution, which will be rapid, fundamental, and permanent. According to Interactive Advertising Bureau's (IAB) recent report, in the first half of 2012, Internet advertising revenues climbed to an all-time high of $17 billion, representing a 14 percent increase year-over-year. This represents a 14% (or $2.1 billion) increase from 2011's $14.9 billion. Hence, this event automatically placed huge opportunities for the search engine to be the main and strategic platform for advertisers to put their ads on the Web.

More to the point, it is forecasted that Global online ad spending will overtake newspaper expenditures next year (USD $101.5 billion vs. $93.2 billion), and will exceed the combined total of newspapers and magazines in 2015 ($132.4 billion vs. $131.7 billion), predicts ZenithOptimedia, (2012) in a new forecast. That equates to almost 1 in 4 ad dollars being spent globally on the Internet in 2015, second only to TV (40%). Breaking down the forecast by media, Internet advertising will jump from $76.9 billion last year to $132.4 billion in 2015, growing its share of the total from 16.1% to 23.4%, Figure 2 shows.

Share of Global Ad Spend, by Medium
(% share)
2011-2015

	2011	2012	2013	2014	2015
Newspapers	20.3%	18.9%	17.8%	16.8%	15.9%
Magazines	9.4%	8.8%	8.3%	7.8%	7.3%
TV	39.9%	40.2%	40.1%	40.1%	40.0%
Radio	7.1%	7.0%	6.9%	6.7%	6.6%
Cinema	0.5%	0.6%	0.5%	0.6%	0.6%
Outdoor	6.7%	6.6%	6.5%	6.4%	6.3%
Internet	16.1%	18.0%	19.8%	21.6%	23.4%

MARKETINGCHARTS.COM
Source: ZenithOptimedia

Figure 2: Share of Global Ad Spend, by Medium (% share) 2011-2015

Narrowing down the scope to the Malaysian market itself, Figure 3 shows the ad spending (2011 – 2012) growth on the Internet advertising in Malaysia. According to Kumar (2012), Internet spending in Malaysia would continue to grow strongly for the rest of the year, and as per estimates by IPG Mediabrands (Asia world market) it has been growing at similar rates for the last two years. Hence, this statement affirmed that the Malaysian advertising landscape is gradually shifting its traditional media forms to the emergent new media advertising (which is Internet advertising) but still at a nascent stage. The latter shows much room for growth, as the industry fuels to content digitization on Web applications, according to Malaysian Communications and Multimedia Commission (MCMC).
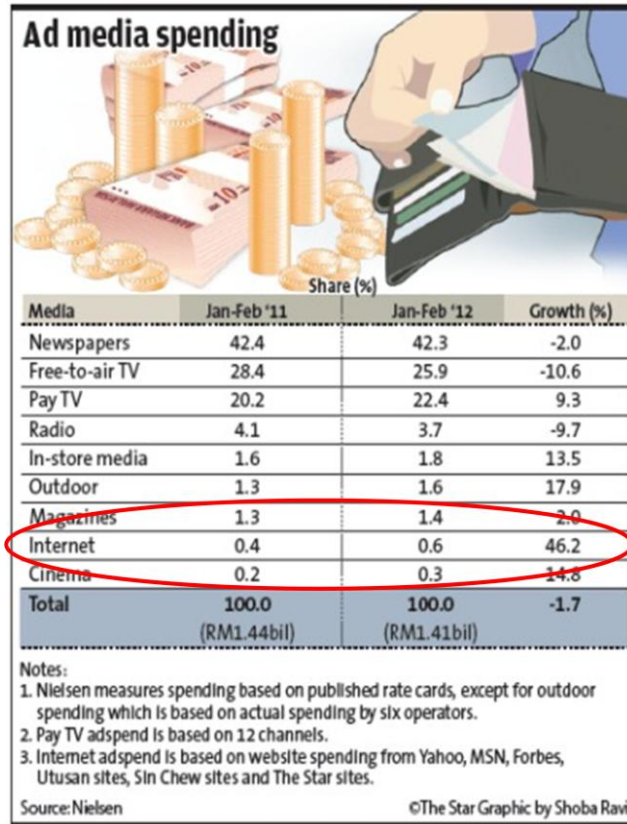
Ad media spending

Media	Share (%) Jan-Feb '11	Jan-Feb '12	Growth (%)
Newspapers	42.4	42.3	-2.0
Free-to-air TV	28.4	25.9	-10.6
Pay TV	20.2	22.4	9.3
Radio	4.1	3.7	-9.7
In-store media	1.6	1.8	13.5
Outdoor	1.3	1.6	17.9
Magazines	1.3	1.4	2.0
Internet	0.4	0.6	46.2
Cinema	0.2	0.3	14.8
Total	**100.0** (RM1.44bil)	**100.0** (RM1.41bil)	**-1.7**

Notes:
1. Nielsen measures spending based on published rate cards, except for outdoor spending which is based on actual spending by six operators.
2. Pay TV adspend is based on 12 channels.
3. Internet adspend is based on website spending from Yahoo, MSN, Forbes, Utusan sites, Sin Chew sites and The Star sites.

Source: Nielsen   ©The Star Graphic by Shoba Ravi

Figure 3: Malaysia ad media spending (2011 – 2012)

1.1.2 Historical View on Internet advertising

Prior to 1998, online advertising consisted of banner advertisements, generally priced by the number of impressions delivered, for example; Cost-Per Mile (CPM) pricing. CPM is used to calculate the relative cost of an advertising campaign or an ad message in a given medium, in which advertisers are charged per thousand impressions. The first sponsored search auction was created by GoTo.com (renamed Overture in 2001, and finally acquired by Yahoo! in 2002), and Google created its first sponsored search auction in 2002.

Between them, Yahoo! and Google have advanced the sponsored search auction format since, as shown in Figure 2. In addition, new platform, or several other players in the sponsored search market start to become known including Ask, LookSmart, and Microsoft. Sponsored search allows advertisers to buy certain keywords to promote and advertise their business when users use such a search engine, and contributes greatly to its free service.
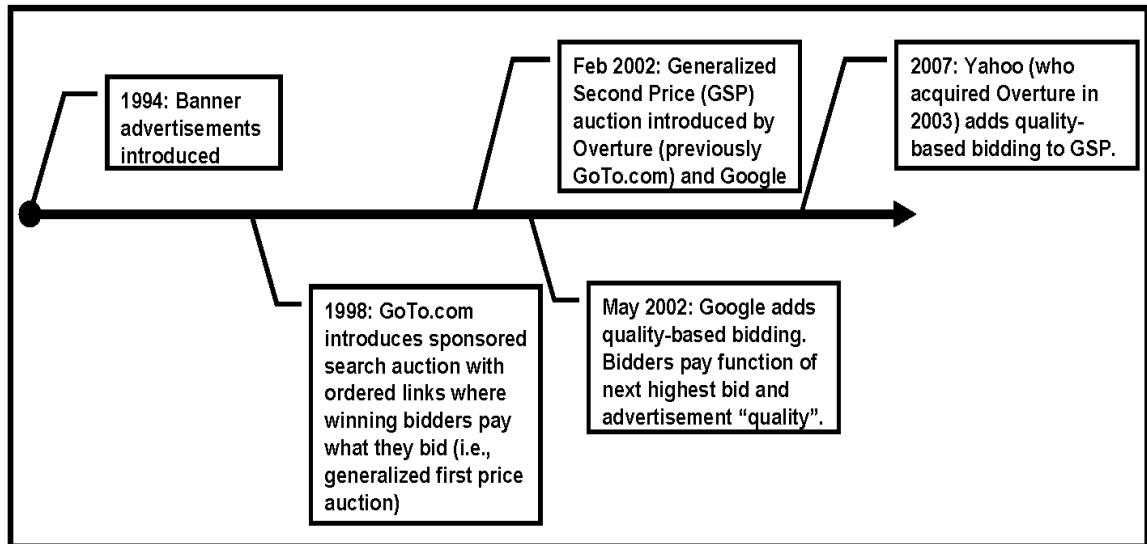

Figure 4: Timeline for sponsored search auctions

The trends evolve, where Contextual Advertising history began in 1988. Oingo, a private held company located in Los Angeles, developed a proprietary search algorithm that was based on word meanings and built upon an underlying lexicon called Wordnet. This company, started by Gilad Elbas and Adam Weissman, which was later renamed to Applied Semantics (happened in 2001). Google bought out Applied Semantics in April 2003 for $102 million and renamed it as AdSense. Besides, MSN AdCenter and Yahoo! Publisher Network are alternative Contextual Advertising companies that provide these types of services. These sites give small business owners the opportunity to advertise without having to deal with each individual sponsor. Often, large sponsors will not advertise on small sites on an individual level.

14

As the trends grow, the Contextual Advertising platforms adapt to a more affluent media environment, such as video, audio and mobile networks with geographical information. Blocks of space on publishers' Web pages, video clips and applications are being sold to make money, through these platforms. Generally such services are called an advertising network or a display network. They are not necessarily run by search engines and the network can contain a huge number of individual publishers and advertisers. For instance in Malaysian contexts, some examples of ad network are Adsense, Nuffnang, Innity, Admax, SkyAd and Advertlets. These ad networks are originated from Malaysia and their businesses are currently up and running smoothly.

Subsequently, in order to lubricate the rolling wheels of the existing online advertising industry, new platforms like ADSDAQ, AdECN, DoubleClick Advertising Exchange, adBrite amd Right Media Exchange, are being introduced. These platforms are called ad exchanges and they are focusing on real-time buying and selling impressions were created. When ad exchanges opened, they brought more liquidity to the marketplace for online inventory. 2007 was a pivotal year for ad exchanges.

Three major exchanges were acquired that year: Yahoo! bought Right Media in April, Google bought DoubleClick in May and Microsoft bought AdECN in August. Each company quickly made vast pools of inventory available, which greatly improved the experience for many parties to transact in online display (Google, 2011). In the mean time, some comparable platforms come into the scene, known as demand side platform (DSP) and supply side platform (SSP), whereby DSP provides a more proficient way for advertisers to purchase ads from many different ad exchanges, and SSP allows Web publishers to automate and optimize the selling of their online media space.

1.1.3 Characteristics and Players of Internet advertising

The cast of characters in the Internet advertising industry can be divided into four camps: advertisers, publishers, ad networks and users. Figure 3 illustrates the simplified ecosystem of Internet advertising.
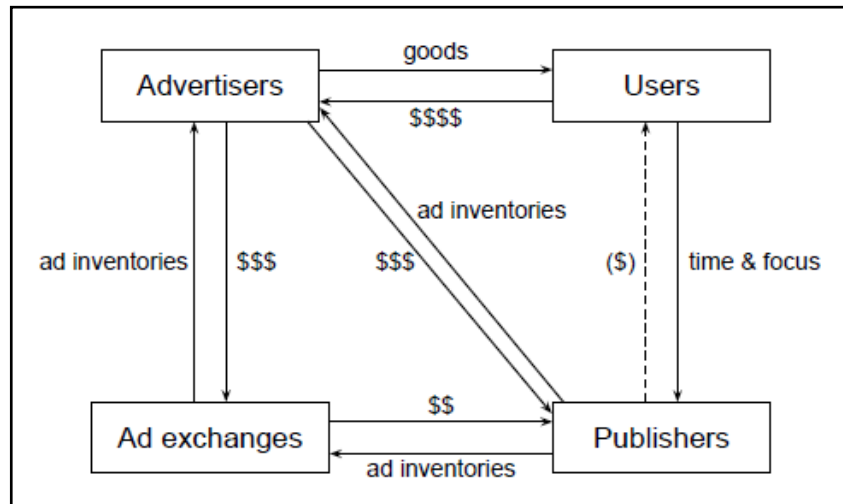

Figure 5: The simplified ecosystem of Internet advertising. Advertisers spend budget to buy ad inventories from ad exchanges and publishers; ad exchanges serve as matchers for ads and inventories; publishers provide valuable information to satisfy and keep visitors; users read ads and purchase goods from the advertisers. Note that normally users would not receive cash from publishers.

If we take an analogy from the economic perspective, ad inventories are traded based on the force of demand and supply. In the case of contextual advertising, a content publisher reserves some space for ads.

- *Publisher* is the owner of the Web pages on which the advertising is displayed. The publisher typically aims to maximize advertising revenue while providing a good user experience.
- *Advertiser* provides the supply of ads. Usually the activities of the advertisers are organized around campaigns which are defined by a set of ads with a particular temporal and thematic goal (e.g. sale of digital cameras during the holiday season). As in traditional advertising, the goal of the advertisers can be broadly defined as the promotion of products or services.

16

- *Ad network* is a mediator between the advertiser and the publisher and selects the ads that are put on the pages. The ad-network shares the advertisement revenue with the publisher.
- *Users* visit the Web pages of the publisher and interact with the ads.

As the industry evolve recent years, new platforms are being introduced in order to get the advertising business running smoothly and more competent, besides giving it unique and valuable criteria. The new platforms include Demand side platforms (DSP), Supply side platforms (SSP), Ad Exchanges (ADX) and Data Exchanges (DX). As shown in Figure 4 below:
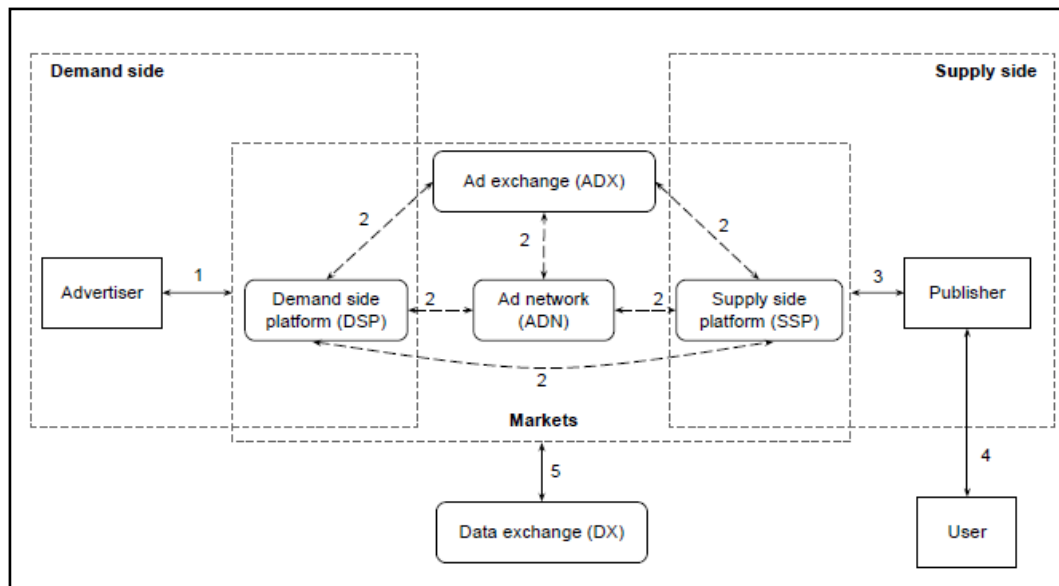

Figure 6: The various players of Internet advertising and the trading process: 1) The advertiser creates campaigns in markets 2) The markets can trade campaigns and impressions to balance the demand and supply for better efficiency 3) The publisher registers impressions with the markets 4) The user issues queries or visits Web pages 5) The markets can query data exchanges for user profiles in real-time bidding.

- *Demand side platforms (DSP)* serve advertisers or ad agencies by bidding for their campaigns in multiple ad networks automatically;
- *Supply side platforms (SSP)* serve publishers by registering their inventories (impressions) in multiple ad networks and accepting the most beneficial ads automatically;

17

- *Ad exchanges (ADX)* combine multiple ad networks together (Muthukrishnan, 2009). When publishers request ads with a given context to serve users, the ADX contacts candidate Ad Networks (ADN) in real-time for a wider selection of relevant ads;

- *Data exchanges (DX)* sometimes called Data Management Platforms (DMP), serves DSP, SSP and ADX by providing user historical data (usually in real-time) for better matching.

1.1.4   Computational advertising: The Brief Concept

Internet advertising is a kind of operational advertising released on the Internet with carriers of digital codes and algorithms. In this project, the types of Internet advertising that the author going to focus on are sponsored search and Contextual Ads. Given that, these types of advertising have the central challenge of finding the "best match" between a given user in a given context and a suitable advertisement, through principled way of computational methods.

Hence, it is also referred as Computational advertising. Computational advertising is an emerging scientific discipline, at the intersection of large scale search and text analysis, information retrieval, statistical modeling, machine learning, optimization, and microeconomics. It is said that, Computational advertising will outshine classic advertising to become the fifth major media, following the traditional television, radio, newspaper, and magazine.

According to Kushal Dave (2011), computational advertising, popularly known as online advertising or Web advertising, refers to finding the most relevant ads matching a particular context on the Web. It is a scientific sub-discipline at the intersection of information retrieval, statistical modeling, machine learning, optimization, large scale search and text analysis. The core problem attacked in computational advertising (CA) is of the match making between the ads and the context.

Therefore, the unique features and characteristics of computational advertising compared to the traditional method of advertising include:

- *Cost-effective*, whereby advertisers may choose the best billing methods for their online advertisements which the most popular are; cost-per-click (CPC), cost-per-mille (CPM) and cost-per-acquisition (CPA). Furthermore, advertisers should consider two factors before they can actually publish their ads, which are: their main purpose(s) for advertising, and the size of their advertising budget. Only from that, advertisers will be able to publish their ads, make trade-offs via online and achieve their advertising purposes, besides maintaining their budget within boundary. But even if the billing methods are chose, the cost is still variable due to the competition in auctions. The auctions are done every time the ads need to be displayed, differ from traditional advertising media, where the cost is usually negotiated and fixed before the ads are placed.

- *Real-time*, whereby an ad is delivered in real-time 24 hours per day, 7 days per week and 365 days per year. Furthermore, an ad campaign can be launched, updated or cancelled immediately online. This is a big difference from print, where an ad cannot be changed until the new edition of the publication is published, or; television, where the high costs of ad development make frequent changed prohibitive.

- *Targeted*, whereby online advertisers have an entirely new range of targeting capabilities. They can focus on users from specific companies, or geographical regions or nations, as well as by time of day, computer platform, and browser. They can target using the databases that serve as the backbone of direct marketing. Furthermore they can even target based on a person's personal preferences and actual behavior. Compared to the classic advertising methods they are not specifying any targeted audience, where ads are displayed publicly, for everyone.

- *Measureable*, whereby advertisers can measure the response to an advertisement (through the number of times and ad is clicked on and the number of purchases or leads and ad generated), which is difficult to do with traditional television, print and billboard advertising. From this, marketers can track how users interact with their brands and learn what the current interests of the users are. This encourages advertisers to improve their campaigns in all aspects.

1.1.4 Contextual Advertising

As per mentioned in the earlier section, there are two types of Internet advertising, which are Sponsored Search and Contextual Advertising. In this project dissertation, the author focused on contextual advertising that refers to the placement of ads within the content of a generic Web page for a user. Contextual advertising or Context Match (CM) refers to the placement of commercial textual advertisements within the content of a generic Web page (Andrei et al. 2007). Besides, contextual advertising is a concept different than normal paid inclusion. Where methods such as pay per click and other links appear on search engine results pages, contextual advertising campaigns place banners and links on Web pages that have a contextual relationship to the product or service being sold. For example, a banner ad for baking pans might be placed on a Web page that has a recipe for a cake on it.

