Decision Support System for Search Engine Advertising Campaign Management by Determining Negative Keywords

Başak Tavşanoğlu

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of the requirements for the degree of
Master of Science

Sabancı University

July, 2018

DECISION SUPPORT SYSTEM FOR SEARCH ENGINE ADVERTISING CAMPAIGN MANAGEMENT BY DETERMINING NEGATIVE KEYWORDS

APPROVED BY:

Assoc. Prof. Dr. Kemal Kılıç ....
(Thesis Supervisor)

Assoc. Prof. Dr. Abdullah Daşcı ....

Assoc. Prof. Dr. Gürdal Ertek ....

Date of Approval: 31.07.2018

Acknowledgements

Foremost I would like to express my appreciation and gratitude to my advisor Dr. Kemal Kılıc for the continuous support of my study and research, for his patience, motivation, and immense knowledge. His encouragement was there all the time even I did not know or decide what to do next. He has been a role model for me both in personal and academic life.

I also want to thank Dr. Abdullah Daşcı and Dr. Gürdal Ertek for accepting to be part of thesis jury and their valuable feedback.

Special thanks to my fellow officemates for all the fun we have had. Especially, Didem and Soner were always there for helping me and motivating me at every step of this journey.

I am greatefull to all of my friends that were beside me during my MSc research, especially Ece, Özlem and Çağla for making our dorm room a cozy home.

Last but not the least, I would like to thank my family for their continuous support and patient.

Arama Motoru Reklam Yönetimi için Negatif Anahtar Kelime Tespiti ile Karar Alma Sistemi

Başak Tavşanoğlu

Endüstri Mühendisliği, Yüksek Lisans Tezi, 2016

Tez Danışmanı: Kemal Kılıç

Anahtar Kelimeler: Arama Motoru Reklamcılığı, Negatif Anahtar Kelimeler, Tıklama Potensiyeli, Makina Öğrenmesi Algoritmaları

Özet

Arama motoru reklam verenleri, kampanyaları için en iyi anahtar kelime grubunu bulmayı hedefler. Her şirketin Arama Motoru Reklamcılığı'ndan (AMR) belirli kısıtlamaları ve beklentileri vardır. Bu tez çalışmasında, AMR kampanya yönetiminde kullanılabilecek bir Karar Destek Sistemi (KDS) geliştirdik. KDS önceki kampanyalardan elde edilen verilere dayanarak, negatif anahtar kelimeleri (performansı artırmak için anahtar kelime kümesinden kaldırılması gereken) belirler. Negatif anahtar kelimeleri belirlemek için kullanılan mevcut metrikler, hemen çıkma oranı, kalite puanı gibi diğer önemli özellikleri içermediğ inden yeterli değildir. Bu özellikler trafiği değerlendirmek için çoğunlukla reklamverenler tarafından kullanılır, ancak genellikle bunlar dönüşüm oranına dayanır.

Araştırmamızda öncelikle, çeşitli makine öğrenme teknikleri kullanılarak kampanyayı olumsuz yönde etkileyen ve / veya olumlu yönde etki eden unigram setini tanımlamak için anahtar sözcükleri (literatürde mevcut olan bazı yaklaşımlara benzer şekilde) unigram düzeyinde analiz edilmektedır. Naïve Bayes, Karar Ağaçları, Lojistik Regresyon (olduğu gibi veyaonlarla ilişkili temel kavramlar) kullanılmıştır. Ayrıca, bu sürecin bir parçası olarak reklamverenler tarafından gerçek hayattaki AMR kampanyalarında kullanılabilecek ve daha fazla yön içeren yeni metrikler de sunulmaktadır. Yaklaşımımızın performansı, büyük bir hızlı tüketim malları üreticisinden elde edilen gerçek hayat verileri üzerinde yapılan deneysel bir analiz ile değerlendirilmektedir.

Decision Support System for Search Engine Advertising Campaign Management by Determining Negative Keywords

Başak Tavşanoğlu

Industrial Engineering, Master's Thesis, 2018

Thesis Supervisor: Kemal Kılıç

Keywords: Search Engine Advertising; Negative Keywords; Online metrics; Click Potential; Machine Learning Techniques

Abstract

Search engine advertisers need to determine the best keyword set for their campaigns. Every company has particular constraints and expectations from the Search Engine Advertising (SEA). In this research we worked on a Decision Support System (DSS) that can be used in SEA campaign management. The DSS determines the negative keywords (which should be eliminated from the keyword set in order to improve the performance) based on the data obtained from the earlier campaigns. Current metrics used to determine the negative keywords are not sufficient/adequate, since they don't incorporate other important aspects such as bounce rate, quality score etc. which are often used by the advertisers in order to evaluate the traffic but rely mostly to conversion rate. In our research first we analyze the keywords at unigram level (similar to some of the existing approaches available in the literature) in order to identify the set of unigrams which are negatively and/or positively effecting the campaign by using various machine learning techniques (either as is or used the core concepts associated with them) such as Naïve Bayes, Decision Trees, Logistic Regression.We further extended these algorithms by incorporating ideas borrowed from Greedy Randomized Adaptive Search (GRASP). We also introduced novel metrics which incorporate more aspects used in real life SEA campaigns by the advertisers as part of this process. The performance of our approach is evaluated with an experimental analysis conducted on real life data obtained from a major FMCG producer.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background Information and History

Internet advertising (or online advertising) is a rapidly growing advertising and marketing platform. It total revenue increased 21.4 % with respect to the previous year in 2016. Since 2016 internet has become the largest advertising media when it is com compared to other U.S. ad- supported media as shown in Figure 1.1(Silverman, 2018).

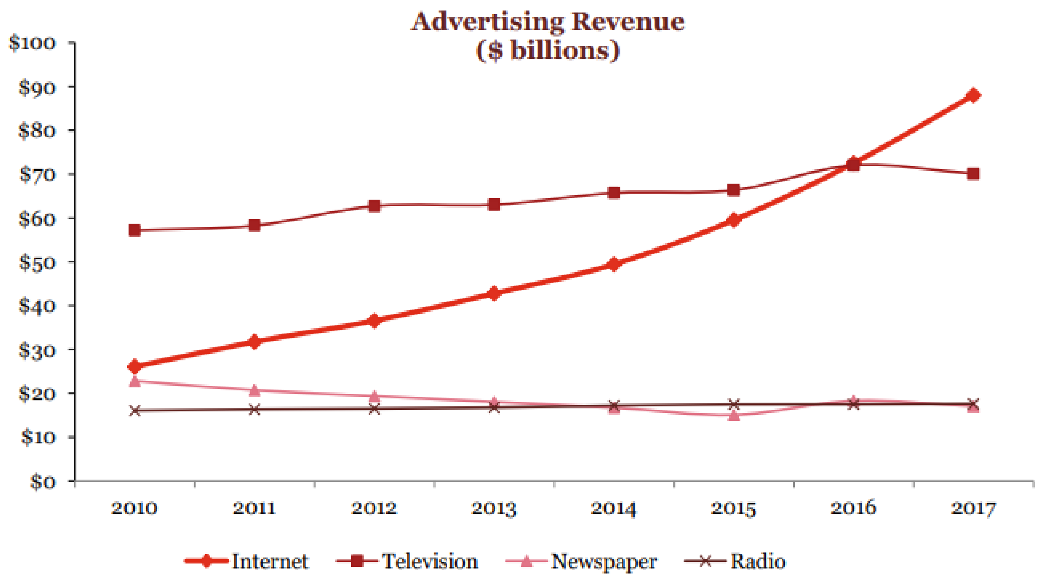

Figure 1.1: Advertising Revenue (Silverman, 2018)