Furthermore, contextual advertising does not require a keyword search to bring up the ad; it is simply there every time a particular page is viewed. Contextual advertising is targeted advertising that typically occurs on a banner or pop-up ad on a Website. Contextual ad systems target advertising to a specific user based on the keywords on the page he or she is visiting (hence, the context of the ad comes into play). Contextual advertising works when Webmasters order Java code from companies such as Google AdWords that will provide them with advertisements that pertain to the content of their Web pages. Once this code is entered into the HTML of the Web page, each time the page is viewed, an advertisement, usually a banner ad, will appear on the page.

Contextual advertising systems scan the Web pages that it has been written into for certain keywords and phrases that have a direct connection to the product or service being sold, and then place the appropriate ad on the page. The banner ad that appears on the Web page can be different each time a person visits the page, but many advertisers keep it to one or two per page, so that interested parties can always come back to the ad if they do not feel like reading it the first time it appears. Advertising with contextual ads are one of Google's major sources of revenue. As the advertisement is contextual, and actually has something to do with the topic of the Web page, people are more likely to click on the ad and visit another Website, or even become a conversion.

More to the point, developers are seeking to improve the relevance of targeted advertisements displayed to the viewer. Currently, the relevance of an advertisement is determined by technology such as Googlebot and Mediabot. This technology works by inserting JavaScript into the Webpage in order to display relevant advertisements to the user. The algorithms are currently being improved to more accurately match advertisements to the preferences of the consumers. In contextual advertising, there is usually an intermediary commercial ad-network entity in charge of optimizing the ad selection with the twin goal of increasing revenue (shared between the publisher and the ad-network) and improving the user experience. With these goals in mind it is preferable to have ads relevant to the page content, rather than generic ads.

There are three different types of contextual ads: separate ads that appear in specific areas on a page, inline or in-text contextual ads, or pop-up ads. Some recognizable contextual advertisers are Google's Adsense, Overture's Content Match and AdBrite. Inline contextual advertisers who are willing to pay top dollar for Pay Per Click (PPC) advertising can have their ad placed at the top of the "sponsored listings" section of a search results page. In-text contextual advertisers, advertisements appear in the actual body text of a page and allow users to view the ad if they choose to click on the link.

Figure 7: Various contextual advertisers in today's Internet advertising network.

Small to medium business owners can save expenses by using a contextual advertising program because the targets are so specific. Certain sites, like news and publisher sites benefit from contextual advertising because the content on these sites is so specific that the returned ads will be targeted to the search engine user. Directories are also great places to display contextual ads. Because they are so targeted, the ads that appear in each category will be relevant. Publishers and advertisers are encouraged to work together to maximize the amount of money that they both earn through contextual advertising. If they work together, both parties will earn a profit. Advertisers enjoy the ability to set their own prices for their product.

However, companies are trying to determine if a price can be set for all members of the network. Google introduced the Smart Pricing scheme that will allow prices to be dynamically set. Yahoo! has opted to set one price for each company rather than using software to determine a price for the product. The problem with this option is that most companies are compelled to set the price to the least valuable site in the network.

Inline contextual ads appear under the "sponsored links" area. Research shows that search engine users are four times more likely to click on a site displayed on the free side of the first page of search engine through organic search engine optimization, by a Website design company, than they are to click on the sponsored sites1.Because of the irritability of pop-up ads, many search engine users use the pop-up ad blocker meaning that they will never even view the ad. Also, contextual advertising has caused controversy by sometimes using third-party hyperlinking, the process of installing software on a search engine user's computer to return links to the user that are not paying the Website to advertise on its site.

1.1 Problem Statements

Based on Yahoo! Research Labs (2012), since Computational advertising is a new discipline in Internet advertising, therefore it has many challenges that still need to be considered. This is because in this new age, new application domains and new techniques are emerging every day, demanding extensive researches and explorations in successfully make the best ad selection that will optimizes a goal that balances the utilities of the four main players in Internet advertising.

More to the point, the traditional form of Internet advertising often failed to address the effective way of advertising ads on the Internet. The problem always rises from the management of the advertisements itself. Usually it could not specify the market segment, advertisers always run out of inventories and some other ads management problems that will eventually decrease users' experience when they clicked on the ads.

Hence, this study will mainly focus on these three challenges:

i. Modelling Internet ads that likely match user's interest, whereby the challenge comes in when advertisers need to know user's preferences at the first place, in order to display the most relevant and appropriate ads that match to user's specific interest.

ii. Establishing Internet ads that are more likely to be clicked on, where publishers need to find a technique in establishing the suitable ads closer to the content served on the Web, so that the ads match user's preferences and eventually will increase the possibility of the ads being clicked by the users when they are online.

iii. Minimizing advertising costs for advertisers with low budget in marketing their products, where advertisers need to successfully display their ads that match to user's preferences, but at the same time, generating as much revenue as possible.

1.2 Purpose of Study

The purpose of this study is to explore the theory of Computational advertising in order to successfully make the best Contextual Ads selections that match to the Web Page contents that will consequently addresses the problem statements.

1.3 Objectives and Scope of Study

The objectives of this study are:

- To successfully make the best Contextual Ads selections that match to the Web Page contents through the concept of Computational advertising.
- To ensure there is a valuable connection between the Web pages and the Contextual Ads.

Thus on the other hand, this research study will focus on four **scopes** as per below:

- Theory of Computational advertising.
- Contextual Ads.
- Matching Contextual Ads and Web pages.
- Finding the most feasible way in creating the valuable connection between Contextual Ads and the Web pages.

1.4 Research Questions

- How does the concept of Computational advertising can be used to successfully make the best Contextual Ads selections that match to the Web Page contents?
- How does valuable connection between the Web pages and the Contextual Ads can be created?

1.5 Significance and Feasibility of Study

Conducive research is conducted on the Computational advertising theory and the theory will later be proved on a Website that demonstrates the selection process of Contextual Ads that match to the content of the Web pages. This study aims to explore the theory of Computational advertising in order for the advertisers to find the "best match" between the given contexts (Web Page contents) with the suitable advertisement (Contextual Ads), as well as to persistently generate as much revenue as possible.

Moreover, it is believed that Internet advertising is likely to propel in the emerging advertising industry in Malaysia as the financial scale is huge. More to add, computational advertising offers a lot of new benefits and advantages that publishers and advertisers can gain compared to the traditional method of Internet advertising. The optimal solution requires use of external knowledge and continuous research and development in this particular area which also explains the significance of this research study.

The time frame of the project development comprises two semesters of study; whereby the first semester (during FYP-I: September 2012 – January 2013), the author focused on the extensive research, planning and analysis phases, while the second semester (during FYP-II: January 2013 – May 2013), the author mainly focused on designing, modeling, prototyping and developing the ads matching system. Also, the time frame given is feasible and practical with the scopes of the study, as being outlined in the previous subtopic.

CHAPTER 2
LITERATURE REVIEW

This chapter discusses on the general idea of the project. Several keywords are discussed to address the statement of problem in the previous chapter. Furthermore, this chapter also expresses certain theories and views on the nature of the Computational advertising and the detailed perspectives of each player in Internet advertising, besides current views and methods of matching Contextual ads with Web page contents, in addressing problem statements and objectives.

2.1 Recent Trends in Advertising

Advertising is a form of communication used to encourage or persuade an audience (viewers, readers or listeners) to continue or take some action. Traditionally, advertising has been defined as a form of controlled communication that attempts to persuade consumers, using strategies and appeals, to buy or use a particular product or service (Defleur & Dennis, 1996). Most commonly, the desired result is to drive consumer behavior with respect to a commercial offering, although political and ideological advertising is also common. The purpose of advertising may also be reassured employees or shareholders that a company is viable or successful. Advertising messages are usually paid for by sponsors and views via various traditional media; including mass media such as newspaper, magazines, prints, television commercial, radio advertisement, outdoor advertising or direct mail; or new media such as Websites and text messages.

Commercial advertisers often seek to generate increased consumption of their products or services through "Branding", which involves the repetition of an image or product name in an effort to associate certain qualities with the brand in the minds of consumers. Non-commercial advertisers who spend money to advertise items other than a consumer product or services include political parties, interest groups, religious organizations and governmental agencies. The advertisement will speak about the salient features of the product on offer and the benefit the user can derive out of the product. It can also educate the target audience about the various other details such as the products cost, availability, usage modalities, problems that may arise while using it and the probable solutions to those problems.

The advertising industry has a long history of creativity and innovation. Nowadays, Internet and mobile phone technology is in the process of transforming the way societies not only communicate with each other, but introducing many new platforms of Internet advertising. According to Zinkhan and Watson, (1996), new ideas and technologies spread rapidly in Western- style democracies, in part due to the advanced nature of the communication industry, including advertising. In the 1990's, consumers perceived online advertising with banner, pop-up, and e-mail advertisements.

Today, online advertising is emerging as a compelling marketing component with rapid growth in the United States, Western Europe, and parts of Asia. It is becoming abundantly clear that the media environment into which advertising is placed is changing, and because of this trend, the nature of advertising is changing as well. Many new channels of mass communication were developed during the latter part of the 20th century that exposed the public to an ever-increasing number of mediated messages (Fitzgerald, 1999; Lombard et al., 1997).

Online advertising, once being associated with banner, pop-up, and e-mail advertising, has now evolved into a versatile advertising channel with massive growth prospects and bigger opportunities. Today, online advertising includes not only evolved banner, pop-up, and e-mail ads but also search marketing, sponsored search, pay-per-click, pay-per-action, rich media, Contextual Advertising, geo-targeting, behavioral targeting, social marketing, video advertising, and user-generated online video. Advertising also is appearing in online games, in-line text, social media, blogs, and mobile formats (Boone, Secci, and Gallant, n.d.).

In Malaysia itself, there have been as many as seven new online advertising network companies entering the local scene. According to Tan (2009), this proves that online advertisers realize the importance of ad networks in their digital media plan mix and are allocating more budget to ad networks as they provide higher audience reach in the relevant environment and give better ROI (return on investment) with media or creative optimization technology. Hence, digital media in Malaysia are on the edge for "explosive growth" and are finally gaining credibility as mainstream – as opposed to niche (classic advertising methods). Furthermore, according to MCMC, the digital world is changing not only consumers' habits in Malaysia, but also encouraging advertising opportunities, especially online advertising. Figure 8 shows the share of advertising expenditure (as per February 2011), in Malaysian context.
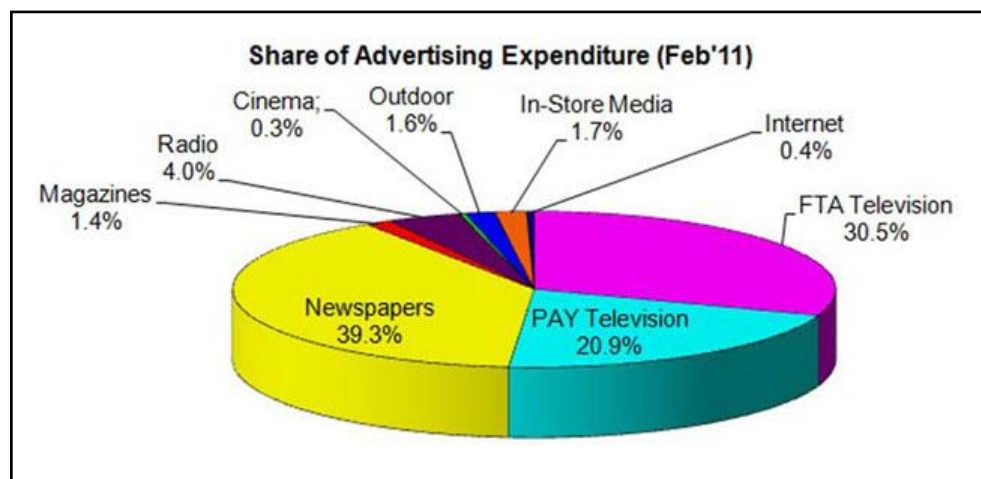
Figure 8: S*hare of advertising expenditure (as per February 2011), in Malaysian context*

Therefore nowadays, advertising industry has evolved from classic/traditional advertising channels (radio, television, print, billboards, outdoor advertisements), to Internet advertising due to the technology advancements, and also due to the consumers' habit. Moreover, the developments of computer mediated communication technology, brings up the opportunity for marketer to get direct feedback from consumers and Website visitors, along with the ability to customize individual messages (Haque, et al. 2007). It is believed that soon in the future, Internet advertising will change the current trends and starts to dominate the existing media as the preferred medium for placing ads.

2.2 Computational Advertising Theory

In this study, since the types of advertising that are being focused on (sponsored search and Contextual Ads) require computation and principled way of finding the best match between a user in a given context and available ads, they are also referred to as Computational advertising. Computational advertising is an emerging scientific discipline, at the intersection of large scale search and text analysis, information retrieval, statistical modeling, machine learning, optimization, and microeconomics. (Broder, 2008).

The central challenge of Computational advertising is to find the "best match" between a given user in a given context and a suitable advertisement. Thus, depending on the definition of "best match" this challenge leads to a variety of massive optimization and search problems, with complicated constraints. Moreover it is aimed that Computational advertising will optimizes a goal that balances the utilities of the four main players in Internet advertising; users, advertisers, publishers and ad exchange.

According to Dave (2011), Computational advertising, a term recently coined in, is about using various computational methodologies to do contextually targeted advertising. As discussed, Computational advertising mainly addressed: retrieving a set of ads that best matches the context and then ranking these ads.

Based on the context, Computational advertising can be divided into following two areas: 1) Sponsored search 2) Contextual Advertising. Also, Dave (2011) proposes that sponsored search refers to the placement of ads on search results page. Here the context is the query issued by the user and the problem is to retrieve top relevant ads that semantically match the query. 2) Contextual Advertising deals with the placement of ads on third-party Web pages. It is similar to sponsored search, with the ads being matched to the complete Web Page text as opposed to a query.
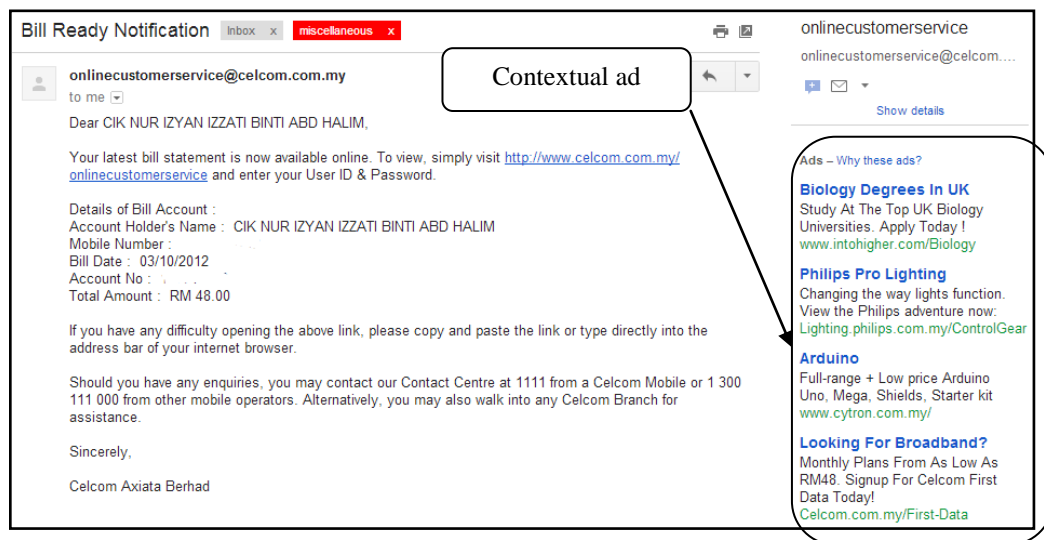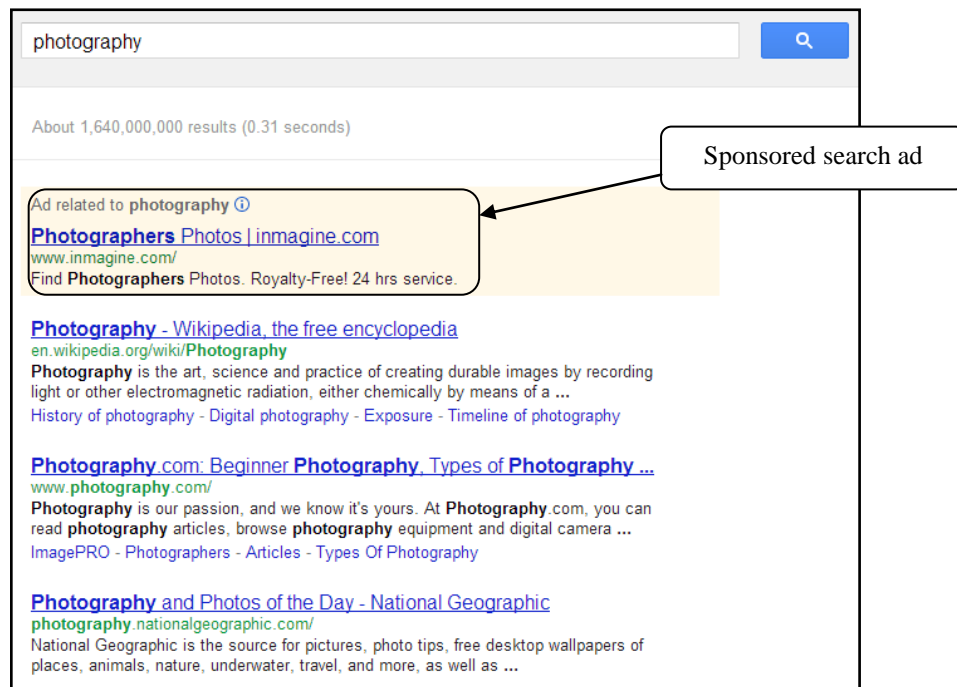
Figure 9: Common types of ads on the Internet. At the top is sponsored search ad relevant to the query issued by the user. And the below is contextual ad.

30

There are two key technological innovations that trigger the sponsored search process, both of which depend on advanced economic and mathematical methods, and which ultimately help determine the nature of the market structure. They include: 1) Search-ad platform using "keyword bidding system" 2) Maximizing revenue from selling slots (David S. Evans, 2008). There are two main reasons why keyword-based ads have come to dominate online advertising - better targeting of customers and better measurability. (Gopal, Li, and Sankaranarayanan, 2011). First, the advertisement brings a product or service to the attention of a potential customer, at the right time, and in aligning users' interests with relevant ads. Second, the business model for keyword-based ads marks a radical departure from previous ad-business models, in that advertisers pay only when a user clicks on an ad (at a cost per click (CPC) that is set by the advertiser), instead of merely when an ad impression is served. Consequently, both Yahoo! And Google run continuous auction by using the keyword bidding system, in which a competitor willing to pay more can displace other advertiser at any time (Fain, and Pedersen, n.d.)