Internet advertising started with the banner ads in 1994. In this platform ads are often priced by the number of impressions delivered, which is known as Cost-per-thousand (CPM) pricing (Fain and Pedersen, 2006). As of now, there are wide variety of advertising formats available in internet advertising such as search engine, banners, mobile etc., among which search engine advertising has the largest share in terms of revenue. Search Engine Advertising (SEA) is referred to as sponsored search, paid search, or keyword advertising in literature (Rangaswamy et al., 2009; Bucklin et al., 2009). It is an ad format that from search engine queries and placed on search engine result page. Among different ad formats search engine has the largest share in 2016 with 48% and also in 2017 it is the greatest by 46 %. Although there is a 2% decrease in the market share, search engine advertising also has a sizable share in internet advertising. As can be seen in the IAB 2017 final report in Figure 1.2.

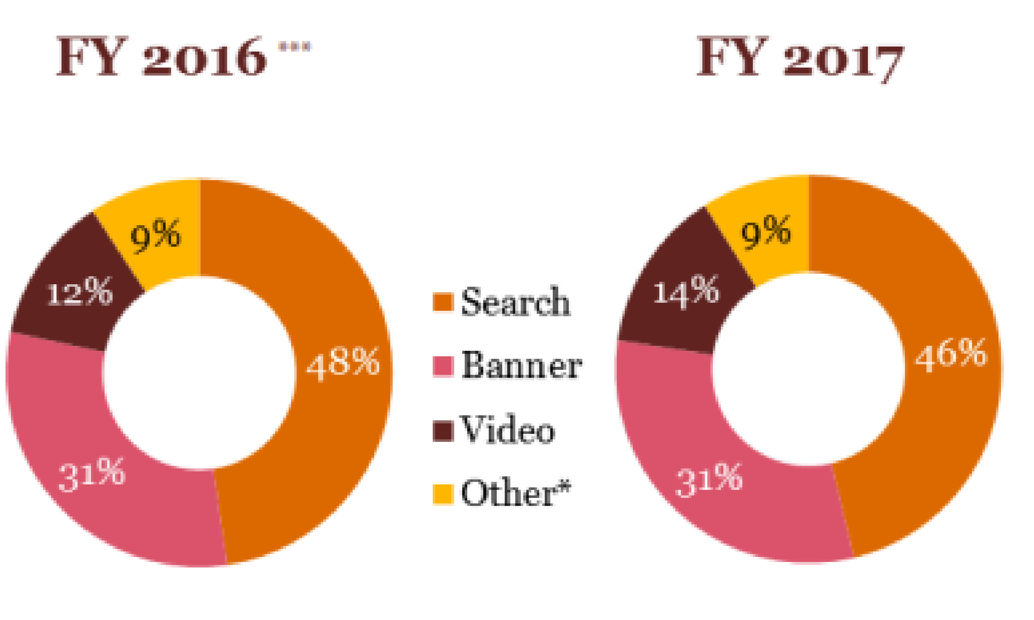

Figure 1.2: Market shares according to ad formats in online advertising (Silverman, 2018)

The process of SEA is continuously evolving and getting complex and sophisticated (Jansen and Mullen, 2008). Briefly speaking, in SEA a text only ad that appear on the Search Engine Result Page (SERP) after customer enters a search query. Advertiser provide a list of keywords which are related to their ads to the search engine providers. These keywords go in to an auction with the competitor advertisers' keywords. According to the auction type search engine provider selects the best one or few of the bests and displays it on the SERP to the customers whose queries matched with the keywords. In 1998 first search engine advertisement was created by GoTo.com which is acquired later by Yahoo in 2003 (Lee, 2009). It was using first prize auction mechanisim, i.e., higher bidder wins, and pays the highest bid. Google come after them in 2002 with generalized second price auction (Jansen and Mullen, 2008), i.e., higher bidder wins and than pays the second highest bid which is also referred as Vickrey auction.

SEA benefits the users, the advertisers, and also generates revenue for the search engine provider. These three actors has different benefits from the process. Firstly the users (i.e. potential customers), find the product that they want to purchase, sometimes with discount options. On the other hand advertisers which is the company that is selling the product or providing the service, reach to their potential customers who are searching for them directly. Also the search engine which is the service provider in the process gain a revenue and provides a useful search experience to their users which improves their market share (Cui et al., 2015).

Search engines play a pivotal role in online shopping and also are the primary medium to promote internet advertising (Mukherjee and Jansen, 2014). 90% of the revenue of the search engines comes from advertisements besides 10% comes from paid inclusion and offline projects. Thus SEA is the primary source of income of search engines (Yang and Ghose, 2010). Therefore much of this online economy is driven by consumer search queries and the advertisements, served by the search engines in response to these queries on SERP which can be quite profitable for online businesses, as they direct consumers to i.e. relevant websites (Mukherjee and Jansen, 2014). According to the literature, 15% of the time the paid ads are clicked and the consumers are ”converted” to the business website who sponsored the clicked paid ad (Jansen and Spink, 2008). Since ad keywords advertisement has high conversion rates one can deduce that they are not annoying the consumers, far less intrusive than online banner ads or pop-up ads (Ghose and Yang, 2009).

1.2 Glossary

In order to understand SEA process one needs to get familiar with the terminology regarding to the metrics used, bidding strategies and match types. In this section we will provide on glossary that will ease the understanding of the reader.

- **Keyword:** The words or the phrases that represents the ad and match with the users' search queries to make the ad visible.
- **Quality Score:**1-10 metric that is provided by the search engine provider to the advertiser for each keyword it is an aggregated estimate of the overall performance of the advertiser with respect to the corresponding keyword in ad auctions. It can not be used directly at auction time to determine Ad Rank (qua, 2017). The details of how Quality Score is calculated usually unknown for the advertiser. However, it is a formulation of relevance between ad, keyword, landing page and user experience. So the components are expected click through rate, ad relevance, and landing page experience. Therefore it is a determinant in position of the ad. Higher quality ads have tendency to have better positions in SERP and lower prices.
- **Ad Relevance:** Provides the relevance between keyword and the ad. It is the information about how closely the searchers query relates to the keywords and how well the keyword matches the message in the ad. It shows the relationship status between advertisement and search query; as above average, average, or below average. (**?**) (Miller, 2016).
- **Landing Page Experience:**Determines how useful the website to the users. Landing page should offer elevant, useful and original content.(Miller, 2016)
- **Click Through Rate (CTR):** Click through rate is a metric that measures the number of clicks advertisers receive on their ads per number of impressions.

- **Conversion Rate (CR):** conversion rate is a metric that measures the number of clicks advertisers receive on their ads per number of conversions.
- **Ad Rank:**Determines the ad position. Ad Rank is calculated using bid amount, the components of Quality Score (expected click through rate, ad relevance, and landing page experience), and the expected impact of extensions and other ad formats.(adr, 2017)
- **Ad Group:** Set of ads that share set of keywords. The ads in the same Ad Group target same set of keywords. It helps to categorize and organize the ads by a common theme.
- **Bounce Rate:** Percentage of the users who click on the ad but immediately leaves the landing page.

Bidding Strategies

- **Cost per Click (CPC):** When CPC is used as bidding strategy, advertiser pays to the search engine provider when the customer clicks on the advertisement. Advertiser offers the maximum amount of bid which is *max. CPC* to the search engine provider. Advertiser usually does not pay this amount because after the auction actual CPC's are calculated. General formula of actual CPC is:

$$\frac{Competitor's AdRank}{Advertiser's QualityScore} + 0.01 = ActualCPC \tag{1.1}$$

- **Cost per Thousand (CPM):**In this bidding strategy advertisers pay when their ads are able to be seen. Number of impression is the most decisive factor.
- **Cost per Acquisition (CPA):**If customer purchase the product after he/she clicks the advertisement and reach to the landing page, advertiser pays to the search engine provider.

Match Types

- **Broad Match:** It is the default match type, user search query matches with the keyword if the query includes any word that is available in the keyword, in any order. This helps to attract more visitors to the advertisers website so this match type reaches the widest audience.

- **Modified Broad Match:** Similar to broad match but advertisers can lock some key phrases by adding '+' sign in front of the keyword and locked them. So this locked terms has to occur in the query to match with the keyword.

- **Phrase Match:** Search engine provider matches the keyword that advertiser list with the query which include the specific keyword or its close variants. There can be additional words before or after the keyword, in the query.

- **Exact Match:** It is the most specific and restrictive match type. The user search query matches with the keyword if the query exactly contains the same phrase or close variants. This match type gives advertiser the most control over the customer so it can provoke an increment in the CTR.

- **Negative Match:** Negative keywords are the ones that the advertiser wants to prohibit. Negative match types (broad, phrase, exact) help the advertiser to disallow the match of these keywords with the customer search query. Using negative match types increase the return on investment as a consequences of better targeting.

1.3 Search Engine Advertising Process

Search Engine Advertising (SEA) process basically start, with the company which wants to advertise. They open an account in the search engine system, i.e. create a *"campaign"*. In the campaign, advertiser provides the keywords, match types, bidding strategy, the market (i.e. the geographical area), the language of the ads going to serve and create its ad groups for each product or service. After this stage; user, (i.e. the customer), takes an action and searches the product or the service with a query through the search engine.

If the query matches with the keywords by the ad selection algorithms of the search engine, the ad presented to the user on the SERP. While matching the keywords and queries, an auction takes place at the background in order to determine the ad which be displayed to the user. This auction is usually the generalized second price auction (Edelman et al., 2007; Varian, 2007). Finally if the user clicks the ad; according to the bidding strategy of the campaign, the advertiser pays a price to the search engine for the click and user directed to the landing page of the ad. Therefore this process also called pay per click advertising (PPC) under search engine advertising. There are different bidding strategies alternative to PPC such as CPM, CPA as described earlier. Figure 1.3 depicts this process.

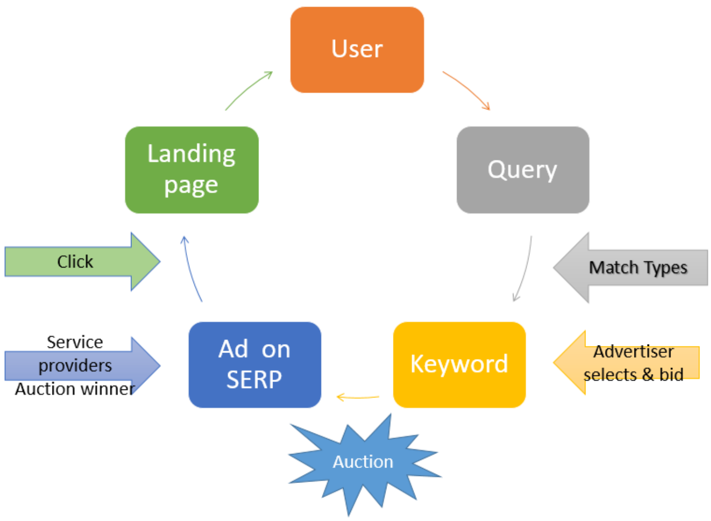

Figure 1.3: Search Engine Advertising Process

Advertisers need to know the auction mechanism very well in order to be successful in SEA. Auction determines ads' value and position on the SERP. Advertiser provides a keyword list that identifies their ad and the max bid (i.e., max CPC) that they are going to pay if their ad clicked. This list consist of phrases, words that describes their product or service which is going to match with the queries that users search. Search engine selects the most relevant and the max bid keyword from the keyword list that advertiser provide for the specific ad and the auction starts. The quality scores of the keyword for every ad are delimited by the search engine provider by their algorithm and ad ranks are calculated. One with the best ad rank is the winner of the auction. However because it is generalized second price auction, winner pays according to the ad rank of the ad bellow it. Figure 1.4

illustrates how the auction mechanism is performed for a case with four advertisers. The advertisers provides keywords and the max bids. Quality scores of the keywords of the advertisers are determined by the search engine. The Ad Ranks calculated by multiplying the max bids and quality scores and the advertisers are ranked based on Ad Ranks. Actual CPC's that the advertiser pats are calculated by dividing the Ad Rank of the next advertiser with the quality score of the advertiser.

Figure 1.4: Auction Mechanism
auc (2017)

According to AdWords, Google is calculating ad rank and determining cost per click (CPC) by multiplying quality score and max bid but the exact auction calculations are not given by any search engine, they are undisclosed. Although the formulas are not provided, definitions of the metrics are given. Quality Score is one of them, it is a combination of relevance and user experience that search engine determines how well the ad related to the users query. Quality score depends on multiple factors such as expected click through rate (CTR), keyword relevance, landing page quality and ad format. In this sense in order to find user experience search engine measure the landing page experience of the user and expected click through rate. How useful is the web site the user directed, how long did the user stay in the web site and how many pages in the website did the user travel through, are the actors that are used to determine landing page experience (Miller, 2016). Its mostly about the effectiveness and the relevance of the web site to the user. On the other hand expected click through rate is another metric that affects the quality score (Miller, 2016). Search engines use expected CTR to determine how likely users are going to click the ad. Expected CTR is based on CTR. CTR is calculated by the number of visitors divided by the number of ad displays on SERP for the keyword. The relevance between the keyword, ad and landing page is the other important determinant that affects the quality score. Keyword relevance or ad relevance is a metric about how closely the advertisers keywords related to the ad (Miller, 2016). Ads are the objects that reaches to the customer so the relation between ad and keywords is important in reaching the right customer which is searching for the specific product or service.

Ad type is another factor that affects the quality score. Search engine services has many types of online advertising formats and also in SEA they have several types too. For example in Figure 1.5 there is three types of ads for only one query, namely shopping network ad, dynamic search ad and basic search ad. Selecting the best fit one for your product and customer has an important role in budgeting and reaching the customer.