In Contextual Advertising, instead of displaying the same ads to everyone, different ads are shown with regard to the geography, language, device and other characteristics of visitors, to maximize the utilization of advertising opportunities (Yuan, et al. 2012). Contextual Advertising, such as Google AdSense, is a type of text-based search ad which often presents three to five different ads in a frame, each including title, description, and display URL (Yung-Ming Li and Jhih-Hua Jhang-Li, 2008). By completing an application form, logging into their private accounts, and putting a block of HTML into Web pages, publishers can apply to Google AdSense. Then, when users view the Web Page containing the code, the user's browser downloads the Web Page and displays it. More requests and responses will not effuse until all materials are ready to show, since the code indicates that some components (i.e., ad content) must be downloaded from a particular Web server. According to Broder, et al. (n.d.), In Contextual Advertising usually there is a commercial intermediary, called an ad-network, in charge of optimizing the ad selection with the twin goal of increasing revenue (shared between publisher and ad-network) and improving user experience.

2.3 Perspectives of the Internet advertising: Players

As per discussions in the project background (*part 1.1.3*), there are four major players in the Internet advertising ecosystem, which are: 1) users, 2) advertisers, 3) publishers and 4) ad exchange. Each and every player in the ecosystem has their own utilities that need to be considered. Hence, Computational advertising facilitates Internet advertising to successfully make the best ad selection that will optimize a goal that balances the utilities of the four main players in Internet advertising. The next four sub-topics will discuss on the perspectives of each of the player (their utilities, and how do each player creates balance to satisfy each other's utilities, in order to create win-win situation between all the four players.)

2.3.1 The Perspective of a User

User is a person who browses the Web, consumes media content and performs searches using search engines. (Yuan, et al. 2012). It is also can be said that user is a person that visits the Web pages of the publisher and interacts with the ads (Broder, et. al, n.d.). Hence, the utility of a user is, having to be provided with suitable ads that best match their preferences and interest, while he/she is online. While browsing on the Internet, users seem to acknowledge that the presence of ads is what allows them to view much of the content on the Web for free and have accepted Web advertising as a way of life on the Internet (McDonald and Crano, 2009). Nevertheless, Enquiro (2008) verified that users show a reluctance and unwillingness to click on Web advertisements and make an active effort to avoid doing so, viewing advertising as a visual obstruction, impeding the content of the page. To further exemplify this point, Jansen and Resnick (2006) demonstrate that users performing e-commerce searches were more likely to click on a non-sponsored link for a Website than the same link contained in an advert. Besides, Fox et al. (2005) found that implicit measurements, in particular click-throughs, time spent on page and exit type, observed whilst users performed Web searches and read news articles, correlated with explicit user satisfaction measurements and thus could be used as an approval metric to determine if the online advertisement is in fact working.

Apart from that behavioral targeting also can be used to achieve the utility of a user, whereby deducing the intent of a search in a search engine, and displaying ads accordingly (Jansen et al., 2008). This can be done through, by making use of an array of Personally Identifiable Information (PII) about users at the same time as they are accessing the Website, derived from cookies, flash cookies, Web beacons, browser and other meta-data. This PII is used to profile users and so deliver relevant, targeted ads to them, which has been shown to be effective by Yan et al.(2009) and Jaworska and Sydow (2008),who were able to grant empirical evidence of improvement using behavioral targeting. On the other hand, even in the absence of cookies and other tracking files, it is possible to perform user profiling based on browser data alone (Eckersley, 2010). Furthermore, another study by Gord Hotchkiss and Edwards, (2005) shows the usage of eye tracking method to investigate how users navigated the Google search interface and this has revealed much about how sponsored add on the search page were considered by the user. Thus, the result from the studies will help advertisers and publishers to improve their advertising campaign in order to satisfy user's utility and preferences, besides increasing user's reach by targeting the right users.

2.3.2. The Perspective of an Advertiser

Advertiser provides the supply of ads. Usually the activities of the advertisers are organized around campaigns which are defined by a set of ads with a particular temporal and thematic goal (Broder, et al. n.d.). In order to generate demands for its products or service form its customers, advertiser will venture into commercializing ads that will attract more of its customers to buy the products and services offered. Thus, the main utility of an advertiser in Internet advertising ecosystem would be maximizing the click-through-rate (CTR) by viewing each impression or click as an asset with future returns. In most networks, the amount paid by the advertiser for each sponsored search click is determined by an auction process where the advertisers place bids on a search phrase, and their position in the tower of ads displayed in conjunction with the result is determined by their bid (Broder, et al. n.d.).

Usually, advertiser participates in keyword auctions, where keywords are selected before the auction starts. The keywords selection must be done carefully, in order to keep up with the budget, and at the same time, profits are still at the maximum point. As the keyword statistics change from time to time, making keyword selection can be overwhelming. Apart from that, Naldi, et al. (2010), explore the application and implications of the use of the generalized second-price (GSP) auction mechanism for sponsored link slot assignment and pricing in sponsored advertisement keyword auctions. The GSP's auction mechanism allocates advertising slots in the descending order of advertisers' bid prices. Those advertisers who bid the highest prices are given the most attractive sponsored advertising slots at the top of the Web Page, where the click-through rates have been shown to be the highest (Brooks, 2004).

2.3.3 The Perspective of a Publisher

Publisher is the owner of the Web pages on which the advertising is displayed. The publisher typically aims to maximize advertising revenue (Return of Investment (ROI)) while providing a good user experience (Broder, et.al., n.d.). Publishers run Websites, and these Websites have specific ad spots where advertising can run. When an advertiser ask the publisher to run a campaign, the publisher ultimately has to make sure that campaign is set up appropriately and getting the promised traffic, as per negotiated in the contract. In order to make the most money, the publisher will make the best use of its available ad inventory.

Therefore a challenge for publishers is to select the optimal contract or estimate the optimal price (Yuan, et al. 2012). Publisher is on maximizing the revenue from the over –the-counter (OTC) contracts. This is due to the strategy here is more perceptible as when the publisher received requests directly from the advertisers, the publisher will have more control on it. Therefore, the publisher's allocation and inventory management have to be efficient if the publisher want to maximize revenue from the OTC contracts.

Work by Roels and Fridgeirsdottir, (2009) and Feige, et al. (2008) incorporated contract guarantees, where an attempt was made to maximize a publisher's revenue in display ads through dynamic optimization. Apart from that, publisher also may schedule and improve its content quality in order to maximize their revenue and profits return. Nakamura and Abe (2005) developed an LP-based algorithm to schedule banner ads, where they presented three features that each ad was associated with; 1) time of day that the ads were preferred to be viewed (e.g. the afternoon), 2) page category (e.g. sports) and 3) the number of impressions. Optimal ad time and location that maximizes overall revenue are successfully determined by using these three features, rather than relying only on the CTR of an individual ad. It is proved that their strategy showed an enhancement over greedy and random methods.

2.3.4 The Perspective of an Ad Exchange

Increasingly, display advertisements (ads) on the Internet are sold via marketplaces that bring publishers and advertisers together in real time when an opportunity arises to present an ad to a viewer (Muthukrishnan, 2009). These marketplaces are Ad Exchange (ADX). According to Yuan, et al. (2012), the ad exchange (ADX) is considered a uniform marketplace for publishers to sell ad inventories, and for advertisers to buy impressions and clicks. In simple words, ADX serves as the interface between publishers and advertisers in the Internet advertising ecosystem.

Therefore, to ensure the valuable connection between publishers and advertisers, an ADX needs to fulfill the twin goal of increasing revenue (shared between publisher and ad-network), and improving user experience. Nowadays the generalized second price auction (GSP) is the most adopted model in ADX (Yuan, et al. 2012). To add to that, Edelman, et al. (2007) demonstrates that GSP is when the advertiser pays the next highest bid instead of their own bid price. In addition, due to the probability that an ad will be clicked (CTR) will determine the value of the ADX revenue, ADX needs to consider the relevancy. If the relevancy is less in value, users will retreat (due to less satisfaction), and refuses to come back. In a follow-up work (Lacerda, et al. 2006) the authors propose a method to learn impact of individual features using genetic programming to produce a matching function.

Furthermore, Neto, et al., (2005) proposed to generate an augmented representation of the target page by means of a Bayesian model built over several additional Web pages. Besides, Radlinski et al. (2008) proposed an online query expansion algorithm of two stages: the offline processing module precomputes query expansions for a large number of queries, and then builds an inverted index from the expanded query features. Later, Broder et al. (2008b) proposed a method for both augmenting queries and ads. Three distinct spaces of different features are used to represent queries: unigrams, classes and phrases extracted using a proprietary variant of Altavista's Prisma refinement tool.

2.4 Matching Contextual Ads and Web Page Contents

The goal of the advertising is simple: to bring the desired message to exactly the right customers wherever they are online. Advertisements provide the necessary revenue model supporting the Web ecosystem and its rapid growth. Contextual Advertising (ConAd) refers to the placement of ads that are contextually related to the Web page content (Kushal, 2011). Targeted or contextual ad insertion plays an important role in optimizing the financial return of this model (Yang, et al. 2010). Contextual advertising is a type of Internet advertising used for content- based Web sites that make targeted ads appear on a Web page based on the page's actual content.

Typically, a code snippet is placed on the page, and the code "figures out" what the page is about and serves appropriate advertisements from a large database of advertisers. This type of advertising works really well for Website owners because the link titles are relevant to the content and they do not get ignored like most advertising. For the advertiser this is good because visitors are already interested with the Web page content, then ads are relevant to that content, so advertisers get a better Return of Investment (ROI).

Since the emergence of Computational Advertising theory, there are many approaches have been proposed by researchers and developers in creating the best ads matching system, in order to get the "best match" between Internet ads and the Web page contents besides leveraging on the revenue part. This new advancement will help advertisers and publishers connect with their potential customers in the right place and at the right time, by placing relevant ads on their Web pages. It is able to bring to a win-win situation for both Web users and advertisers: on one hand, users can have better experiences during browsing Web content and facing relevant ads they might be interested in, instead of annoying irrelevant ads; on the other hand, relevance between content and ads leads users to click and browse the ads and brings the advertisers potential increase in revenue.

Traditionally, a contextual advertising system scans and extracts the content of a Web page for keyword phrases. This system will then return targeted ads based on the content the people accessing the Website are viewing. Lots of research has been done in keyword extraction field. As the key step of contextual advertising, keyword extraction affects the accuracy of the advertising system directly, and a lot of research has been done on English keyword extraction.

Furthermore, Lang (1995) describes a system which separates advertisements and publishers' Web sites by introducing an advertisement agent. The agent sits between advertisers and the user's browser and merges banner advertisement directly into the currently viewed page, independent of the page itself. A number of companies have introduced similar systems, for example displaying advertisement while the user waits for Web content to download, but these have so far failed to gain widespread acceptance. In this 21st century, this technique has now been improved and advanced through the existence of ad exchange and ad network in the Internet advertising ecosystem as explained in Section 2.3.

According to Frank, et. al. (1999), KEA keyphrase extraction algorithm, extracts keywords using a simple machine learning mechanism. It shows a simple procedure for keyphrase extraction based on the naive Bayes learning scheme performs comparably to the state of the art. The naive Bayes learning scheme is a simple application of Bayes' formula. It assumes that the attributes – in this case *TF×IDF* and distance – are independent given the class. Making this assumption, the probability that a phrase is a key phrase given that it has discretized *TF×IDF* value T and discretized distance *D* is:

$$\Pr[key|T,D] = \frac{\Pr[T|key] \times \Pr[D|key] \times \Pr[key]}{\Pr[T,D]}$$

where Pr[*T/key*] is the probability that a keyphrase has *TF×IDF* score *T*, Pr[*D\key*] the probability that it has distance *D*, Pr[*key*] the probability that a phrase is a keyphrase, and Pr[*T,D*] a normalization factor that makes Pr[*key/ T, D*] lie between zero and one. KEA uses the procedure described above to generate a naive Bayes model from a set of training documents for which keyphrases are known. The resulting model can then be applied to a new document from which keyphrases are to be extracted.

On the other hand, Yih, et al. (2006) make a great improvement over KEA. The general architecture of the keyword extraction system consists of the follow four stages:

i. *Preprocessor*

    The main purpose of the preprocessor is to transform an HTML document into an easy-to-process plain-text based document, while still maintaining important information candidate selector

ii. *Candidate Selector*

    The system considers each word or phrase (consecutive words) up to length 5 that appears in the document as a candidate keyword. This includes all keywords that appear in the title section, or in meta-tags, as well as words and phrases in the body. Putting together these possibilities, Yih, Goodman, and Carvalho ended up considering three different candidate selectors:

- *Monolithic, Separate (MoS)* – fragments that appear in different document locations are considered as different candidates even if their content is identical.
- *Monolithic, Combined (MoC)* – reduces the number of candidates by combining identical (case ignored) fragments.
- *Decomposed, Separate (DeS)* – tries to assign a label to each word in a document, as is done in related fields.

iii. *Classifier*

The classifier needs to output some kind of confidence scores or probabilities. The scores or probabilities can then be used later to produce a ranked list of keywords, given a document. The actual probability returned is:

$$P(Y = 1 | X = \bar{x}) = \frac{\exp(\bar{x}.\overline{w})}{1 + \exp(\bar{x}.\overline{w})}$$

iv. *Postprocessor*

After the classifier predicts the probabilities of the candidates associated with the possible labels, our keyword extraction system generates a list of keywords ranked by the probabilities.

Apart from that, Ribeiro-Neto et al., (2005) describes an Impedance Coupling technique for content-targeted advertising which expands the text of the Web page to reduce vocabulary impedance with regard to an advertisement, can yield extra gains in average precision of 50%. For this, they proposed to expand the triggering pages with new terms. Figure below illustrates their intuition.
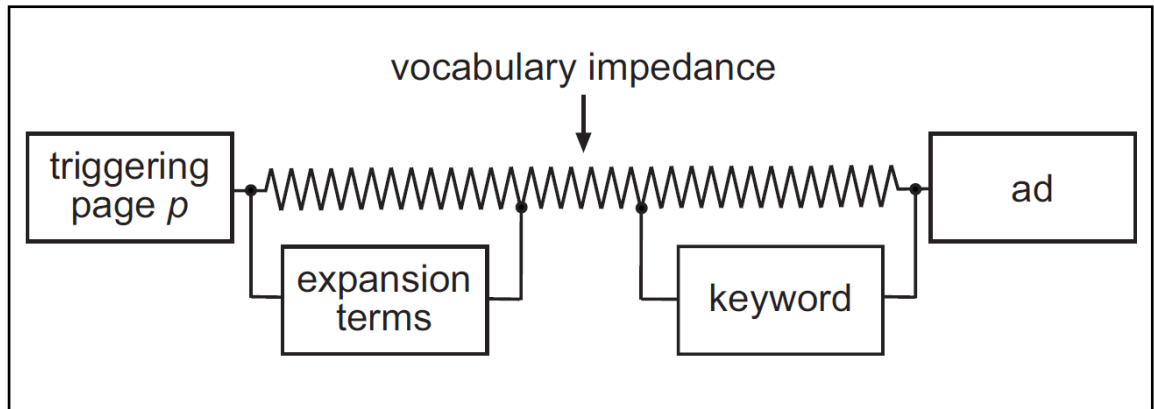
Figure 10: Addition of new terms to a Web page to reduce the vocabulary impedance.

Most importantly, their work focused not on finding keywords on Web pages, but on directly matching advertisements to Web pages. They used a variety of information, including the text of the advertisements, the destination Web page of the ad, and the full set of keywords tied to a particular ad (as opposed to considering keywords one at a time.) They then looked at the cosine similarity of these measures to potential destination pages.

More to add, other works have addressed the ads targeting issue includes a proposed system called ADWIZ that is able to adapt online advertisement to a user's short-term interests in a non-intrusive way, suggested by Langheinrich et al., (1999). Contrary to the work done by Ribeiro-Neto et al., ADWIZ does not directly use the content of the page viewed by the user. It relies on search keywords supplied by the user to search engines and on the URL of the page requested by the user. The ADWIZ system consists of four principal components, *the ad server* (1) handles the selection and actual delivery of the advertisement banner to the user; *a separate database server* (2) provides a central storage facility for all parts of the system, effectively decoupling each component and providing asynchronous communication; *a learning system* (3) runs periodically over aggregated performance statistics and dynamically calculates a set of display probabilities used by the advertisement selection system of the ad server; and *an administration server* (4) for inspecting and manipulating the database content such as advertisements and their properties, advertisement campaigns, and calculated display probabilities.

Meanwhile, Lacerda et al. (2006) proposed to use machine learning to find good ranking functions for contextual advertising. They use the same dataset described in the paper by Ribeiro-Neto et al. (2005). They use part of the data for training a model and part for evaluation purposes. They apply a genetic programming algorithm to select a ranking function which maximizes the average precision on the training data. The resulting ranking function is a non-linear combination of simple components based on the frequency of ad terms in the target page, document frequencies, document length and size of the collections. Lacerda et al. (2006) find that the ranking functions selected in this way are considerably more accurate than the baseline proposed in Ribeiro-Neto et al. (2005).

Additionally, according to Anagnostopoulos et al. (2006), for static pages that are displayed repeatedly, the matching of ads can be based on prior analysis of their entire content; however, often ads need to be matched to new or dynamically created pages that cannot be processed ahead of time. Thus their work focused on the contributions of the different fragments of the pages. Extracting small but informative parts of pages is important because often page content is not available for analysis ahead of time, as is the case for dynamically created or frequently updated pages.

Their approach allowed matching ads to pages in real time, without prior analysis of the page content. They employed text summarization techniques to identify short but informative page fragments that can serve as a good proxy for the entire page. Plus Anagnostopoulos et al. (2006), also use two source of external knowledge. Experimental findings confirmed that using only a small portion of the page text can yield highly relevant ads, and the quality of summary-based ad matching is competitive with that of using the full page.

Murdock et al. (2007) consider machine translation to overcome the vocabulary mismatch between target pages and ads. In more detail, the machine translation features they use correspond to the average translation probability of all words in the target page translated to the keywords or to the description of the ad, and the proportion of translations of the ad terms, or the ad keywords, that appear on the target page.

Murdock et al. (2007) report that the machine translation probabilities produce statistically significant improvements in precision at rank one compared to a baseline, where the cosine similarity between the target page and each of the ad fields is weighted separately.

On the other hand, Chakrabarti et al. (2008) proposed a new class of models to combine relevance with click feedback for a contextual advertising system. Their model is based on a logistic regression and allows for a large number of granular features. The key feature of the modeling approach is the ability to model interactions that exist among words between page and ad regions in a way that is suitable for efficient evaluation over inverted indexes. In fact, they employ a multiplicative factorization to model the interaction effects for several (page, ad) regions in a parsimonious way that facilitates fast look-up of ads at run time. Through large scale experiments, they convincingly demonstrate the advantage of combining relevance with click feedback. In fact, they achieve a 25% lift in precision for a recall value of 10% relative to a pure relevance based model in their experiments.

Broder et al. (2007) notice that the standard string matching approach can be improved by adopting a matching model which additionally takes into account topical proximity. In their model the target page and the ad are classified with respect to taxonomy of topics. The similarity of ad and target page estimated by means of the taxonomy provides an additional factor in the ads ranking function. The taxonomy, which has been manually built, contains approximately 6,000 nodes, where each node represents a set of queries. The concatenation of all queries at each node is used as a meta-document, ads and target pages are associated with a node in the taxonomy using a nearest neighbor classifier and *tf-idf* weighting. The ultimate score of an ad a_i for a page p is a weighted sum of the taxonomy similarity score and the similarity of a_i and p based on standard syntactic measures (vector cosine).

On evaluation, Broder et al. (2007) report a 25% improvement for mid-range recalls of the syntactic-semantic model over the pure syntactic one. Additionally, there are several work and findings that highlight the issue through semantic approach. Ciaramita et al., (2008) presented a machine learning approach to contextual advertising using a novel set of features which aims to capture subtle semantic associations between the vocabularies of the ad and the Web page.

Besides that, Fan and Chang (2009) put forward their ideas with regards to matching contextual ads and Web page through a novel framework for associating ads with blog pages based on sentiment analysis. They investigated the sentiments of blog pages and utilized this information to demonstrate sentiment-oriented contextual advertising. For page-ad matching, they evaluated their framework using 150 blog pages and over 100,000 ads sampled from Google AdSense. First, they compared SOCA with Google AdSense and found that their proposed method with sentiment detection can achieve superior performance (74.1% precision). To analyze their SOCA in detail, they evaluated three matching strategies (i.e., cosine similarity, ontology similarity and the combined approach). The results indicated that the three proposed approaches can assign relevant ads to the positive (and neutral) aspects of a blog page. Besides, the combined approach has a better performance than cosine and ontology independently.

More to add, Hatch et al. (2010) introduced "clickable terms" approach to contextual advertising. This approach involves matching a Web site directly with a set of ad-side terms, independent of the page content. They use log-likelihood ratios (LLRs) to measure the relative clickability of a given ad-side term on a given site. These LLRs are derived from a Bayesian framework for measuring click probabilities of ads on a given site. The LLRs can be shown to be optimal for measuring posterior click probabilities of ads when certain conditional independence assumptions hold.

They use the LLRs as a basis for defining various new attribute representations for an ME click model. They also introduce various techniques for scaling or normalizing the LLR-based features in an ME model. The overall "clickable terms" approach described in this paper achieves significant gains in normalized discounted cumulative gain (NDCG) over a baseline click model that incorporates only lexical match features. The clickable terms model also yields modest, but significant gains in click through rate (CTR) and revenue per impression (RPI) in tests performed on live Web traffic.

Other than that, Wu et al. (2011) suggested a new approach by incorporating the Wikipedia concept and category information into the traditional keyword matching to enrich the content representation of pages and ads. They described how to map each ad (or page) into a keyword vector, a concept vector and a category vector, as well as how to combine the three feature vectors together for making the top-N ads selection. Based on their experiment, the system performance is leveled up. Due to that they used the Wikipedia knowledge to enrich the semantic representation of pages and ads, and use them to measure the semantic similarity between pages and ads; while, compared to the surface text information contained in pages and ads, the semantic information has a better stability, i.e., it can reflect out the similarity between pages and ads more accurately.

Recent views on this matching system had been done by Joshi et al. (2013) where they make contextual targeting more relevant with Extraction of relevant entities from the Web page. They extract the entities from Web page, which is of interest to the consumer. Then they target the interest of Internet user and put up the ads according to their interest. The system is designed in such a way that it can extract entities (Name, Place, Title, Location, date etc) from Web page and ad publisher put up a advertise on that page which include those entities which are extracted from page. They also presented how rule-based systems provide a convenient method of defining extraction patterns spanning over various properties of the tokens and the context in which it resides. Through extraction of different entity can help wide area of products for ad publisher for ad marketing. With reference of multiple entities create closed group of similar entity. It helps to present relevant data on Web pages.

CHAPTER 3
METHODOLOGY

It is crucial to choose the right methodology in developing the products of the research study; which are Web pages and ads matching system. Different methodologies could cater different needs of a project in a period of time. Thus, this chapter will elaborate on:

- Product Development Phases
- Project Activities
- Tasks Schedule and Key Milestones
- Tools for Development

3.1 PRODUCT DEVELOPMENT PHASES

At the end of this study, Web pages and ads matching system will be constructed upon demonstrating the theory of Computational advertising on the Contextual Ads selection process in order to match them with the Web pages contents. Hence, in completing this project, the main methodology used is the Rapid Application Development (RAD) method.

This methodology is designed for faster development and higher-quality results than those achieved with the traditional lifecycle. It is designed to take the maximum advantage of powerful development that has evolved recently (Martin J., *n.a*). Rapid Application Development (RAD) method takes an approach whereby minimal planning is required and focuses more towards rapid prototyping.
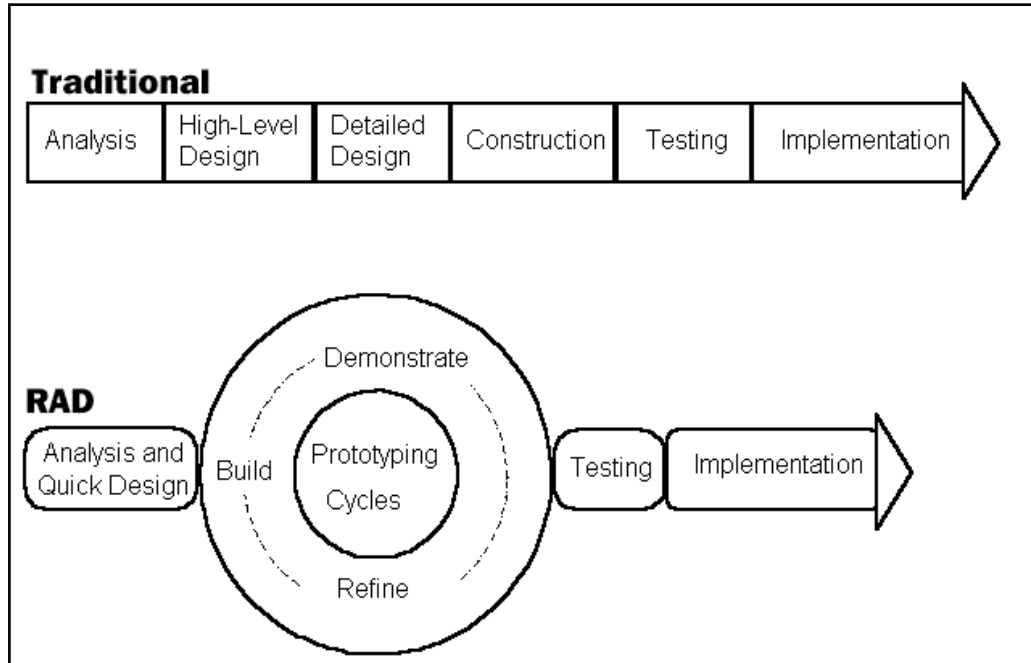
Figure 11: Rapid Application Development Methodology

The reason on why this methodology is chosen is due to the constraint of the time provided to finish this project, which in total is less than 10 months. In addition, there might be possibilities of functionality and performance compromising along the development process. This allows flexibility towards any changes to be made during the development phase if there is needed to be reviewed and rechecked at any phase. Thus, in short, the Rapid Application Development (RAD) method enables quality project to be developed faster and saving valuable resources rather than traditional approach.

3.2 PROJECT ACTIVITIES

This methodology provided four main phases in order to ensure the project progress meets the milestones that have been set in front. The phases with the project activities are subcomponents inside Key Milestones of this project. The tasks involved are shown in the table below:

Phase	Project Activities
i. **Analysis and Quick Design**	• Literature review and research on the subject matters (Advertising, Internet advertising and Its Players, Computational advertising, Contextual Advertising, Matching Ads and Web pages Contents, etc) • Prepare survey questionnaires for data gathering and carry out survey to 65 respondents • Sketch brief test plan • Develop ads matching system flowchart, database model and database design • Design user-interface and create contents for Web pages • Prepare Gantt chart for task accomplishment and progress • Develop key milestones that mark significant points in the development process
ii. **Prototyping Cycles**	i. *Build* • Prototyping the user interface and Web pages contents according to the test plan sketched • Create ads database framework by using OpenX • Write codes and develop ads matching system in order to match ads and Web pages contents

	ii. *Demonstrate* - Run simple test to show the workability of the prototype (the ads matching mechanism) - Ensure all the components in the prototype interrelated and working fine - Publish Web pages and the ads matching system on live server
	iii. *Refine* - Reconstruct and revamp the ads matching mechanism of the prototype according to the test plan prepared for improvement purpose, after reviewing the system faults of the previous version - Revert back to the first and second prototyping cycles (build and demonstrate) if needed, until the system meets all requirements and objectives
iii. **Testing**	- Evaluate the system functionality and usability on live server - Check the specifications of the prototype whether they are aligned with the requirements and meet the objectives of the study
iv. **Implementation**	- The prototype is ready to be used when it has passed all the testing phase - Provide further recommendations for the system

Table 1: Project Activities

3.3 TASK SCHEDULE (GANTT CHART) AND KEY MILESTONES

All main tasks in developing this project have been tabulated in a Gantt chart for a better view. The timeline-based graphical approach is useful to visualize the current progress of the project and any chances of delay, defect or prototype's failure. For that reason, below is the task schedule with respect to the time frame given by the coordinator of Final Year Project II of Computer Information Sciences (CIS).

First Term : Final Year Project I															
No	Tasks	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Analysis and Quick Design															
1	Literature review and research on the subject matters	■	■	■	■	■	■	■							
2	Prepare survey questionnaires for data gathering and carry out survey to 65 respondents								■	■					
3	Sketch brief test plan										■				
4	Develop ads matching system flowchart, database model and database design										■	■			
5	Design user-interface and create contents for Web pages											■	■	■	■
6	Prepare Gantt chart for task accomplishment and progress	■													
7	Develop key milestones that mark significant points in the development process	■													

49

No	Tasks	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Final Term: Final Year Project II														

No	Tasks	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Prototyping Cycles															
-Build															
8	Prototyping the user interface and Web pages contents according to the test plan sketched	▓	▓	▓	▓										
9	Create ads database framework by using OpenX					▓									
10	Write codes and develop ads matching system in order to match ads and Web pages contents						▓	▓							
-Demonstrate															
11	Run simple test to show the workability of the prototype (the ads matching mechanism)								▓	▓					
12	Ensure all the components in the prototype interrelated and working fine								▓	▓					
13	Publish Web pages and the ads matching system on live server								▓	▓					
-Refine															
14	Reconstruct and revamp the ads matching mechanism of the prototype										▓				

No	Task	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14
15	Revert back to the first and second prototyping cycles (if needed)										▓				
Testing															
16	Evaluate the system functionality and usability on live server											▓			
17	Check the specifications of the prototype whether they are aligned with the requirements and meet the objectives of the study											▓			
Implementation															
18	The prototype is ready to be used when it has passed all the testing phase												▓	▓	▓
19	Provide further recommendations for the system												▓	▓	▓

Table 2: Gantt Chart and Key Milestones

3.4 TOOLS FOR DEVELOPMENT

Tools that have been opted for this project are:

i. *OpenX*

OpenX is an ad server that can be used to manage and optimize the advertising space on one or more Websites. In this project, this tool is used to manage the ad server and measure the statistics of ad banners that are placed on the Web pages. OpenX acts as the connector between publisher and advertisers, giving out the same roles as ad exchange.

ii. *Apache HTTP Server*

Apache HTTP Server is an open-source Web server platform. In this project, this tool is utilized to deliverer Web pages through the Internet. When the Apache program receives the request for a file, it looks for the file on its disk, and when found, sends that file to the requester in a stream of data named HyperText Transfer Protocol, which is then decoded by the browser program and rendered as a Web page on the requester's screen.

iii.   *Wordpress*

In this project, Wordpress is used in the development of different Web pages. It is a semantic personal publishing platform with a focus on aesthetics, Web standards, and usability. WordPress is completely customizable and can be used for almost anything.

iv.   *Adobe Photoshop CS5*

Adobe Photoshop CS5 is a graphic-editing program developed and published by Adobe Systems. In this project, this tool is used to design the content and user-interface of Web pages. Besides it is also used to design ad banners that are stored in the ad server in the system.

CHAPTER 4
RESULTS AND DISCUSSION

This chapter is a core component of the whole progress report whereby results and justifications are discussed here. All results gained during online survey are tabulated and analyzed in this section. It also includes the System Flow Diagram, System Modeling, and System Prototype Sketch.

4.1 SURVEY ANALYSIS

The author conducted a survey, mainly aimed to identify what are the Internet users' perspectives and views on Internet advertising. The survey is conducted online, targeting responses from 65 Malaysian Internet users, specifically. The results of the surveys are very important in order for the author to deduce further conclusions on the Malaysian Internet users' perspectives on how they actually perceive and react to the online ads, advertised on the Web pages.

There are 22 close-ended questions being asked in the survey. There are two sections of questionnaires:

i. *Section 1: Demographic Data*

In this section, author has collected the demographic data of the 65 respondents of the online survey with the intention of obtaining the specific information of the respondents. This is because advertising relies heavily on demography, since service and goods providers need specific information to reach the maximum number of potential customers in their target audience. These data are then used in analyzing certain patterns and trends related to the project.

Seven questions are asked, pertaining to the demographic data of the respondents. The questions and analysis are as followed:

No.	Result and Analysis
1.	**What is your gender?** Female [37] Male [28] Male 28 43% Female 37 57% Analysis: Question 1 is referring to respondents' gender. 57% of the respondents are female and the remaining 43% are male. Hence, majority of the respondents that respond to the survey questionnaires are females.
2.	**What age group below you belong to?** less than 18 18 -25 26-35 36-50 over 50 0 11 22 33 44 55 66 less than 18 0 0% 18 -25 57 88% 26-35 5 8% 36-50 2 3% over 50 1 2% Analysis: Question 2 is referring to respondents' age group. Majority of the respondents are from 18-25 age groups (teenagers and young adults) which comprises 88% of the population sample. The remaining are from 26-35 age groups (8%), followed by the 26-50 age groups (3%) and the least, over-50 age groups (2%). Hence, it can be concluded that people from the 18-25 age groups are the major user of the Internet.
3.	**What is your education level?** High school Some college Diploma Degree Masters PHD 0 12 24 36 48 60 72 High school 1 2% Some college 0 0% Diploma 2 3% Degree 61 94% Masters 1 2% PHD 0 0% Analysis: Question 3 is referring to respondents' current education level. 61% of the respondents hold a degree, only 3% with a diploma, and 2% with Masters and the other 2% respondents hold high school certificates. It can be deduced that majority of the Internet users hold a Degree.
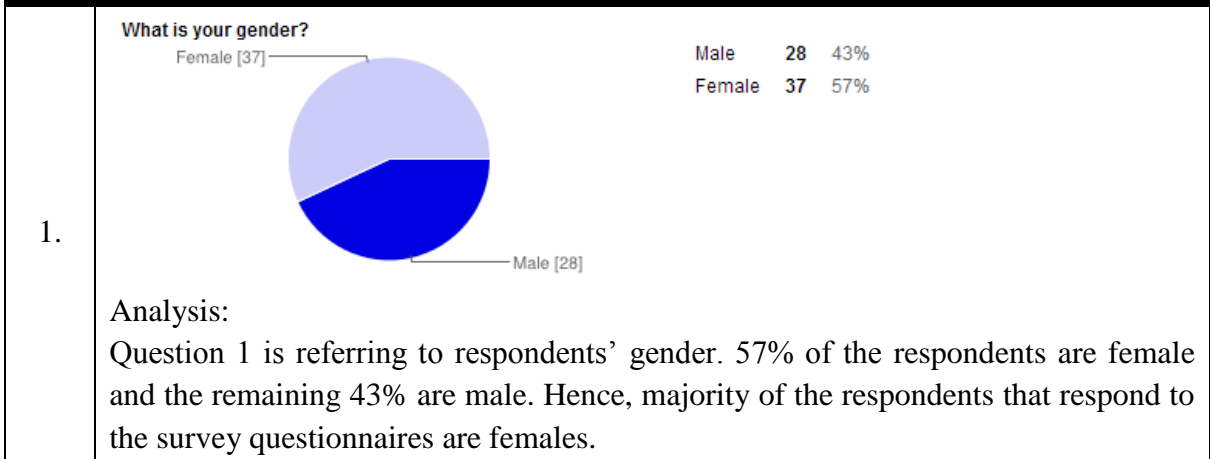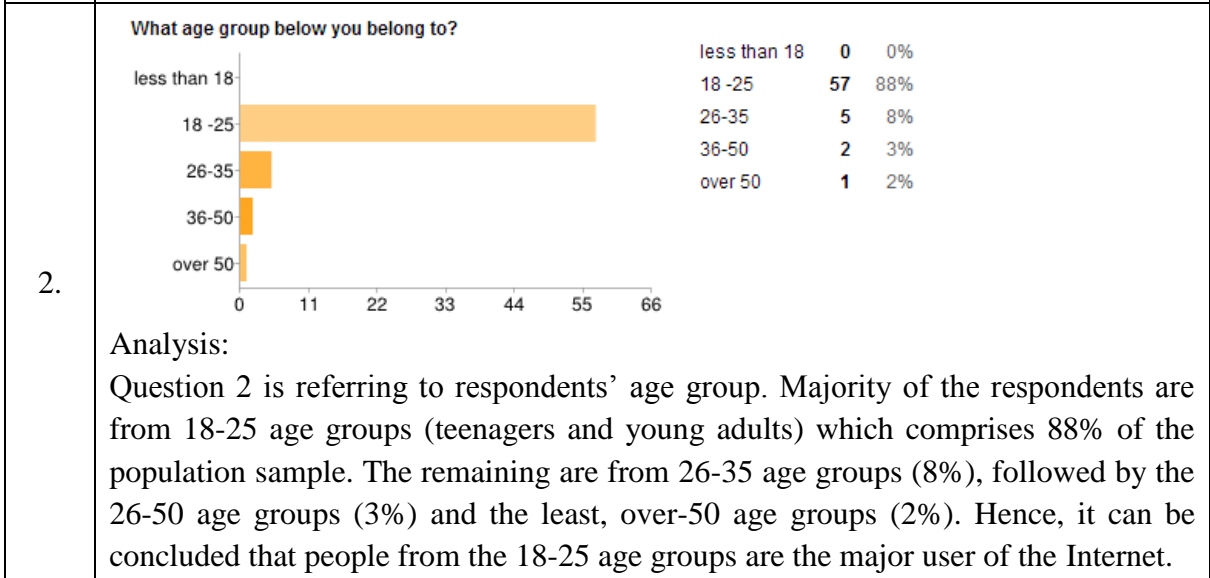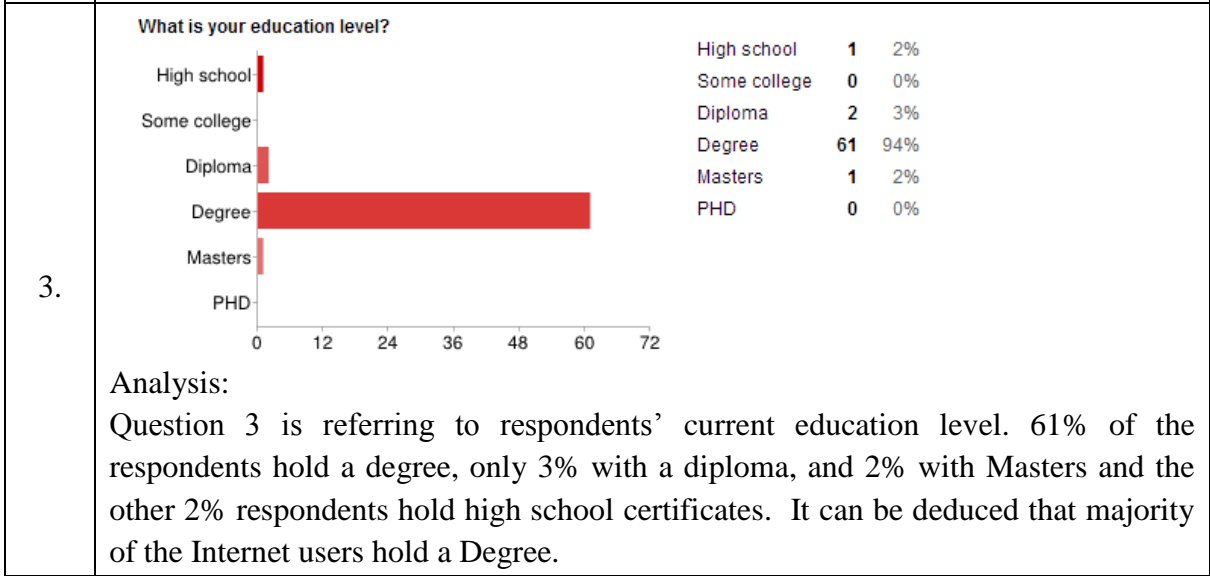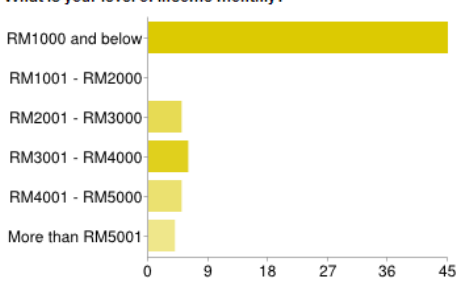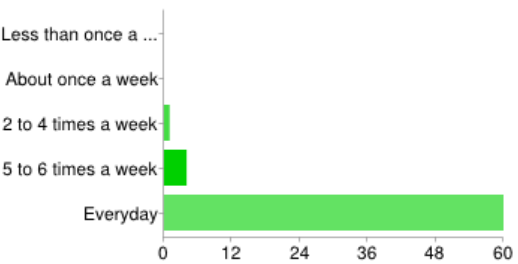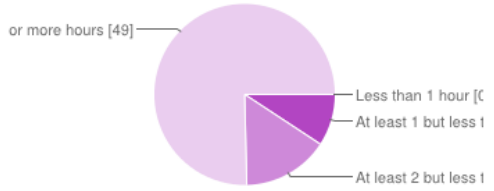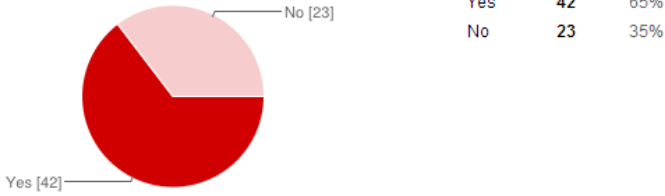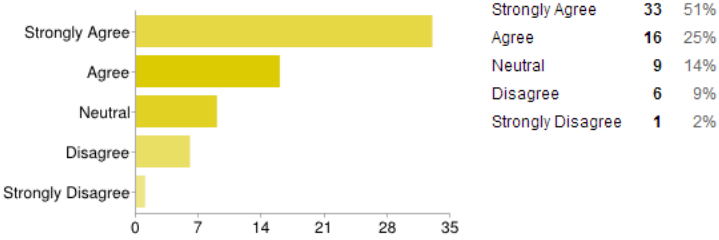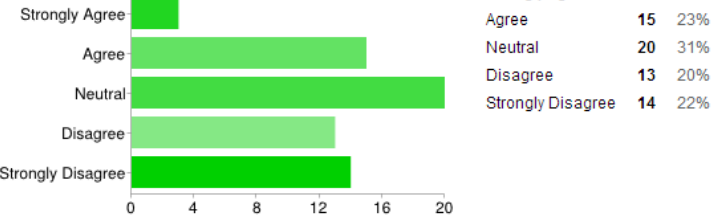
4.	**What is your level of income monthly?** RM1000 and below RM1001 - RM2000 RM2001 - RM3000 RM3001 - RM4000 RM4001 - RM5000 More than RM5001 (0 9 18 27 36 45) RM1000 and below 45 69% RM1001 - RM2000 0 0% RM2001 - RM3000 5 8% RM3001 - RM4000 6 9% RM4001 - RM5000 5 8% More than RM5001 4 6% Analysis: Question 4 is about the respondents' monthly level income. Majority (69%) are currently having monthly income of RM1000 and below. 9% with RM3001-RM4000, 8% with RM2001-RM3000 and RM4001-5000 respectively, while the remaining 6% with an income level more than RM5001 monthly. Hence this can be deduced that most of the respondents have low-income level. Probably because they are still studying and have not gained any fixed monthly earnings.
5.	**How often do you go online and use the Internet?** Less than once a ... About once a week 2 to 4 times a week 5 to 6 times a week Everyday (0 12 24 36 48 60) Less than once a week 0 0% About once a week 0 0% 2 to 4 times a week 1 2% 5 to 6 times a week 4 6% Everyday 60 92% Analysis: Question 5 is asking the respondents on how often do they go online and use the Internet. 92% of the respondents said that they go online and use the Internet every day. While 6% said 5 to 6 times a week and the remaining 2% said 2 to 4 times per week. Therefore this reflects that most of the respondents are online and using the Internet regularly in their daily life.
6.	**When using the internet, about how long do you stay logged on (consider the whole day)?** or more hours [49] Less than 1 hour [0 At least 1 but less t At least 2 but less t Less than 1 hour 0 0% At least 1 but less than 2 hours 6 9% At least 2 but less than 5 hours 10 15% 5 or more hours 49 75% Analysis: Question 6 is about the duration that the respondents stay logged on when using the Internet. 75% of the respondents said they stay logged on the Internet, 5 hours or more in a day. 15% said at least 2 hours but less than 5 hours. While the remaining 9% said they are on the Internet at least 1 hour but less than 2 hours. Thus it can be said that majority spends 21% of their time daily on the Internet.

No.	Result
7.	**What is your most favorite general search engine?** Google [63] — Bing [0] / Yahoo! [1] / Volunia [0] / Other [1] Google 63 97% Bing 0 0% Yahoo! 1 2% Volunia 0 0% Other 1 2% Analysis: Question 7 is asking the respondents regards to their most favorite search engine. Majority voted for Google (97%), while 2% said Yahoo! And the remaining 2% use other type of Search Engine. Hence it can be concluded that, most of the respondents use Google as their search engine and there is probability that they have seen and have been exposed to Internet Ads, as Google is one of the search engines that carry out robust Internet advertising programs.

ii. Section 2: About Internet advertising

In this section, the author had collected the information with respect to the Internet users' perceptions towards Internet advertising. This section is very important in order for the author to recognize the Internet users' insights and awareness about Internet advertising and its usability and effectiveness (in Malaysian context). The questions and analysis are as followed:

No.	Result
8.	**Have you ever heard of Internet Advertising?** Yes [61] — No [4] Yes 61 94% No 4 6% Analysis: Question 8 is asking the respondents whether they have heard about Internet advertising. Majority (94%) said Yes and the remaining 6% of the respondents have never heard of Internet advertising. Hence it can be concluded that majority of the Internet users are already exposed to Internet advertising but somehow there are still minority groups that need more information on what Internet advertising is.

57

9.	Have you ever clicked on an online advertisement (banner pop-up, online ad on your search engine/blog/website)? Yes 42 65% No 23 35% No [23] Yes [42] Analysis: Question 9 is asking the respondents whether they have ever clicked on an online ad. 65% said yes and the remaining 35% said no. Hence, majority said they have clicked on an online ad which reflects more than half of the respondents are familiar with the Internet ad and show some interests by clicking it or probably they were just accidentally clicked it.
10.	Questions 10 are divided into 12 sub-questions that asked the respondents on their level of agreeableness on the statements with respect to Internet advertising. The questions and its result analysis are as follow:
i.	- When you see an online ad, you do not click on it and avoid looking at it because it is annoying/offensive Strongly Agree 33 51% Agree 16 25% Neutral 9 14% Disagree 6 9% Strongly Disagree 1 2% Analysis: Based on the statistics, 51% of the respondents strongly agreed that they tend to avoid and do not click on the online ad. 25% agreed with the statement, 14% said neutral, 9% disagreed with the statement and the remaining 2% strongly disagreed. Hence it can be deduced that majority of the Internet users in Malaysia are avoiding to look or click on Internet ads even though they are aware of it. But still, there are minority groups that still show some interest in clicking and looking at the Internet ads.
ii.	- When you see an online ad, you click on it to seek more information Strongly Agree 3 5% Agree 15 23% Neutral 20 31% Disagree 13 20% Strongly Disagree 14 22%

	Analysis: Based on the statistics above, majority of the respondents said they are neutral with the statement (31%). 23% said they agreed with the statement and 5% with strong agreement. The other 22% and 20% belongs to those who strongly disagree and disagree to the statement, respectively. Therefore it can be deduced that most of the Internet users either clicked on the online ads to seek more information of the ads or they already saw the information on the ad banners and they are not interested to dive in to seek more information.
iii.	- When you see an online ad, you click on it because it matches your interest and preferences |		
---	---	---
Strongly Agree	11	17%
Agree	27	42%
Neutral	13	20%
Disagree	8	12%
Strongly Disagree	6	9%	 Analysis: The statistics show that majority (42%) of the respondents agreed that they clicked on the online ads because the online ads match their interest and preferences. 17% of the respondents are strongly agreed with the statement while 20% said neutral. On the other hand, 12% disagreed with the statement and the remaining 9% strongly disagreed with the statement. Hence it can be said that most Internet users prefer to click on Internet ads that match to their preferences and interest.
iv.	- When you see an online ad, you click on it because it is very interactive (got videos, sounds) |		
---	---	---
Strongly Agree	4	6%
Agree	11	17%
Neutral	14	22%
Disagree	23	35%
Strongly Disagree	13	20%	 Analysis: The statistics above show that 35% of the respondents disagreed and 20% strongly disagreed with the statement. 22% are neutral with the statement and the remaining 17% and 6% belongs to those who agreed and strongly agreed with the statement. Thus, it can be said that commonly, Internet users are not clicking on the Internet ads because of the ad banners' interactivity. In brief, Internet ads are not necessarily need to be interactive in order to attract users to click on it.
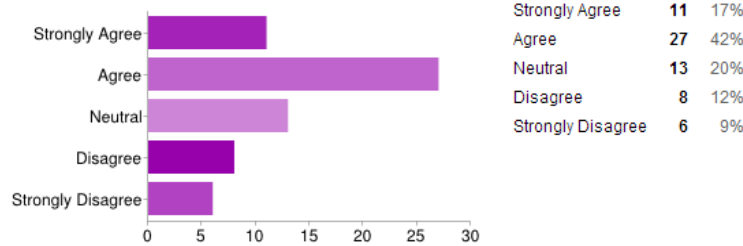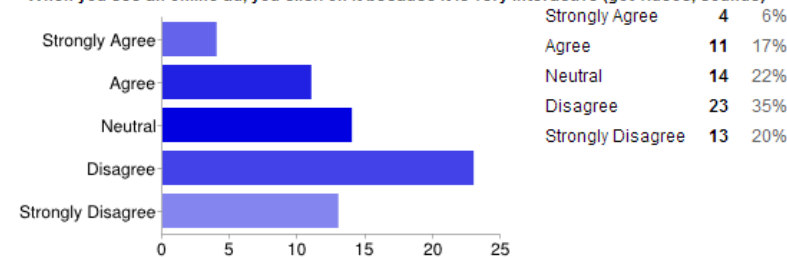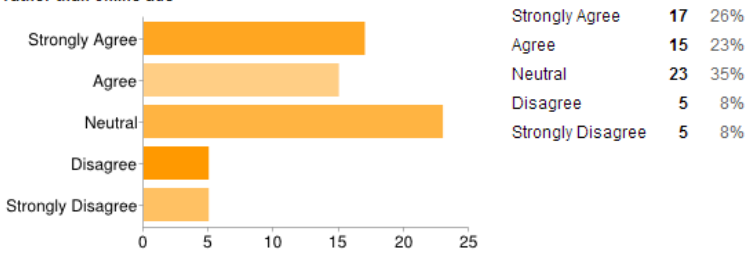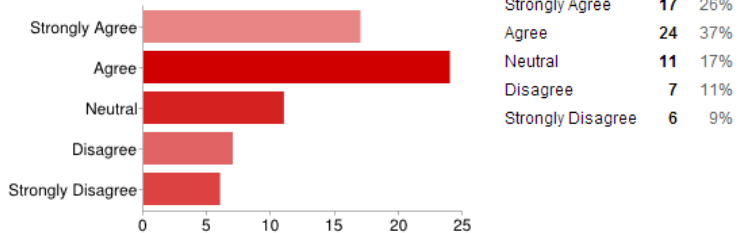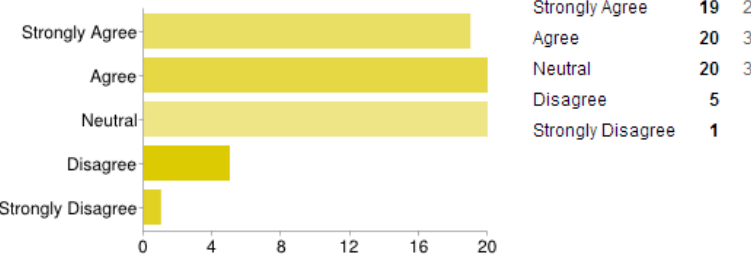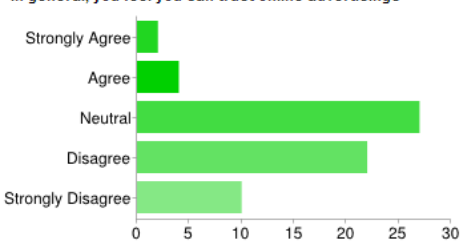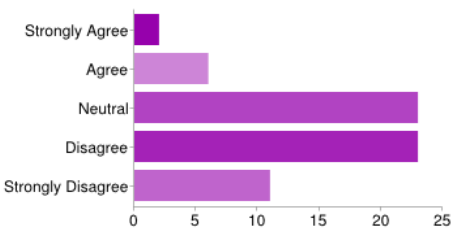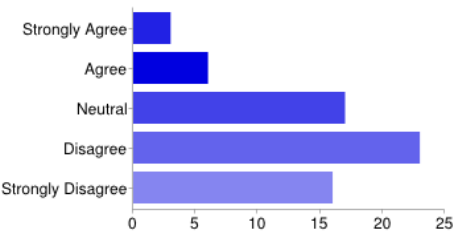
v.	- You always prefer the classic advertisement mediums (on newspapers, television, radio) rather than online ads

Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree bar chart

Strongly Agree	17	26%
Agree	15	23%
Neutral	23	35%
Disagree	5	8%
Strongly Disagree	5	8%

Analysis:

Based on the graph above, 35% of the respondents are neutral with regards to the statement. But 26% strongly agreed and 23% agreed with the statement. While 8% of the respondents disagreed and strongly disagreed (8%) with the statement. Hence it can be deduced that in Malaysia, Internet users still prefer the classic advertisement mediums rather than online ads. |
| vi. | - You always accidentally clicked the online ads

Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree bar chart

Strongly Agree	17	26%
Agree	24	37%
Neutral	11	17%
Disagree	7	11%
Strongly Disagree	6	9%

Analysis:

The statistics show that 37% of the respondents agreed with the statements. 26% strongly agreed, 17% on the neutral par and on the other hand, 11% disagreed and 9% disagreed with the statement. Therefore it can be said that most of the Internet users always accidentally clicked on online ads, when they are not intended to do so. |
| vii. | - You think the online ads always misled you

Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree bar chart

Strongly Agree	19	29%
Agree	20	31%
Neutral	20	31%
Disagree	5	8%
Strongly Disagree	1	2%

Analysis:

The graph shows that 31% of the respondents agreed that online ads always misled them and 29% also strongly agreed with the statement. 31% on the neutral side while the other 8% disagreed and 2% strongly disagreed with the statement. Thus, it can be concluded that majority of the Internet users think Internet ads always misled them from their first intention and purposes on clicking the ads. |

viii.	- In general, you feel you can trust online advertisings Strongly Agree 2 3% Agree 4 6% Neutral 27 42% Disagree 22 34% Strongly Disagree 10 15% Analysis: Based on the graph above, 34% of the respondents disagreed and 15% strongly disagreed with the statement. 42% of the respondents are on the neutral par and on the other hand, 6% agreed and 3% strongly agreed with the statement. Hence it can be deduced that most of the Internet users believe that they cannot put trust on Internet advertising.
ix.	- In general, you feel very confident in using information you see in the ads to make a purchase decision Strongly Agree 2 3% Agree 6 9% Neutral 23 35% Disagree 23 35% Strongly Disagree 11 17% Analysis: The graph shows that 35% of the respondents disagreed and 17% strongly disagreed with the statement. 35% of the respondents are on the neutral par and on the other hand, 9% agreed and 3% strongly agreed with the statement. Therefore it can be said that majority of the Internet users is not confident in using information they see on the Internet ads to make a purchase decision. Hence, transactions cannot occur.
x.	- You feel comfortable about purchasing an item through an online ads by giving your address or phone number Strongly Agree 3 5% Agree 6 9% Neutral 17 26% Disagree 23 35% Strongly Disagree 16 25% Analysis: Based on the statistics, 35% of the respondents disagreed and 25% strongly disagreed with the statement. 26% of the respondents are on the neutral par and on the other hand, 9% agreed and 5% strongly agreed with the statement. Hence, it can be concluded that majority of the Internet users are not comfortable about purchasing an item through an online ad by giving their personal information (address, phone number, account number, etc).

xi.	**- You think products purchased through online ads has much lower prices** Strongly Agree / Agree / Neutral / Disagree / Strongly Disagree bar chart (0–30 scale) Strongly Agree 6 9% Agree 12 18% Neutral 26 40% Disagree 14 22% Strongly Disagree 7 11% Analysis: The statistics show, 22% of the respondents disagreed and 11% strongly disagreed with the statement. 40% of the respondents are on the neutral par and on the other hand, 18% agreed and 9% strongly agreed with the statement. Thus it can be said that most of the Internet users believe that products purchased through online ads can either have much lower prices or much higher prices than the products purchased outside the Internet medium.
xii.	**- In general, I like online advertising and I enjoy it** Strongly Agree / Agree / Neutral / Disagree / Strongly Disagree bar chart (0–30 scale) Strongly Agree 2 3% Agree 12 18% Neutral 27 42% Disagree 14 22% Strongly Disagree 10 15% Analysis: Based on the statistics, 22% of the respondents disagreed and 15% strongly disagreed with the statement. 42% of the respondents are on the neutral par and on the other hand, 18% agreed and 3% strongly agreed with the statement. Hence it can be deduced that majority of the Internet users are feeling neutral about liking and enjoying Internet ads. They are not rejecting it completely, but at the same time, they are not accepting it wholly.