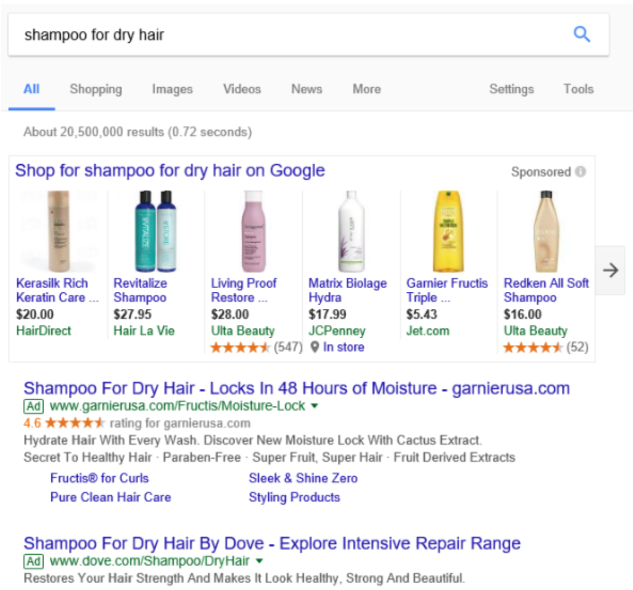

Figure 1.5: Ad Types

The content of the ad has an influential effect, especially the first and second title/headlines. Search ads have two headlines consist of 30 characters. In dynamic search first headline is the query that the user search, this directly takes attention of the customer but in a basic search ad these headlines specified by the advertiser. Also search ads has two lines of description, that comprise of 80 characters. Users distinguish the search ads from organic search results by the little ad sign near the landing pages' url.

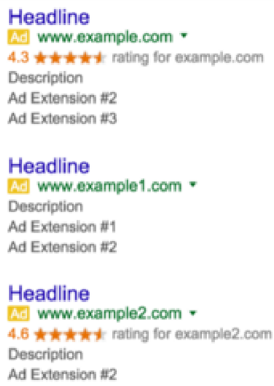

Figure 1.6: Ad Format

1.4 Motivation

It is clear that SEA is (and will continue to be in near future) is a major media for companies in order to reach their target customers. Targeted advertisement enables the companies to increase the awareness of the customers and increase their revenue with smaller budgets. Therefore identifying the relevant keywords in an advertisement campaign as crucial role for the success of the campaign.

The relevant keywords has to be selected for the advertisement campaign by the domain experts. It is hard to detect them manually because advertisers are responsible of many campaigns for many products and the process is not consistent across time. Most of the keywords performance can change according to seasons and special occasions.

Keywords are consisted of word, i.e. unigrams. Some unigrams can have negative effects on the keywords that are used in the campaign. In the literature, determination of the negative unigrams and eliminating the keywords that includes those unigrams is proposed as a solution for the problem that advertisers are facing. These words has to be detected and the keywords that consist of these words has to be eliminated from the keyword set which is going to use in the SEA.

The current proposal that are available in literature use conversion rates in their analysis. However, using just conversion rate to measure the ads performance is not adequate because significance of the conversion can change according to the marketing strategy of the advertiser. For example big brand advertisers usually do not consider conversion but pay more attention to impression and/or clicks of the user. Their aim is to reach as much as customer and show their ads more on prominent slots.

This research bases on the current proposals that aims to determine the negative unigrams by using data mining/ machine learning approaches in order to manage the advertisement campaign efficiency.

A decision support system is developed to automate the managerial decisions regarding to, selection of the relevant the keyword sets process for the campaign. According to the given historical data system split the keywords in to two sets as negative and positive. Since the current metrics used to determine the negative keywords are not sufficient for some of the advertisers, we use *"Click Potential"* which is proposed by Jansen and Clarke (2017) and developed a new metric called *"Bounce Rate Index"*. These two metrics provide us to investigate more deeply the impression, click and bounce rate relationship between keywords and users. These metrics are used to label the keywords as negative or positive in the analysis.

After labelling the negative keywords with these metrics we carry our analysis to unigram level. We detect the negative unigrams with the help of machine learning techniques. For this purpose we use some core concepts associated with Naïve Bayes and Decision Trees and also Logistic Regression as proposed by the earlier research(Bulut, 2015). If a unigram is seen mostly in negative keywords, we eliminate the keywords which contain that specific unigram. Therefore we constitute a keyword set without the negative unigrams, which is our final keyword set to be tested. Further more in order to evaluate the overall campaign performance we compare the campaign which has negative keywords with the same campaign which does not contain negative keywords according to their cost for the advertiser. We also developed our approach to measure and compare the ad campaigns as part of the research.

The structure of the thesis is as follows. Next, in Chapter 2 we will provide the relevant literature. The problem statement will be presented in Chapter 3. The metrics and methods used in our analysis will be detailed in Chapter 4. The experimental analysis where the performance of the various approaches will be provided and results will be discussed in Chapter 5. We will conclude the thesis with our remarks and propose some future research directions in Chapter 6.

Chapter 2

Review of the Relevant Literature

Given magnitude of the economic activities, SEA received a lot of attention from researchers. Therefore an extensive review of literature is prohibitive.

We will first start with review of the literature regarding to the general concepts in context of SEA. We will only cover some of the most cited researches in this part. Next we will focus to keyword generation/ selection problem. Again we wil restrict our review with those that are most prominent.

We will also provide the review of the negative keyword selection literature.

2.1 General Literature Review

Jansen and Mullen (2008) provides a general overview of the concept, history and technology of SEA. Even though, a lot of enhancements has done both in technology and process of the SEA since Jansen and Mullen (2008) was published, their research still provides the key concepts and conceptualize the sponsored search process. The major actors (customer, advertiser, search engine provider), their goals and the process is clearly presented in the paper. Auction mechanism and different auction techniques that has been used is also presented. Payment mechanism, bidder strategies, issues caused by the intersection of the sponsored search participants' profits is discussed in detail. Jansen and Mullen (2008) consider the SEA process as an aspect of information searching, rather than viewing it only as an advertising medium. Queries are searched to reach the right advertisement, and according to the search at the end relevant organic and sponsored listings appear on SERP. Therefore the process can be viewed as seeking information.

Search engine advertisements appear with organic searches on the same SERP after the customer search for the same query. These both organic and non organic listings can effect each other because customer can prefer to click organics rather than advertisements even he/she searched for purchasing. The research that conducted by Yang and Ghose (2010) aimed to model and estimate the interdependence relationship between organic and paid search results of the same firm from consumer' responses. They address to the relation between the presence of organic listings on SERP and the click through rate of the paid search advertisements. Their integrated model which is a hierarchical Bayesian framework, is built to estimate the relationship between metrics such as CTR, CR, CPC and keyword rank. As a result they claim that the CTR of the organic search have a positive interdependence with CTR of the paid search and also organic clicks increases the utility from paid click. Therefore this positive interdependence lead to an increase in the revenue of search engine provider. Throughout the research aggregate consumer response data set used one hundred keywords over three month period. Keywords are divided into three parts according to their characteristics as, retailer specific, brand specific and the length of the keywords. The strongest positive interdependence is detected in least competitive keywords such as retailer specific. They state that interdependence relationship is an important determinant of the effectiveness of the search engine advertising.