Table 3: Survey Results and Analysis

4.2 SYSTEM FLOW DIAGRAM

System flow diagram (flowchart) helps to clarify how things are currently working and how they could be improved. It also assists in finding the key elements of a process, while drawing clear lines between where one process ends and the next one starts. Developing a flowchart stimulates communication among participants and establishes a common understanding about the process in the system. Flowchart also uncovers steps that are redundant or misplaced. Hence, figure below illustrates the brief flowchart of the system:
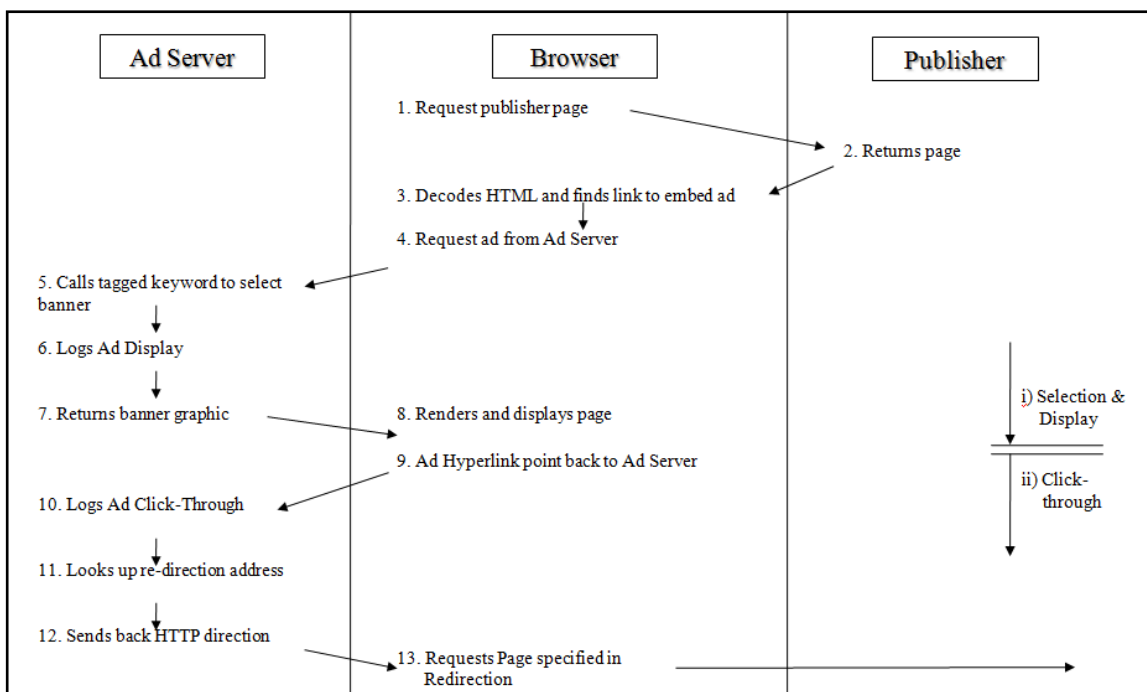
Figure 12: System Flow Diagram

Based on the flowchart above, the system basically has three main elements that will serve the system purpose which are ad server, browser and publisher. Ad server stores all of the ads and delivers them to the requests made by the Web pages through Web browser. Meanwhile Web browser displays the delivered ads form the ad server to the user and publisher returns the requested page from the Web server. More to add, Web server holds all the database of the Web pages and Web pages contents. Also, Web server stores the database of the ad server. Hence the ad server is the subset of the whole Web server.

63

4.3 SYSTEM MODELING

In system modeling phase, the author had model some of the system characteristics and features in order to recognize the behavior of the system as it evolves over the development period. This model will form a set of assumptions concerning the operation of the system. This model also used to study systems in the design stage before the system is being built.

Hence, table below states the system and its components:

System	Entities	Attributes	Activities	Events	State Variables
Ads Matching System	Ads; Web pages	Ad keywords and tags; Web page contents	Matching ads with Web page contents	Ad requests; Ad delivery; Ad display	Type of Web page contents; Ads tag; Ads schedule

Table 4: System and its components

More to add, the modeling phase of the system ranged in three phases based on the behaviors and features of the Web pages and the ads delivery:

i.   **Static Web page contents and static ads delivery method**, whereby the system has fixed Web pages contents and fixed ads display method. The system will only requests and retrieves ads from the ad server and matches them to the Web pages contents according to the ad tags with no manipulations and flexibility in ads display time.

In essence, this phase is developed to provide basic framework of the next phase of the system development stated, as follow.

ii. **Static Web page contents and dynamic ads delivery method**, whereby the system has fixed Web pages contents and the ads delivery and ads display time will vary over point in time. This means that, the system will match ads with the Web pages contents according to ad tags and ads schedule that had been set beforehand. These two variables are being considered in order to satisfy the dynamic feature of the ads delivery method.

In real-life situation, dynamic ads delivery method by scheduling the ads will actually keeping the ads relevant to the users. It will give out good impressions to the Internet users, as the displayed ads are regularly changing and being updated with the new one. Good impressions then will attract more users to click on the ads and this leads to profit gain.

iii. **Dynamic Web page contents and dynamic ads delivery method**, whereby the system has Web pages that are regularly having their contents changed, and simultaneously, the ads delivery and ads display time will vary over point in time. As the contents changed, the system will match ads with the changing Web pages contents according to ad tags and ads schedule that had been set beforehand. The system will be more robust in matching the ads with the Web pages contents as the contents keeps changing from time to time.

It is believe that this third phase of the modeling will simulate 99% of the real-life situations of the system. Therefore at the end of the system development, the outcome would yield the third phase of the modeling part, which is dynamic Web page contents and dynamic ads delivery method.

4.4 SYSTEM PROTOTYPE DESIGN

In simulating the real environment of ads matching system and getting the best match between contextual ads and the Web page, the author proposed to have an ads matching system that have an agent between publisher and advertiser in the Internet advertising ecosystem. The agent in between will act as a platform to serve publisher and advertiser in advertising their ads and to get more targeted customers that will likely click on their advertisements. Hence to set up this agent in the middle, the author used an ad serving management ad server which is OpenX Ad Server. OpenX Ad Server is the world's leading open source ad server. OpenX is a new of trading digital advertising programmatically. And it will soon be the path to strategically applying technology to create a more perfect marketplace for both buyers and sellers.

OpenX Ad Server is an ad server, created and published by the British-American company OpenX.org. It is a system that can be used to manage and optimize the advertising space on one or more Websites. It is a tool for Web publishers, but also for organizations running their own ad network. Furthermore, when managing a Website, or perhaps even multiple Websites (in this project), there can be a lot of work involved in putting advertisements on the site and removing them again, especially when this is done by entering the ads directly into the site's coding. It takes a lot of time and it reduces flexibility. Hence, OpenX can provide a real solution in this scenario.

Thus in this project, the author integrated the OpenX Ad Server with three Web pages. All three websites are developed by using Wordpress and its customized templates. Three Websites under the theme 'Photography' are created to demonstrate ads-matching varieties. Scoping down the 'Photography' theme, the first Website (named The Picstop) is mainly about giving out reviews to latest gadgets for photographers like cameras, lenses and camera accessories. The second Website (named The Shutter Up Photography) is a photography-service Website, whereby the owner of the Website is providing portrait photography service to its clients. The third Website (named Bridalsnaps) also provides photography service, but they are specializing in bridal photography, and taking photos for wedding receptions. Besides, Bridalsnaps also provides wedding photography tips.

Below are a few screenshots of all Websites built by the author:

    i.   The Picstop

Figure 13: The Picstop Website

ii.   The Shutter Up Photography

Figure 14: The Shutter Up Photography Website
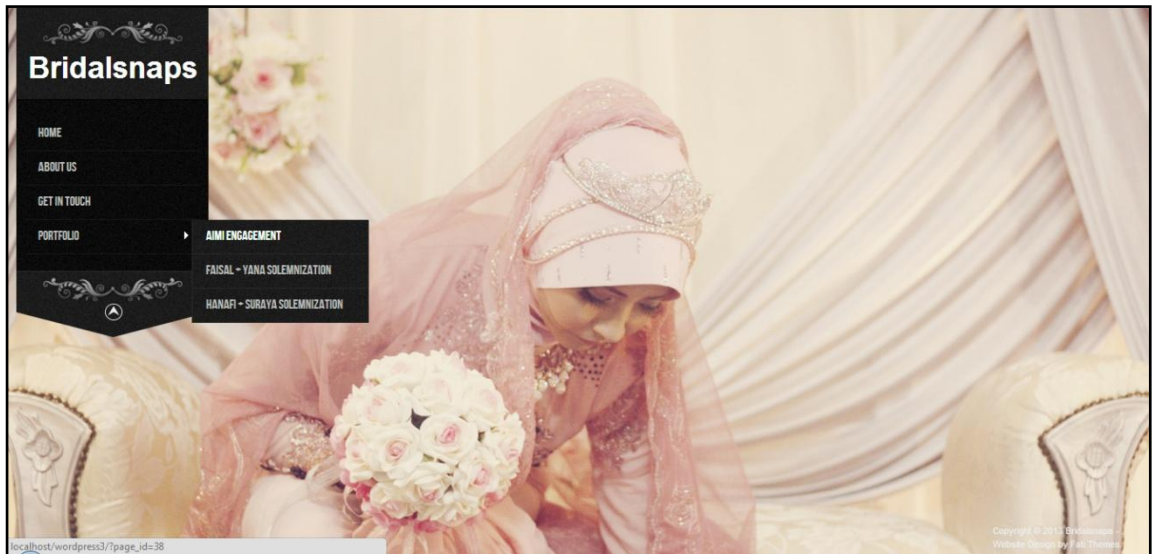
iii.   Bridalsnaps

Figure 15: Bridalsnaps Website

The owner of these Websites will then act as publishers. Later, the ads (contextual ads) paid by advertisers will be displayed on their Websites will mainly related to their Website contents, satisfying the relevancy of the displayed ads. Hence the final step is to link the right campaigns to the right zones, effectively determining which ads will appear where. This can be combined with various forms of targeting, matching the display of ads to specific situations. OpenX Ad Server can handle targeting based on date, day and time, geographic area (country or city, and more), Web browser used by the visitor, and language set by the visitor, to name just a few.

The next section of this topic will discuss more on how does the author use OpenX Ad Server to create an ads matching system.

4.4.1   Advertisers and Campaigns

Author also created advertisers, campaigns and ads in the OpenX Ad Server according to the campaign demand. In OpenX Ad Server, an advertiser has a name, a contact and their e-mail address. Any advertiser can have one or more campaigns. A campaign is a collection of ads that have several properties in common.

Assume that author is creating a campaign for advertiser – Fotokem:

Figure 16: Creating campaign in OpenX Ad Server

In OpenX Ad Server, a campaign has:

- a name
- a type: 'Remnant', 'Contract' or 'Contract Exclusive'
    - Remnant: The word 'remnant' means "leftover, a small part or portion that remains". The most common scenario is that a site owner has one or more remnant campaigns that can be displayed at any time.
    - Contract: Contract campaigns can be used for many scenarios, but most often it is a paid campaign from what is generally referred to as a 'direct advertiser'.
    - Contract Exclusive: Exclusive campaigns are useful in scenarios where multiple Contract campaigns are fighting for the available traffic, possibly even exhausting it. Then an Exclusive campaign is going to be displayed with preference to all other campaigns.
- a start and end date (optional)
- pricing information based on a selected pricing model
- a priority or weight (remnant and exclusive campaigns have weights, contract campaigns have priorities, more on this subject in a later chapter)
- capping (a way to limit the number of ad views per site visitor, more on this in a later chapter)
- companion positioning (optional)

In this project simulation, author used **contract** campaign type. By default, campaigns have no start and end date. They will start delivering (when linked to zones) straight away after they have been created, and they will not expire. When a campaign has a start date in the future, OpenX Ad Server will automatically activate that campaign at midnight of that date. When a campaign has an end date, the campaign will continue to run until midnight of the date. So both the start date and the end date are inclusive.

OpenX Ad Server can report on ad revenues based on several different pricing models. The pricing models available are:

- CPM (cost per mille)
- CPC (cost per click)
- CPA (cost per action)
- Monthly tenancy (A Monthly tenancy campaign has a pricing model where the advertising is paying a fixed amount of money per month, regardless of the number of ad impressions, clicks or conversions)

In all of the 'cost per …' types, the word 'cost' refers to what it costs the advertiser to be running the ads on the sites. These are standard terms from the online advertising industry. In this campaign, CPC is chose as author would like to monitor the amount of click generated by this campaign. Any pricing models will do for any campaigns, depending on the agreement between publishers, advertisers and the ad exchange.

In OpenX Ad Server they also have a weight. The three campaign types (Remnant, Contract and Exclusive) have their own characteristics when it comes to interacting with other campaigns. These interactions are based on campaign weight and priority. Weight is entered as a number between 0 and 127. The specific value is not important. It is the value in comparison to the weight of other campaigns that is significant. Setting a campaign's weight at 0 (zero) will halt that campaign immediately, a very effective way of (temporarily) disabling a campaign.

Next, delivery capping is a way of limiting how often the ads from a campaign can be displayed for every single visitor of the sites. It is possible to specify a maximum number of campaign views per visitor for the life time of the campaign, or just for the length of a single session. Capping works by recording a count of the campaign's impressions in a cookie on the user's computer. If OpenX Ad Server finds that the browser does not accept cookies, then campaigns with capping will not be delivered at all, to be on the safe side.

Capping can also be specified at the banner level, and every banner can even have its own capping specified. When capping is set for a campaign, entering capping for the banner overrules it. Example: campaign capping set to 5 but banner capping set to 3 means that the specific banner can only be delivered three times, even if the campaign's capping hasn't been exhausted yet.

A campaign can be linked to a zone (or to multiple zones). A zone will automatically only show the ads that are part of the campaign and that have the exact same dimensions (width and height) as the zone. The next section will elaborate more on zones and linking zones with Websites.

4.4.2 Zones and linking zones with Websites

In the OpenX Ad Server, the author can define Websites and also one or more zones for each Website. A zone represents a space on the Web pages where ads are supposed to be displayed. For every zone there is a little snippet of HTML code, which must be placed in the site, at the exact spot where the zone should go. This forms the integration between the site and the OpenX Ad Server. It is a one time job that will take little time for an experienced Webmaster or developer.

The word 'zone' is a term in OpenX, other systems refer to it as 'location', 'spot', 'placement' or 'position'. A zone in OpenX Ad Server has a few properties: name, description and size (width and height). And after creation, every zone gets a unique ID. There are 4 types of zones:

- Banner – for displaying ads on a Web page
- Interstitial/floating DHTML – for displaying an ad in a 'layer' that's floating in front of the actual Web page
- Text ad – for displaying ads in a paragraph of text on a Web page, looking identical to any other links
- Email/newsletter – a special version of the 'banner' type for use in e-mail newsletters

Hence in this project, author used the first type of zones, which is **banner** to display related ads on Websites. In OpenX Ad Server these are called 'banners' which shows the age of the software because back when phpAdsnew was develop that was what online adverting was all about. An ad is a creative that is part of one campaign.

There are 5 different types of ads:

- Local banner (stored in the Web server's file system)
- Local banner (stored in the ad server's database)
- External banner (hosted by the advertiser, agency or ad network)
- Generic HTML banner (an ad that's actually a snippet of HTML code that, when delivered to the visitor's Web browser, will be rendered to show the ad)
- Generic Text banner

All ads have in common that they contain an element that will be visible on the Website and a destination URL. The Website visitor will be taken to that URL when he/she clicks the ad. In the online advertising business, this destination URL is also referred to as the landing page.

In this project, the author mainly will use **local banners – banners that are stored in the ad server's database** as the author's perception is to create a middle party between publishers and advertisers. Thus, an ad exchange party needs to have its own ad server in order to hold in all the ad inventories from both publishers and advertisers and serve both parties well in terms of delivering relevant ads to targeted customers.

Furthermore, ads generally also have a specific size (width and height). Figure 13 on the next page shows the standard sizes of ad banner (units are in pixel).
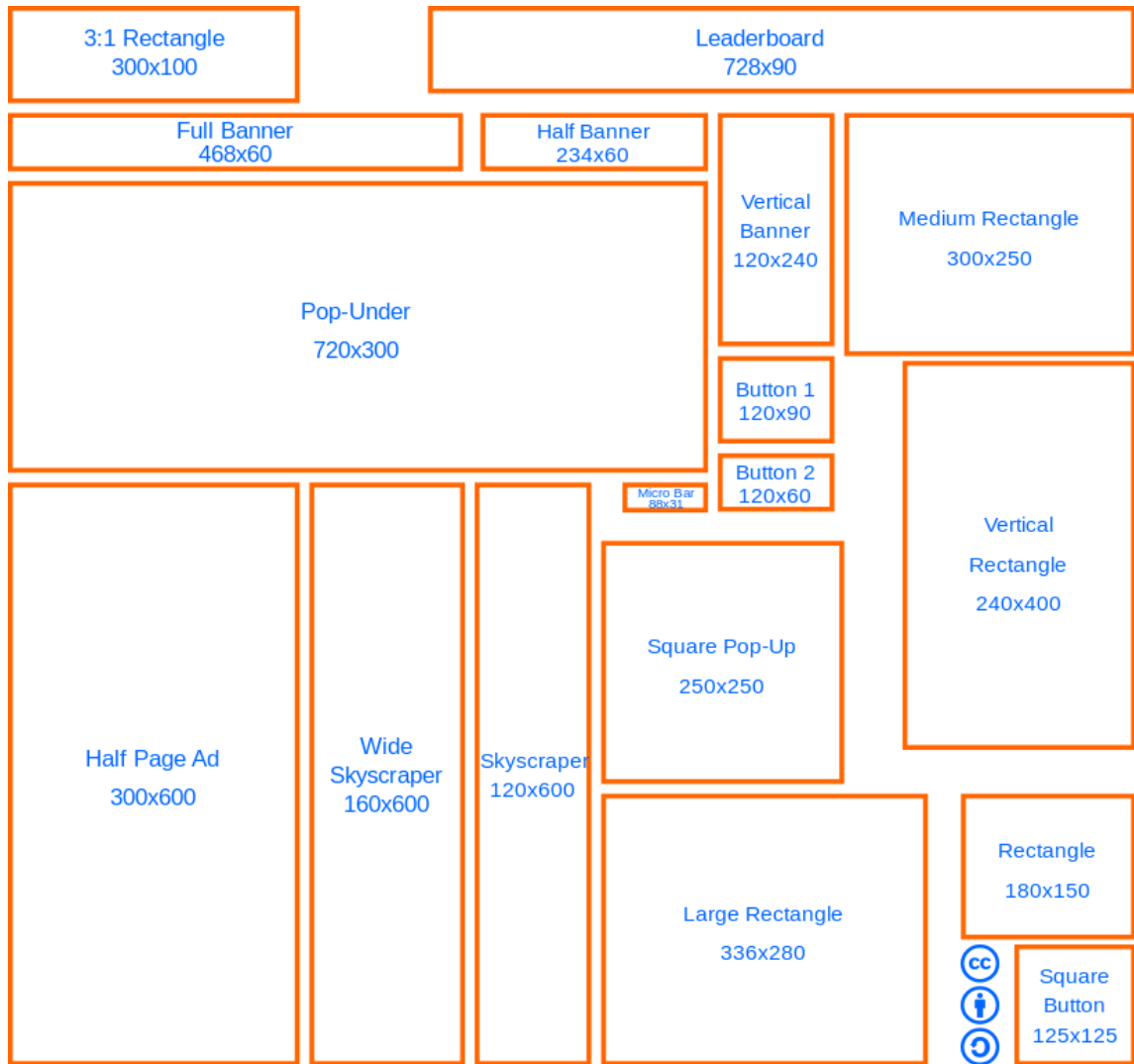
Figure 17: Standard Sizes of Ad Banners (units in pixel)

In this project, the author decided to use only a few popular banner sizes in creating banners for advertisers, which nicely compatible to the design of the Websites. The standard sizes chose are Leaderboard (728x90), Square Button (125x125), Vertical Rectangle (240x400), and Full Banner (468x60). Figures on the next page show a few banners prototype designed by the author. All banners are designed in .gif format, as OpenX can take up various forms of image formats, besides making ads looks more attractive with animations. These banners will then be displayed on the publisher's Websites, according to the designated campaigns, that have been set earlier by the author that acts as a middleman between publishers and advertisers.
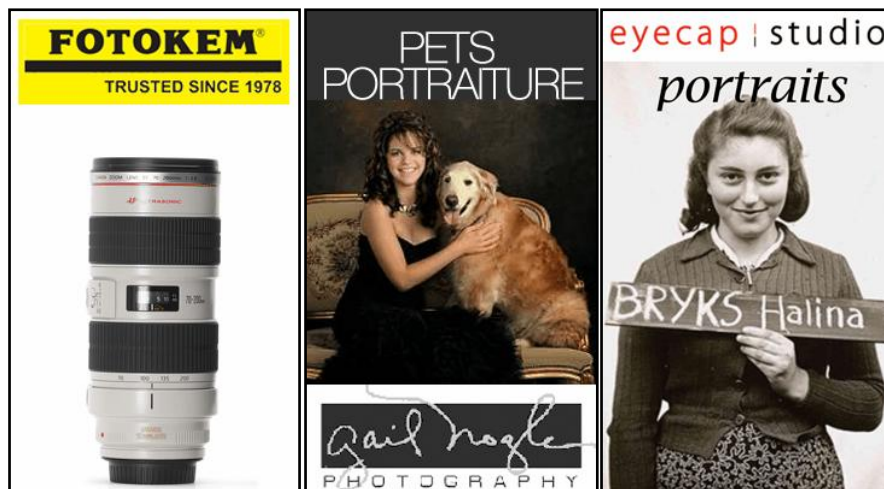
- Leaderboard (728x90)

- Square Button (125x125)

- Vertical Rectangle (240x400)

- Full Banner (468x60)

Once the author has created some advertisers, campaigns and ads, and also some Websites and zones, the final step is to create the links between the campaigns and the zones. That way, the author can determine that the ads of a campaign should be shown in certain zones.

When multiple campaigns are linked to a single zone, OpenX Ad Server assigns a probability percentage to each of the banners in those campaigns, but only for those banners that have the exact same dimensions as the zone. The probability percentage determines how likely it is that an ad is delivered when the zone is called.

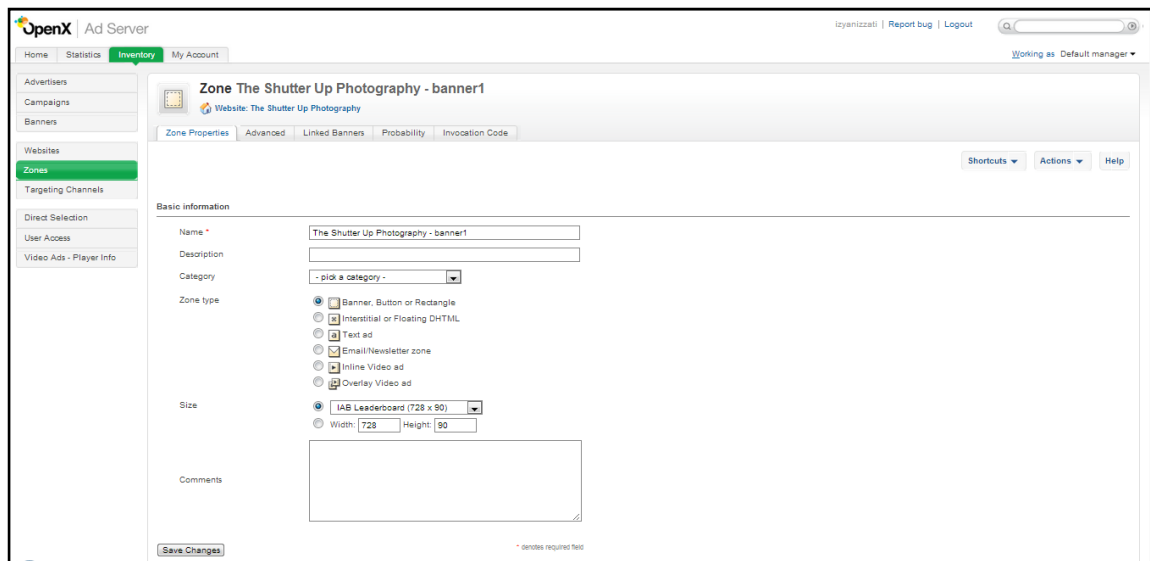Assume that author is creating zones for The Shutter Up Photography Website:

Figure 18: Adding zones in OpenX Ad Server

To add a zone,

- Name: zone should have a unique name. If this field is left blank, OpenX gives it the name of its parent website and adds the suffix "-Default"
- Description: enter information to make it easier to see what the zone is used for, e.g. 'full banner (468 x 60)'
- Zone type: select a zone type
- Size: specify the size of all zones, apart from text ad zones. Only banners that fit these dimensions will be displayed in the zone. This dropdown includes all standard dimensions as defined by the IAB
- Media/Technology Cost: The costs incurred to show ads in the zone
- Comments: this field is provided for you to make notes relevant to this zone

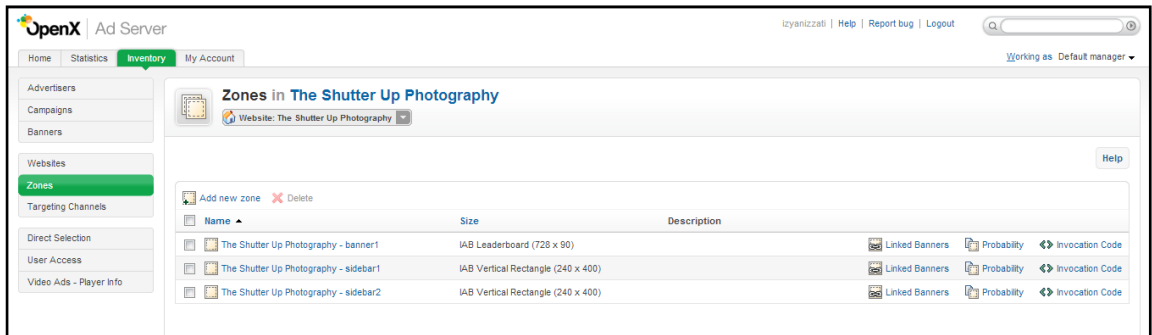The author created three zones for The Shutter Up Photography Website, as per illustrated below:

Figure 19: Zones created for The Shutter Up Photography Website

Next, the Zones > Linked Banners allows author to control which banners author would like to link to a particular zone (with reference to the agreement with publishers and advertisers):
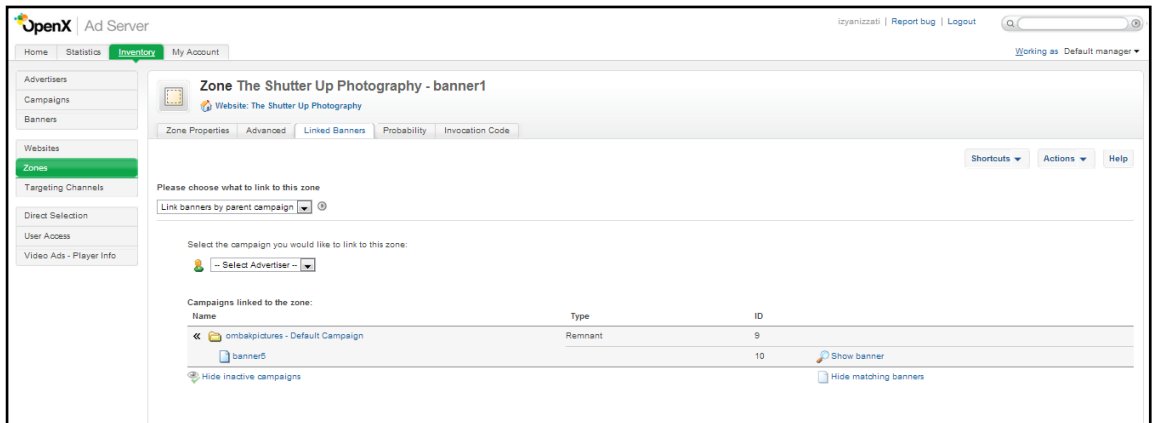

Figure 20: Linking zones in OpenX Ad Server

The first dropdown gives author two choices:

- Link banners by parent campaign: select this option to link all banners associated with a particular campaign to this zone
- Link individual banners: select this option to assign individual banners to this zone

The second dropdown allows author to choose from the advertisers that have banners compatible with this zone in their campaign. Once author have made her selection further dropdowns appear listing the relevant campaigns and individual banners. To finalize the choice, author then click the arrow to the right of the final dropdown:
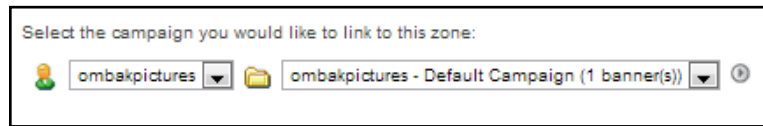
The linked banner and its parent campaign/advertiser will then be listed at the bottom of the page:

Now, it is time for action – serving the banner on the website. Once advertisers, campaigns and zones are created, it is time to link zones with Websites. Assume that the author is linking campaigns of ombakpictures (the advertiser) with The Shutter Up Photography Website (publisher). In addition to linking campaigns to zones, the author can also link an individual ad to a zone. Consequently, it is also possible to unlink one ad from a zone even thought the campaign that the ad is part of is linked to that zone. The ability to link individual ads rather than whole campaigns provides extra flexibility, but it comes at the price of requiring more work to set everything up.

The Zones > Invocation Code tab is where ad tags are generated. These tags are then pasted into HTML pages which invoke (call) OpenX to deliver ads. The types of tags available depend on the type of zone you have defined. The tag types (listed in order of popularity) are:

- JavaScript Tag: the most popular and versatile tag type (includes a noscript option for the very small number of users that are not JavaScript-enabled).
- Local Mode Tag: enables server-side processing for banners that are stored locally, e.g. to change the site template when a user clicks on an ad.

- This tag works only when OpenX and the site calling the code are on the same physical machine. The website must be written in PHP, or be able to parse PHP.

- XML-RPC Tag: similar to the local mode tag, but can be used for ads stored on a remote server.

- iFrame Tag: this tag permits an ad to be imbedded in an iFrame on a web page.

- Image Tag: useful if you want to display banners to an audience that is not JavaScript-enabled. Width and height attributes are not defined in this tag.

- If you want to allocate space for an ad before it is displayed then this information must be added directly into the tag.

- No Cookie Image Tag: this image tag does not use cookies. This tag is used by email zone types because most email servers do not support the use of cookies.

Now, we are on the invocation code settings page. Notice that the Javascript Tag option and invocation Bannercode is selected by the author:

Figure 21: Bannercode generated in OpenX Ad Server

When one select the tag type, one can set various optional parameters in the Tag settings section of the page which change the banner code for the zone. These parameters vary depending on the type of zone and the type of zone tag one is using.

OpenX Ad Server generates snippets of html or JavaScript code, called 'invocation code' in OpenX. These snippets of code, also referred to as 'tags', are pasted into the template, which enables the pages that use these templates to show zones and the campaigns and ads linked to those zones. Hence, after that, the author would need to paste this code on the section where advertisers want the advertisements to appear (with the agreement of publishers). In this project, the banner code is pasted on the sidebar widget (titled: Sponsors) in the Wordpress admin page:

Figure 22: Bannercode pasted in Wordpress widget

After the banner code has been saved in Wordpress, one would wait for around 10 minutes, if one doesn't see the advertisement served there immediately. Then refresh the website again. Finally, the advertisement should appear there, as assigned from OpenX Ad Server. The delivery of ads can be limited by various settings, which allows the user to target the ad to a specific location, audience, or condition. In almost all cases, zones are part of the overall site layout that's being implemented through some kind of page template.

Thus per illustrated below, Ads displayed on the web page are pulled from OpenX Ad Server that match the dynamic Website contents as per assigned. When a user clicks on these ads, they will be redirected to advertisers' landing page automatically.

Figure 23: Zones in The Shutter Up Photography Website

Notice that all ads that appeared on the Website are related to the Website content.

As for instance, the first zone (Leaderboard) shows ad from advertiser ombakpictures.com. The ad highlights the portrait photography service it offers at ombackpictures.com. This is related to the Website content since The Shutter Up Photography is also doing portraitures photography service. The same thing applies for ads in zone 2 and zone 3. Both ads are offering photography service.

When a user visits The Shutter Up Photography Website, he/she also being served with ads that are related to the Website. Thus the likeability of the user to click on the served ads will be high, as it matches their preferences. Plus, when ads are matched to user's preferences, the user will perceive ads with high value and not easily being ignored.

Furthermore, the same mechanism is used for the other two websites – The Picstop and Bridalsnaps. Both figures below illustrate the ad zones. All ads that are displayed mainly related to the website content. For The Picstop Website, all zones advertised camera products and accessories.

Figure 24: Zones in The Picstop Website

As for Bridalsnaps, all zones are set to advertise ads that are mainly related to wedding, such as wedding planning, wedding photography and bridal products and accessories. Zones specified are as per below figure:
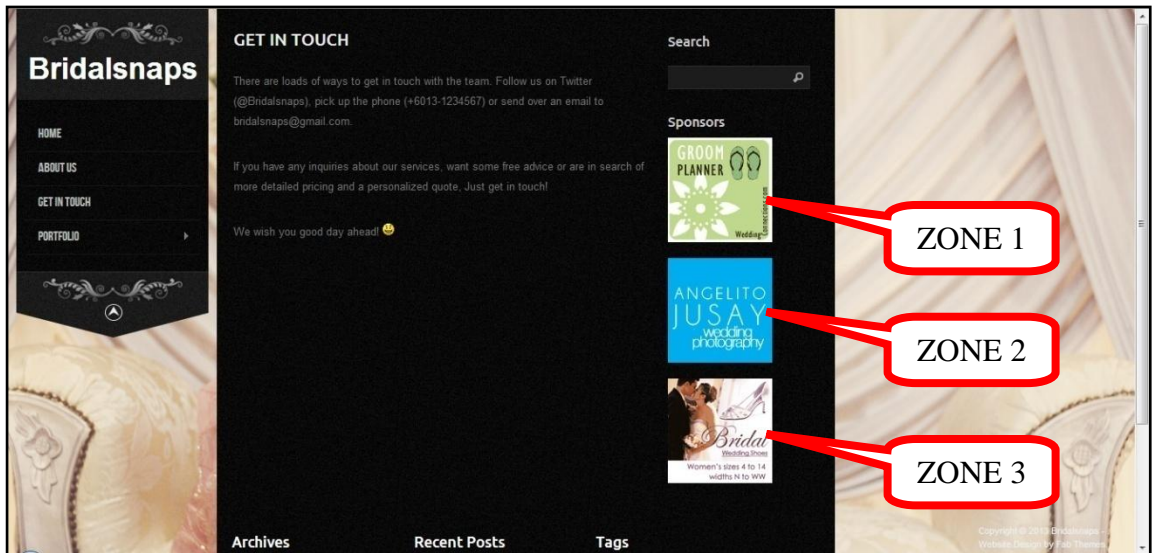

Figure 25: Zones in Bridalsnaps Website

Many sites also have different templates for different sections, like the home page, the news pages and the forum pages. Whatever the case, the zones are part of the template, and as a result will automatically appear on all the site's pages using that template.

4.4.3   Targeting: the right ad at the right spot

OpenX Ad Server is capable of targeting ads to specific sections of a Website, using a feature called the 'source parameter'. Source is a way of 'labeling' a zone. The word entered in the source field is used to determine which ads will be displayed. Targeting is a great way to increase ad space revenue. There are also advanced targeting options, using known information about members or customers (like their gender, education or age range), or from known details about the site's content and structure (for instance: display the ad only in the 'news' section or anywhere but the 'sports' pages).

More to the point, in the online advertising industry, targeting refers to the practice where an ad is not just displayed anywhere and everywhere, but only when specific conditions occur. Since this limits the situation when the ad can be delivered, OpenX Ad Server calls this 'delivery limitations'. OpenX Ad Server presents the available types of delivery limitations in a list:
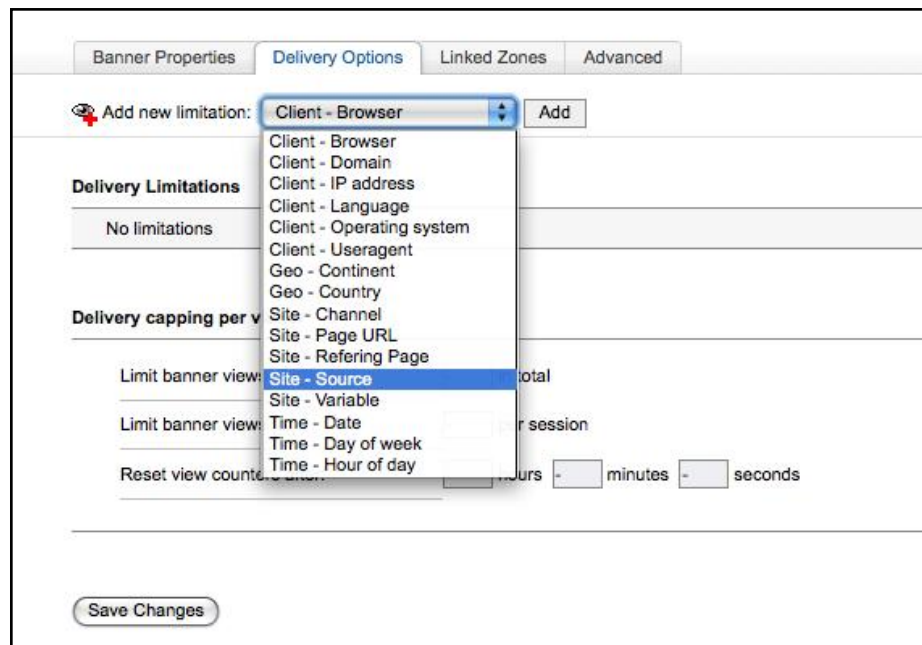
Figure 26: Types of delivery limitations

There are four different sets of targeting mechanisms:

  i. *Client (computer and browser):*

    The word 'Client' is a technical term from internet technology that refers to everything that has to do with the computer and the web browser that a person is using when visiting a site.