Rutz et al. (2011) focus on the dynamics of the online customer search process. Some customers return to the web page by directly without using the search engine ads after they reach the web page by pay per click ads once. This cause an indirect effect on the search engine advertisement revenue. Therefore they conceptualized SEA as a hybrid between direct responses and indirect effects. They analyzed a commercial automotive industry keyword level data set to find out the previous paid search visits on current direct type-in visits. Therefore the semantic characteristics examine to recognized the keyword effectiveness on direct type-in traffic. Bayesian elastic net model applied and keywords clustered into their textual attributes such as branded and broad. Branded keywords are directly relevant with the ad but as a result broad keywords have more value in long term because broad search is an early stage of information gathering so it might cause return visits.

Companies can give their brand names as keywords to the search engines but on the other hand competitors of that company can bid on the same brand names too. This is called "piggybacking" in the context of SEA. Although there are so many cases about piggybacking in the court, it is allowed in U.S., Canada, Ireland, and the U.K. since 2009. Rosso and Jansen (2010) investigate piggybacking with three types as competitive, bidding to competitors brand name, slogan etc. directly, second promotional, re-seller or the retailer of the brand bidding on the brands name etc., and third orthogonal, offering the user different products/ services offered that offered by the brand. In their research three main internet search engines were covered; Google, Yahoo! And MSN Live. According to the results most popular search engine Google does not display competitive ads on the north side, which is the most attractive part of the SERP. This attitude is because of the negative impact of competitive piggybacking on users. It causes confusion on users and search engines want to provide the most relevant results to its consumers. Another result of the research was, large advertisers want to self-bid and own the pages that result from searches for their brand names, but prefer to pay lower prices.

Using brand names as keywords makes the companies advertisements more relevant to the customer they target so their ad relevance is great. Recall that ad relevance is one effective component of the quality score which effects the auction results, i.e. the cost per click that the advertisers pay. Unfortunately search engines do not provide the formula of it. Hillard et al. (2010) tried to formulate the ad relevance and predict it through the query ad pairs that maintained from search engine logs. Machine learning techniques are implemented on the query ad pairs to predict the sponsored search ad relevance by Hillard et al. (2010). A binary classifier trained to detect relevant and irrelevant advertisements, maximum entropy, AdaBoost decision tree stumps and gradient Boosting Decision trees are used. The model extended with the past user clicks to learn semantic relationship between queries and ads. The semantic relationship information provides the highly related words which is helpful in predicting ad relevance.

Various methods/techniques are used in the literature to classify and predict the performance metrics, i.e. CTR,CPC. However there is limited research on the application of data mining techniques to classify the search engine campaign performance on profit. King et al. (2015) combine ensemble learning techniques with classification models to fill the gap in the application of advanced data mining techniques on pay-per click campaign classification. Four popular base classifiers as Naïve Bayes, Logistic Regression, Decision Tree and Support Vector Machines are used to predict profitable pay-par click campaigns. Search engine advertising data has ad impression, CTR, the average page position and the conversion rate as dependent variables. Therefore base classifiers combined with ensemble learning techniques such as voting, boot strap aggregation, stacked generalization and Meta Cost on Rapid Miner to get more accurate results. During the process text mining is used to preprocess the textual content of the advertisements by tokenizing them in to stylometric, sentiment and semantic features. As a result Meta Cost ensemble models and logistic regression models performed well in classifying whether a new campaign is profitable or not. (King et al., 2015).

2.2 Keyword Generation and Selection Literature Review

Keyword generation and selection is the most indispensable part of Search Engine Advertising. Main part of the SEA is the keyword auction which the advertisers bids for the specific keywords that they select to advertise their product or service. The keywords have to be selected meticulously. There are many factors such as budget constraints, relevance and popularity of keywords etc., affecting the keyword selection and generation which makes this process complicated. Also CTR, CPC, CR, AdRank, quality score, bounce rate and the interdependence among them should be considered during the keyword selection process. Search engines force the advertisers to bid for highest ranks by the means of the auction mechanisms. Bidding on the popular, most relevant keywords or bidding for the highest price is not the best decision because other auctioneer's bids, ad relevance and the quality score of the advertisement effects the CPC and the CTR of the corresponding keyword advertisement pair.

Advertisers want to bid on relevant but also at the same time economic keywords. Often advertisers tend to bid on popular keywords, and this inclination makes these keywords competitive so they become expensive. As a result, advertisers come up with a more economical strategy by bidding on nonobvious yet relevant keywords, which are economically more affordable. Joshi and Motwani (2006) developed TermsNet to find terms that are related with the content but do not occur so much in the context(uncommon) and they define a new measure of nonobviousness. TermsNet is a novel graph-based technique to identify relevant yet nonobvious terms and their semantic associations. It aims to find a solution to this challenging task and also can be used for clustering terms in study of word relationships. Their approach creates a characteristic document as a corpus which contains text-snippets from top 50 search hits for that specific term. Joshi and Motwani (2006) finds out relevance relationship between terms from the characteristic document with directed relevance notion in association rule mining context. Such as if A is related with B but B does not need to be related with A. This relevance found out by looking at these terms frequency in the other characteristic document. A directed graph is constructed using this measure of similarity. The outgoing and incoming edges for a term are explored to generate suggestions.(Joshi and Motwani, 2006)

Commonly occurring terms are more expensive than infrequent ones because there is a strong correlation between the search volume and the cost per click. Another research similar to TermsNet that aims to find related but low volume and inexpensive terms is coined by Abhishek and Hosanagar (2007). They created a platform called Wordy which uses a web based kernel function to find the semantic similarity between the terms. Wordy targets to find these long tail terms that generate same amount of traffic cumulatively but are much cheaper than popular words by using the semantic similarity between terms. (high volume keywords vs large number of terms from tail). The research has three main objectives; generating a large number of keywords, establishing semantic similarity between these keywords and suggesting relevant keywords that might be cheaper. A corpus is created from the advertisers (merchants) website as HTML format with an Information Retrieval package website parsed and pre-processed. Tfidfs (vector weightings) of all words are computed, the ones with lower vector weightings than the threshold eliminated and the other ones constitute the dictionary. Each term in the dictionary queried on a search engine, the top documents are retrieved and added to the corpus. Updated corpus is analyzed and the important terms added to the dictionary. This dictionary contains the convenient terms that the advertiser can use for SEA as keywords. Relevance score of each keyword pair in the dictionary calculated to find the semantic similarity between terms. Each term in the dictionary queried and a large amount of data drag from the web. Returned documents create a context vector for each term in the dictionary. These context vectors compared by dot in order to product to find out the semantic similarity between terms. Semantic similarity between terms projected to an undirected graph. The proposed algorithm finds out the low frequency terms by breath first search from the graph that is suggest cheaper keywords to the advertiser.

The above mentioned two researches, determine the semantic similarity based on the co-occurrence frequency of the two keywords and assumees that if two terms are semantically similar this fact cause them to co-occur together. Therefore semantic similarity adopted *indirectly* on these researches. Chen et al. (2008) tried to fill the gap of semantic information by suggesting new keywords based on concept hierarchy. This approach aim to minimize the low coverage and lack of disambiguation in keyword suggestion especially for the terms that has two or more meanings. The approach presented in the research, match the given keyword with the relevant concepts, and look for phrases related with the concepts. Concept hierarchy derived from a high quality manually defined web directory which is called Open Directory Project (ODP) and has the definition of the concept relationship and content. After the matching stage the related concepts categorized with their distinct meanings to avoid the interference. Relevant concepts categorized into clusters in this stage by hierarchic truncating algorithm. Therefore these categorized suggestions provide new phrases tightly relevant to the distinguishing meaning which are the suggested keywords. Concept hierarchy enables advertiser to choose its campaigns keyword scope either general or special by pruning. The major disadvantage of this method is the relatively static structure of the concept content which comes from ODP. However there can be words that has not occur in the content. (Chen et al., 2008)

Therefore, Kiritchenko and Jiline (2008) used past query information which provides more detail about the customers' purchase behaviour and can identify words that does not occur in the static structure of the concept content. Kiritchenko and Jiline (2008) aim to have an optimized keyword list by using an algorithm that is based on feature selection. The method used on all possible word combinations and past user queries of original keywords which are selected by the advertiser. Objective of the research is to gather the most profitable and specific combinations of the original keywords in a keyword list which make the advertisers' campaign more profitable. Generally single words have broad meanings and multiword phrases have more specific meanings so they can be more discriminative keywords. So all possible combinations of words appealing in past user queries gathered together then filtered by feature selection algorithm. In this approach all word combinations are sorted by their importance, top n scoring ones are selected from this new set of keywords. The n selected according to the revenue cost analysis of the advertiser. Highly ranked negative phrases are ranked and added to the list in reverse order or original set of keywords which is given by the advertiser are replaced with extended keywords that were generated from the list. When the algorithm is compared with other classification methods, it yields appreciable results. However this research ignores the individual costs of the keywords, their placement in the auction, max daily budget of the advertising campaign and similar critical information. So it has to be designed for the specific task with the conditions that effecting the SEA campaign. (Kiritchenko and Jiline, 2008)

Another approach to keyword generation problem is proposed by Thomaidou and Vazirgiannis (2011). which is a semiautomatic system that make multiword keyword recommendation. Similar to other researches, Thomaidou et. al. also assumes single words that frequently used has broad meaning but multiword phrases are more specific which makes them more representative to be suggested as advertising keywords. The research aimed to optimize the human resource effort and improve quality, quantity and variety of the proposed keywords. To achieve this goal the system proposed by Thomaidou and Vazirgiannis (2011) divides keyword generation in to two sub-tasks as keyword extraction and suggestion. The system extracts relevant terms consisted of two or three words from landing page to match potential search query. Vector space model is used to segment the text as weighted keywords. Special weights assigned to each different Html tag and TF-IDFs evaluated according to their semantic meaning and relevance to the landing page. These scores are calculated by the terms frequency. A word occurrence matrix is constructed to pull together possible combinations of two word phrases by the extracted single word terms. This process is repeated in order to construct three word phrases. These extracted keywords are given as the input to the search engine API which returns short text snippets so most relevant new support vectors are produced as suggested keywords. Experiments and system evaluation is conducted by experts using scale of 1-5 on relevance, specify and non-obviousness. Accrding to their results, the automated system proposed to find the appropriate keywords and recommending multiword terms outperformed the other competitive keyword suggestion tools especially in prominent competitive industrial categories(Thomaidou and Vazirgiannis, 2011).

Thomaidou et al. (2014) later advanced the above mentioned research to a PhD thesis and propose an automated system that creates and optimizes online advertising campaigns. They designed a system that is capable of automated advertising campaign creation, management and also monitoring profit optimization under budget constraints. The system organized as three parts and the first component of this system is the keyword generation which is explained earlier. The second component is the budget optimization which is a challenging task, done during the campaign running time. Finally a fully implemented system was developed for Google AdWords platform; using Global monthly searches and Competition of Google AdWords tools for observing and predicting the campaign behavior of the advertiser.

In the budget optimization component they present a methodology for selecting the most effective keyword bid pairs aiming to maximizing the *profit* for specific budget and the *traffic rate* to the website separately. Selecting keyword and bid combinations is a kind of multiple choice knapsack problem (MCKP). Optimal solution of the MCKP will indicate the best possible choice of keyword bid combinations. The approach is a combinatorial optimization problem, and the authers propose a genetic algorithm for the solution. Within this process, proper statistics collected from previous time periods and most profitable options kept for the next time period. These proper statistics are CTR, impression average CPC and conversion rate of the keyword. Thus genetic algorithm finds the proper options of MCKP for profit or traffic maximization. Thomaidou (2014) use MCKP model to focus on clicks that each keyword gain rather than the placement of the ad on the ad slots so Thomaidou (2014) gives more importance to number of clicks than position of the ad.

Test of the system done for four budget optimization scenarios; with prediction or without prediction for both traffic and profit. Every result of the genetic algorithm application for each optimal keyword- bid combination; provides data of clicks, cost, profit, keywords used in and average bid. The system is evaluated with the data taken from a website then tested on two websites. Also compared with the manually created campaigns. Automated system with prediction performed better than the manual campaigns which shows that the system is useful for the search engine advertisers. (Thomaidou et al., 2014)

2.3 Negative keywords Generation/Selection Literature

Çoğalmış (2016) designed a semi-supervised tool called AdScope for filtering out the unprofitable user queries from the search campaign while at the same time allowing the profitable ones . For Çoğalmış (2016) the profitable queries are the ones which has high relevance and high conversion rate. Therefore in order to determine the unprofitable user queries they use relevance feedback and classify the user queries as relevant and non-relevant. Queries that classified as relevant included in the campaign as regular keywords on the other hand non- relevant classified queries are excluded. Two source of feedback is used to classify the queries; one is a feedback of the domain expert which is done mostly by expert knowledge and the classifier that uses the labels based on the conversion rate of the query which can be observed from the search query logs provided by the search engine provider. These two steps built up on Binary Independence Model and constitute the conversion model. BIM is used because it has good performance in Information Retrieval tasks on short documents. To label each query the Relevance Status Value (RSV) of the queries determined by the log odds ratio which is calculated using the odds of probability of a term appealing in relevant and non-relevant documents. Sum of the odds ratio of the terms in the query give RSV of that query. These unlabeled queries with high and low RSV are shown to the advertiser for relevance feedback to be labeled by the advertiser.

The model tested with pre-processing and without pre-processing. Their findings indicate that advertisers should use pre-processing in order to maintain more valuable information. To measure the advertiser's feedback several experts' with same prior knowledge used the model (AdScope) and tagged the queries as relevant or non-relevant. In their research AdScope is compared with multinomial Naïve Bayes, Binary classifiers and Markow-Chain models. As a result AdScope gives competitive classification accuracy compared to these offline training models.

AdScope is a conducive research to RightScope which is a study developed on the objective of selection of right keywords (Alchalabi et al., 2016). In this study the queries are also classified relevant and non-relevant. RightScope is a supplement of AdScope so the relevance and profitable concepts has the same meaning in both of the researches. As in the AdScope, BIM model is built using both statistical and machine learning approaches. These are RSV from the probability Ranking Principle model which used in the previous research that mentioned above, Logistic Regression Model's probabilities and weighted cost model. Preprocessing done to the data and only terms that appear at least three times included. Conversion, relevance and number of converted clicks of the each query taken in to account while labeling them.

RSV used as mention in Çoğalmış (2016) and logistic regression is a good fit to the problem because there are two classes. Weighted Cost Model which they modeled calculates two values for each term in the query. One is for the terms that contributing to the total cost resulting all relevant queries and the other is for non-relevant queries. With these values and total CPC of the terms a weighted CPC score of a query is calculated. According to these methodologies two different BIM models were built; a term based weights combination and query based score combination. The accuracy of the models show that two models worked better in finding non converting queries/ terms than converting ones(Alchalabi et al., 2016).