     • Client – Domain: using this option the ad can be delivered to just those users browsing the internet from a specific domain (for instance: only when the visitor uses a computer that is connected through the example.com domain name). Note: this is not the same as the domain name of the website!

     • Client – IP Address: enables the advertising manager to show ads only to specific IP addresses.

- Client – Language: every browser sends the preferred language as part of the request for a page. Ads can be targeted to the language preferred language set in the browser.
- Client – User agent: a browser used to visit a website identifies itself, by make and version. Common user agents are: MSIE for Microsoft Internet Explorer , Safari for Apple's Safari browser and Mozilla for Firefox.

ii. *Geographic targeting:*

The world 'Geo' refers to the type of delivery limitations where the geographic location of the person visiting the site is used for targeting. Depending on the geo-targeting features available and enabled in the OpenX Ad Server, targeting by country, by city, by continent, region, province, state or DMA can be available.

iii. *Site targeting:*

The set of delivery limitations grouped as 'Site' are used to respond to specific information coming from or about the website where the ads are displaying.

- Site – Channel: see the next paragraph for a description of channels, also known as 'targeting channels'.
- Site – Page URL: this delivery limitation can be used to restrict ad delivery to a page or page that has a specific word in its URL, or perhaps matches an exact URL like the website's homepage.
- Site – Source and Site – Variable are delivery limitations that respond to information available in the invocation code on the website.

iv.   *Time targeting:*

The delivery limitations grouped as 'Time' can be used to restrict ad delivery based on time and date settings.

- Time – Date: an ad will only be delivered is a date condition is mate.

- Time – Day of week: the ad will only be delivered on specific days of the week

- Time – Hour of day: ad delivery is restricted based on the hour of the day, with each of the 24 hours available separately.

Note that in combining delivery limitations, each ad can have multiple delivery limitations, and they can be combined using the AND and OR operators. Using two delivery limitations with the 'AND' operator means that both conditions will have to be met for the ad to be delivered. The 'OR' parameter, on the other hand, means that just one of the two conditions has to be met.

4.4.4   Statistics: to measure is to manage

OpenX Ad Server continuously records every single ad impression and ad click. The data it collects is kept in back end of the database, and is not visible through the OpenX Ad Server user interface. The OpenX developers refer to this data as 'raw data', because it has not been summarized into the on-screen statistics. In most cases, there is more than one zone on any single Web site, so viewing a Web page will cause multiple ad impressions being logged.

As per instance in this project, the statistics shows how the campaigns are doing based on the number of impressions, clicks, CTR, revenue generated and ECPM:
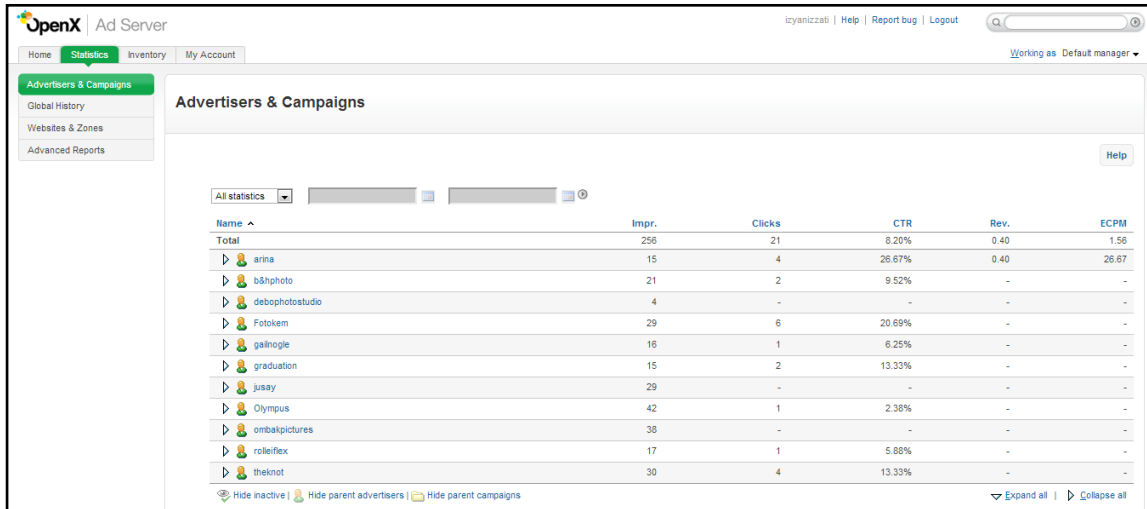

Figure 27: Campaign statistics generated by OpenX Ad Server

The logging of raw data is called 'bucket logging'. Raw data is being recorded for every combination of an ad and a zone, and the software keeps a running total. The first impression of an ad in a zone creates a record with an impression count of 1. The second impression increases that count to 2, and so on.

The raw data is logged in buckets for every hour, per ad/zone combination. At the top of the new hour, OpenX Ad Server starts with new buckets. In addition to being able to have buckets by ad/zone, the system can also have additional buckets with more levels of detail. This can be achieved by developing a new plug-in.

Thus in a nutshell, OpenX Ad Server is written in the PHP programming language and stores data using the MySQL database, both of which are also open source software. Author has Apache HTTP Server (that comes with MYSQL database) installed in building the Websites and integrating them with OpenX Ad Server. Except for time and hardware, there are no additional costs, and no license fees. The system has been designed to enable installation on a cluster of Web servers, allowing it to grow alongside the growth of the sites.

CHAPTER 5
CONCLUSION AND RECOMMENDATION

In conclusion, this research project is focusing on to successfully make the best Contextual Ads selections that match to the Web Page contents through the concept of Computational advertising. Furthermore, this research project also focuses on to ensure there is a valuable connection between the Web pages and the Contextual Ads during the selection process.

It is believed that Internet advertising has emerged as the potential channel for e-commerce companies to reach their target customers in marketing and promoting their products. Furthermore, addressing the issues raised in this project is very essentials to make the best Contextual Ads selections that match to the Web Page contents of the business so that ads are more likely to be clicked on by the users. More to the point, computational advertising theory could be applied in creating more effective Internet advertisements in online advertising that are more targeted and best-matched with the Website contents.

For recommendations, future work in this area needs to be done in order to refine the potential of establishing Computational advertising, as it is believed that this area of business is likely to propel in the local market in the near future. Besides, future analysis and research on balancing goals for online advertising market players need to be in continuation to enhance the credibility and explore the possibilities of Internet advertising, to be implemented in the local scene. It would be very great if future works will explore more features on making content matching more personalized compared to what is done in this project.

Also, it is suggested that this area of research will take a migration towards behavioral targeted advertising as it will likely boost up online advertising performance. Behavioral Targeting refers to a range of technologies and techniques used by online Website publishers and advertisers which allows them to increase the effectiveness of their campaigns by capturing data generated by Website and landing page visitors. This area of research needs more attention on consumers'' behavior analysis, Human-Computer Interaction (HCI) and consumers' privacy.

Thus, this research is another contribution to the Advertising Industry itself, to be exploitable and acceptable by the local community and contributes to bringing forward the potential of Internet advertising in Malaysia. Hence, the author really hopes that the project could be expand further in term of technology used and also can be realize as one of problem-solving approach in optimizing Internet advertising . Lastly, hopefully this project could open more opportunity towards all organizations and individuals, especially small and medium business entities to get involved in progressing their businesses

REFERENCES

− Leo Burnett. (n.d.). *BrainyQuote.com*. Retrieved October 24, 2012, from BrainyQuote.com Web site: http://www.brainyquote.com/quotes/quotes/l/leoburnett103284.html

− Xian-Sheng Hua; Tao Mei; Shipeng Li; , "*When multimedia advertising meets the new Internet era,*" Multimedia Signal Processing, (2008) IEEE 10th Workshop on , vol., no., pp.1-5, 8-10 Oct. 2008 doi: 10.1109/MMSP.2008.4665039 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4665039&isnumber= 4665023

− Yu Hu; Jiwoong Shin; Zhulei Tang; , "*Pricing of Online Advertising: Cost-Per-Click-Through Vs. Cost-Per-Action,*" System Sciences (HICSS), (2010) 43rd Hawaii International Conference on , vol., no., pp.1-9, 5-8 Jan. 2010 doi: 10.1109/HICSS.2010.470 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5428411&isnumber= 5428274

− Interactive Advertising Bureau. (n.d.) *IAB Internet advertising Revenue Report conducted by PricewaterhouseCoopers (PWC)*. Retrieved October 25, 2012, from http://www.iab.net/insights_research/industry_data_and_landscape/adrevenuerep ort

− Jansen, B.J., Mullen. T., "Sponsored search: an overview of the concept, history, and technology," Int. J. Electronic Business, Vol. 6, No. 2, 2008. URL: http://tmullen.ist.psu.edu/pubs/overview_sponsored_search.pdf

− Wei Li, Xuerui Wang, Ruofei Zhang, Ying Cui, Jianchang Mao, Rong Jin. (2010). *Exploitation and Exploration in a Performance Based Contextual Advertising System*. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10). ACM, New York, NY, USA, 27-36. DOI=10.1145/1835804.1835811 http://doi.acm.org/10.1145/1835804.1835811

– Shuai Yuan, Ahmad Zainal Abidin, Marc Sloan, Jun Wang. (2012). *Internet advertising: An Interplay among Advertisers, Online Publishers, Ad Exchanges and Web Users*. London: University College London. Retrieved October 26, 2012,from http://arxiv.org/pdf/1206.1754v2.pdf

– Andrei Z. Broder. (2008).*Computational advertising and Recommender Systems*. Retrieved on October 24, 2012 from http://delivery.acm.org/10.1145/1460000/1454009/p1-broder.pdf?ip=203.135.190.8&acc=ACTIVE%20SERVICE&CFID=134707582&CFTOKEN=13369803&__acm__=1351716028_04733dfb8b852cc306c79fadb95f3267

– Yahoo! Research Labs (n.d.). In *Key Scientific Challenges: Computational advertising.* Retrieved on October 25, 2012 from http://labs.yahoo.com/ksc/Computational_Advertising

– McBride, R., & Schostak, J. (n.d.). *Quantitative Versus Qualitative Research.* Retrieved from http://www.enquirylearning.net

– Zinkhan G.M., Watson, R.T. (1996). *Advertising Trends: Innovation of Creative Destruction and the Process*. Retrived on October 25, 2012 from http://is.gd/nKUbeY

– Boone, G., Secci, J. Gallant, (n.d.). *Emerging Trends in Online Advertising*. Retrieved on October, 25, 2012 from http://www.doxacomunicacion.es/pdf/artculoboonesecciygallant.pdf

– Evans, D.S., (2009). *The Online Advertising Industry: Economics, Evolution, and Privacy.* Retrieved on October 25, 2012 from http://www.intertic.org/Policy%20Papers/EvansEOAI.pdf

– Dave, K. (2011). *Computational advertising: Leveraging User Interaction & Contextual Factors for Improved Ad Retrieval & Ranking. Hyderabad, India*. Retrieved on October 25, 2012 from http://wwwconference.org/www2011/proceeding/companion/p349.pdf

– Evans, D.S., (2008, January). *THE ECONOMICS OF THE ONLINE ADVERTISING INDUSTRY*. Retrieved on October 25, 2012 from http://www.intertic.org/Policy%20Papers/Evans.pdf

- Broder, A., et. al. (n.d.). *A Semantic Approach to Contextual Advertising*. Retrieved on October 25, 2012 from http://fontoura.org/papers/semsyn.pdf

- Mutukrishnan, S. (2009). *AdX: A Model for Ad Exchanges*. Google Inc. Retrieved on October 25, 2012 from http://is.gd/R7IjMh

- Haque, A., et. al. (2007). *Internet Advertisement in Malaysia: A Study of Attitudinal Differences.* Retrieved December 8, 2012 from http://www.ejisdc.org/ojs2/index.php/ejisdc/article/viewFile/233/208

- Li, W., et. Al (n.d.). *Exploitation and Exploration in a Performance based Contextual Advertising System.* Retrieved on December 12, 2012 from http://www.cs.cmu.edu/~xuerui/papers/ee.pdf

- Ciaramita, M., et. al. (2008*). Semantic Associations for Contextual Advertising*. Retrieved on December 12, 2012 from http://goo.gl/n8IoU

- Fan, T.K., Chang C.H., (n.d.). *Sentiment-Oriented Contextual Advertising.* Retrieved on December 12, 2012 from http://in2.csie.ncu.edu.tw/~chia/pub/SOCA.pdf

- E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. (1999). *Domain-specific keyphrase extraction.* Retrieved on December 12, 2012 from http://ijcai.org/Past%20Proceedings/IJCAI-99%20VOL-2/PDF/002.pdf

- Shengwen Yang, Jianming Jin, Parag Joshi, Sam Liu. (2010). *Contextual Advertising for Web Article Printing*. Retrieved on December 12, 2012 from http://www.hpl.hp.com/techreports/2010/HPL-2010-79.pdf

- Anagnostopoulos A. et al. (n.d.). *Web-Page Summarization for Just-in-Time Contextual Advertising.* Retrieved on January 05, 2013 from http://cs.brown.edu/~aris/pubs/jit-journal.pdf

- Eibe Frank, Gordon W. Paynter, Ian H. Witten. (2005). *Impedance coupling in content-targeted advertising.* Retrieved on January 05, 2013 from http://dl.acm.org/citation.cfm?id=1076119

- C. Wang, P. Zhang, R. Choi, and M. Daeredita. (2002). *Understanding consumers attitude toward advertising.* Retrieved on January 05, 2013 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.8755

- Wu Z. et al. (2009). *Leveraging Wikipedia Concept and Category Information to Enhance Contextual Advertising.* Retrieved on January 05, 2013 from http://vbn.aau.dk/ws/files/59464685/Leveraging_Wikipedia_Concept_and_Category_Information_to_Enhance_Contextual_Advertising.pdf

- Langheinrich et. al. (1999). *Unintrusive customization techniques for Web advertising.* Retrieved on January 05, 2013 from http://www.ra.ethz.ch/cdstore/www8/data/2159/pdf/pd1.pdf

- Teng-Kai Fan and Chia-Hui Chang. (2009). *Sentiment-Oriented Contextual Advertising.* Retrieved on January 05, 2013 from http://in1.csie.ncu.edu.tw/~chia/pub/SOCA.pdf

- Wen-tau Yih, Joshua Goodman, Vitor R. Carvalho. (2006). *Finding Advertising Keywords on Web Pages.* Retrieved on January 05, 2013 from http://research.microsoft.com/pubs/73692/YihGoCa-www06.pdf

- Feng Zhang, Guang Qiu, Jiajun Bu, Mingcheng Qu, Chun Chen. (2009). *A Novel Approach to Keyword Extraction for Contextual Advertising.* Retrieved on January 05, 2013 from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5175966&tag=1

- Xiaoyuan Wu, Alvaro Bolivar. (2008). *Keyword Extraction for Contextual Advertisement*. Retrieved on January 05, 2013 from http://wwwconference.org/www2008/papers/pdf/p1195-wu.pdf

- Alexander Pak, Chin-Wan Chung. (n.d.). *A Wikipedia Matching Approach to Contextual Advertising.* Retrieved on January 05, 2013 from http://islab.kaist.ac.kr/chungcw/interJournal_papers/AlexPak.pdf

- Asmita Joshi, J.S.Sodhi, Roopali Goel. (2013). *Contextual Advertising through Entity Extraction.* Retrieved on February 25, 2013 from http://www.ijeat.org/attachments/File/v2i3/C1220022313.pdf

- Lang et. al. (2011). *Efficient Online Ad Serving in a Display Advertising Exchange.* Retrieved on February 25, 2013 from http://research.yahoo.com/pub/3445

- *OpenX Ad Server: Beginner Guide.* Retrieved on February 25, 2013 from http://goo.gl/mlfc7

APPENDICES

Appendix A: Survey Questionnaires

Internet advertising - What say you?

The questions below deal with the topic of Internet advertising. Please take a moment to read the questions and complete the survey. There are no right or wrong answers and your identity is anonymous. This survey will take about 5 minutes to complete. You must complete all the questions in order to submit your answers.

1. What is your gender? *
   - ○ Male
   - ○ Female

2. What age group below you belong to? *
   - ○ less than 18
   - ○ 18 -25
   - ○ 26-35
   - ○ 36-50
   - ○ over 50

3. What is your education level? *
   - ○ High school
   - ○ Some college
   - ○ Diploma
   - ○ Degree
   - ○ Masters
   - ○ PHD

4. What is your level of income monthly? *
   - ○ RM1000 and below
   - ○ RM1001 - RM2000
   - ○ RM2001 - RM3000
   - ○ RM3001 - RM4000
   - ○ RM4001 - RM5000
   - ○ More than RM5001

5. How often do you go online and use the Internet? *
   - ○ Less than once a week
   - ○ About once a week
   - ○ 2 to 4 times a week
   - ○ 5 to 6 times a week
   - ○ Everyday

6. When using the Internet, about how long do you stay logged on (consider the whole day)?*
   - ○ Less than 1 hour
   - ○ At least 1 but less than 2 hours
   - ○ At least 2 but less than 5 hours
   - ○ 5 or more hours

7. What is your most favorite general search engine? *
   - ○ Google
   - ○ Bing
   - ○ Yahoo!
   - ○ Volunia
   - ○ Other: [＿＿＿＿＿]

8. Have you ever heard of Internet advertising? *
   - ○ Yes
   - ○ No

9. Have you ever clicked on an online advertisement (banner pop-up, online ad on your search engine/blog/Website)? *
   - ○ Yes
   - ○ No

10. Rate your agreeableness on these statements: *

		Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
i.	When you see an online ad, you do not click on it and avoid looking at it because it is annoying/offensive	○	○	○	○	○
ii.	When you see an online ad, you click on it to seek more information	○	○	○	○	○
iii.	When you see an online ad, you click on it because it matches your interest and preferences	○	○	○	○	○
iv.	When you see an online ad, you click on it because it is very interactive (got videos, sounds)	○	○	○	○	○
v.	You always prefer the classic advertisement mediums (on newspapers, television, radio) rather than online ads	○	○	○	○	○
vi.	You always accidentally clicked the online ads	○	○	○	○	○
vii.	You think the online ads always misled you	○	○	○	○	○
viii.	In general, you feel you can trust online advertisings	○	○	○	○	○
ix.	In general, you feel very confident in using information you see in the ads to make a purchase decision	○	○	○	○	○
x.	You feel comfortable about purchasing an item through an online ads by giving your address or phone number	○	○	○	○	○
xi.	You think products purchased through online ads has much lower prices	○	○	○	○	○
xii.	In general, I like online advertising and I enjoy it	○	○	○	○	○