Another conductive research on finding the negative keywords has come from Bulut (2015). As mentioned in Bulut (2015) 10% of the campaign budgets spent to irrelevant queries therefore the study aimed to find out the characteristics of these irrelevant queries using the past data of the campaign from the unigrams. Naïve Bayes classifier, decision tree classifier and association based methods such as Jaccard similarity index is performed on search query data. The data contains the historical campaign performance, relevance judgment per user query and number of conversions per query that used to distinguish converting and non-converting queries. After the query set divided *manually* into two classes (converting/ non-converting), according to the conversion rates of the queries, keywords split into unigrams. Non-converting queries has the negative unigrams and these unigrams which occur more than the threshold constitute the initial candidate set.

The four data mining methods performed on this set. As a result eliminating the negative keywords that detected by the study, decreases the CPA of the campaign 25% and logistic regression outperformed all other methods. (Bulut, 2015)

A system is created by Hubinette (2012) in order to detect the negative keywords in a search engine advertising campaign. The proposed system developed to be used with an advertising service in order to enhance and maximize the relevancy of the advertisement presented which will increase the user's action rates (CTR etc.) and conversion. Mainly the system evaluate one or more search criteria that associated with one or more keywords, to identify the negative keywords. First of all it identifies the search criteria that is relevant to the advertisement with the help of Bayesian networks. This search criteria can include, match at least one of the words or phrases in the presentation of the advertisement. Then determine the negative keywords in unrelated search criteria. The server of the system consist of a memory and a processor. The memory includes keyword files, query files, off-topic queries, models, evaluation criteria and negative keywords. Moreover processor constitute the evaluation engine which determines the keyword file and an associated query file for an advertisement, then map them to a set of related clusters using one or more models. Afterwards, evaluation engine identify the search criteria that are not associated with the advertisement from keyword and query clusters using evaluation criteria. Then calculate relevance scores associated with search criteria, according to semantic similarity and concept hierarchy. Conferring to relevance scores the off topic search criteria is detected and selected as negative keywords. The system identifies the negative keywords with the help of semantic concepts and probability of occurrence and at the end maximizes the relevance of the advertisement.

A recent article on search engine advertising is Jansen and Clarke (2017), which propose several metrics as click potential and conversion potential for the SEA to plan and evaluate the campaign performance. The paper underlines the need for a metric to evaluate the traffic and conversion simultaneously. Jansen and Clarke (2017) highlight that the current metrics for SEA are misleading and not sufficient for the managers to understand the overall campaign because SEA campaigns use a multitude of independent variables. Therefore they introduce a combined metric called Conversion Potential (CvP) which is managerially useful and uses multiple SEA factors. To develop the CvP they begin with the click potential. CP is a predictor factor or summation of predictor factors that influence the possibility of an ad attracting a searcher's attention and generating possible clicks. Consequently it can be defined as the overall opportunity of an ad to be viewed and clicked. CP is an absolute metric hence it does not provide sufficient detail to understand the change in the number of impressions with the change in another attribute they use relative impressions and propose relative click potential. RCP informs advertisers about the overall traffic goals. From these metrics they compose CvP to evaluate the change in number of conversions based on the change in some other attribute. This metric gives opportunity for future conversions to occur based on past traffic and sales. As a result they found out that the significant difference in CvP was due to the change in CP. Evaluating the site traffic and conversion simultaneously provided a richer interpretation of the advertising data. CvP measures both campaign traffic and sales using ad rank. Results indicates that there is a dramatic fall in CP from ad ranks 4 to 16. Also these two metrics (CP, CvP) simultaneously incorporates a sales ratio and a traffic ratio, hence provides managers to make more informed decisions more effectively targeted bidding behavior and budget management.

Chapter 3

Problem Statement

A search engine advertising campaign process involves three actors namely, the user, the search engine provider and the advertiser as illustrated in Figure3.1. Each actor contemplate different results from the campaign and each has different objectives. The user wants to attain the product or service she/he searching for without much hassle. The search engine provider aims to give the best experience to the user by displaying the best ads related to the users search query on SERP (which is crucial for long term) and at the same time wants to maximize their immediate advertising revenue. On the other hand the advertiser aspires to reach the relevant users(i.e. potential customers) with their ad.

Each advertiser has their own particular constraints and expectations from the SEA. According to their market expectations and perspective some advertisers want to maximize their conversion rates and some others want to make maximum possible impressions or clicks. As Nabout et al. (2014) assert that in highly competitive markets the brand aware companies measure their advertising effectiveness by higher number of clicks. Whereas big-brand advertisers want to display their ads in more prominent slots without considering that the ads are economically justifiable or not (Yuan et al., 2015). Therefore measuring the campaign performance only with conversion rate is not sufficient. However most of the studies and advertisers usually limit their attentions to the conversion rate as a measure of the campaign performance which might be misleading in some cases.

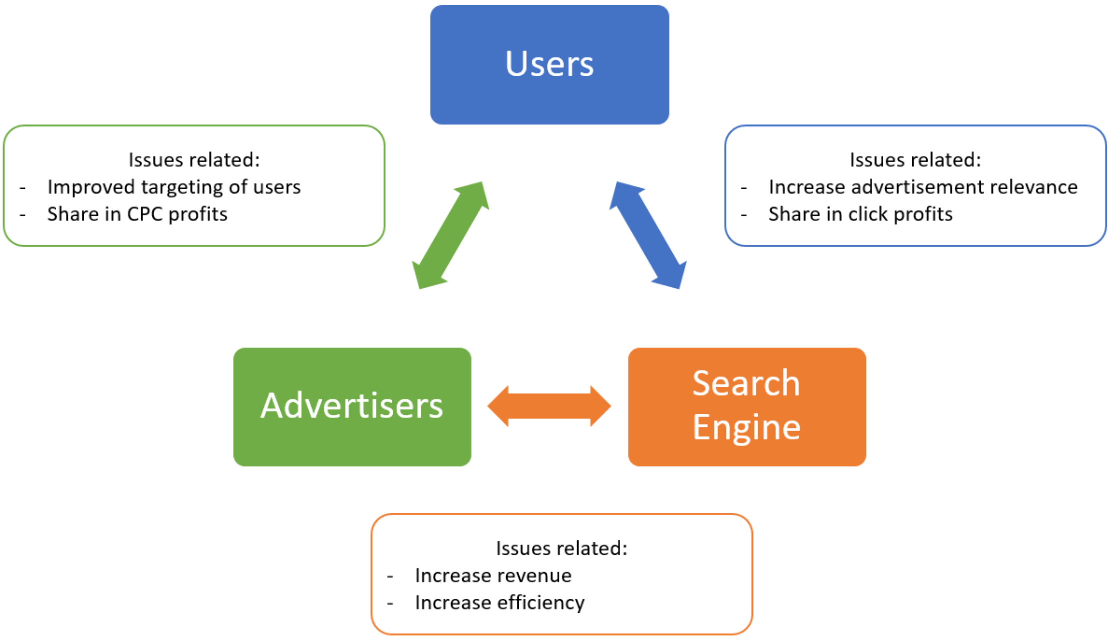

Figure 3.1: Actors

Conversion can be defined differently in terms of the desired goal or behavior of the searchers action (Jansen and Clarke, 2017). For example completing a form, downloading a content, making a purchase or order can be considered as conversion in different contents. In this research we focused on impression and click as major factors that affect the campaign performance in SEA. We consider the case where exposure of an ad to the customer has an advertisement value. Also by clicking the ad, being directed to the landing page and viewing the page for a certain amount of time, means that the customer is interested in the advertiser's product or service. In this sense considered impression and click as a proxy of conversion for the advertiser. Also bounce rate of the user is an important aspect while counting on the clicks. Bounce rate is related to the time that user spend during his/her visit of the landing page. If the user spend adequate time on the landing page, it shows that she/he is involved in the advertisement and advertisement achieve its goal. Otherwise, the user bounces back to the SERP which is an indicator of the disinterest of the user.

Impression, click through rate, bounce rate and other metrics of the SEA are not independent. They all effect each other in the campaign process. For example as shown in Figure 3.2 advertisements with higher quality scores will lead more impressions and better positions therefore lower cost per click values (Yuan et al., 2015). It is known that high per click values make advertisers to raise their bids which will lead them to have better ranks and higher CTR. Then the CTR cumulatively effect QS to raise which will decrease the CPC. At the very end lowered CPC and increased CTR, grow the revenue which will afford advertiser to bid more aggressively (Yuan et al., 2015). Thus to decide which metric to consider is a hard problem, taking only some of them in account can be misleading for the managers to evaluate the traffic and conversion going through in the campaign. Unfortunately there is limited to no academic research regarding SEA metrics which is a standard measurement for accessing in this manner (Jansen and Clarke, 2017). In our research we propose a solution to this problem by using conversion potential and introducing modified click potential index to determine the best keyword set.

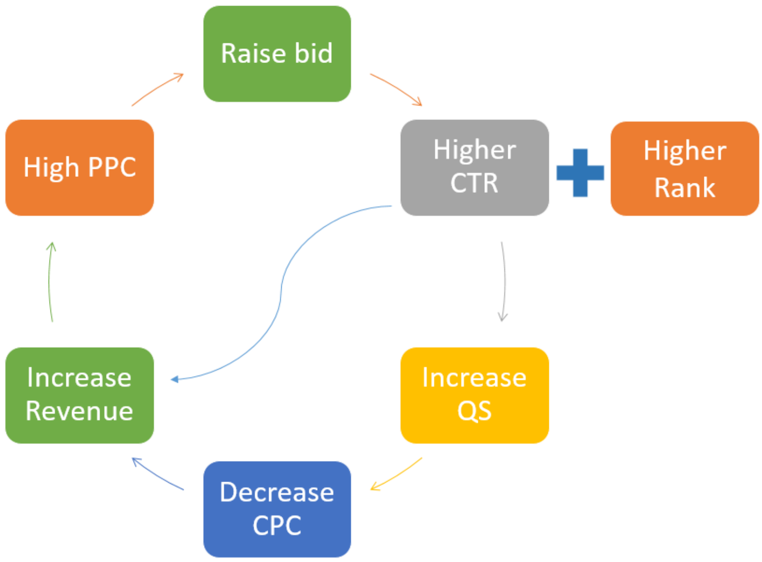

Figure 3.2: Interaction between metrics

Determining the best keyword set is a major problem for the advertisers in order to target the right set of users. Advertisers do not want to spend their limited budget on irrelevant keywords because each click cost money to the advertiser. Thus, in order to target the right set of users and make the campaign more profitable, right keywords has to be selected. In practice the keyword lists are usually determined by the domain experts. They periodically go through their list manually and based on their experience expand or shrink them (Bulut, 2015); (Çoğalmış, 2016). Beyond this process is not consistent across time. A query can specified as positive but after 6 months a similar one can specified as negative. Also advertisers usually responsible for many campaigns for many products for many seasons so the classification of a keyword as positive or negative needs high experience. Therefore manually conducting the process is inadequate to detect such inconsistencies and provide a clear solution (Çoğalmış, 2016). An automated tool such as our decision support system is needed to detect such keywords as positive or negative. Bulut (2015) on his research with historical campaign data and relevance judgment try to devise o scheme to this problem. Rather than just relevance judgment in our study we quantify campaign with a measure. We try to estimate campaigns overall performance from historical data that we have by eliminating the negative keywords.

The main problem is to detect negative and positive keywords. The historical data which arranged and served by search engine provider has click, impression, quality score, CTR, bounce rate and such attributes about the campaign. By using these we constitute two metrics called *"Click Potential"* and *"Modified Click Potential Index"* to distinguish negative and positive keywords and form two keyword sets. Then we conduct our research on unigram level and detect the words that occur frequently in negative keywords by algorithms based on naïve bayes, decision trees and logistic regression. After determining the negative words we eliminated the keywords that contains them according to the elbow criteria that the words have. Most of the research that has done to distinguish keywords are not done in unigram level. We tried to give this problem a solution on unigram level. Also there is limited research on the application of data mining techniques for SEA (King et al., 2015) by using classification techniques we attempt to close this gap.

Chapter 4

Methodology

4.1 Workflow of Decision Support System

4.1.1 Pre- Processing and Identification

Search engine providers periodically provide advertisers the search terms reports. These reports have the keyword query matches and many other statistics such as CTR, impression etc. per query. In this thesis keyword sets, keywords and unigrams in the search term reports are represented as shown below.

$$K = \{K_1, K_2, K_3, ..., K_n\}\, t = 1...n \quad K_t \quad \text{is the keyword set for period} \quad t$$

$$K_t = \{k_1, k_2, k_3, ..., k_m\}\, i = 1...m \quad k_i^t \quad \text{is the} \quad i^{th} \quad \text{keyword set period} \quad t$$

$$k_i^t = \left\{u_{i,1}^t, u_{i,2}^t, u_{i,3}^t, ..., u_{i,j}^t\right\} j = 1...l \quad u_{i,j}^t \quad \text{is the} \quad j^{th} \quad \text{unigram in the keyword} \quad i$$

Statistics of a keyword is the attributes of that keyword. Let's assume keyword set is

$K_{t=1}$ each keyword k_i^t has attributes as,

$$k_1^1 = \{Click_1^1, Impression_1^1, CTR_1^1, Avg.CPC_1^1, Avg.CPM_1^1, Cost_1^1, Avg.Pos_1^1, QS_1^1, ...\}$$

In order to identify the relevance of the keywords, CP and MCPI of the keywords calculated. The keywords that have calculated metric greater then the mean of that keyword set labeled as positive and the other ones labeled as negative. More formally,

$$K = \{(k_1, +), (k_2, -), ..., (k_n, +)\}$$

$$K^- = \{(k_i, c_i) : c_i = - \wedge k_i \epsilon K\}$$

$$K^+ = \{(k_i, c_i) : c_i = + \wedge k_i \epsilon K\}$$

$$K = K^- \cup K^+$$

Where each item of *K* (keyword set) represented as a tuple *(k,c)*. First attribute *k* is a keyword and *c* is a boolean indicator for relevance. When *c* is ”-” it indicates that the keyword has a positive relevance and if it is ”+” it indicates the opposite. Emphasized notations are represented in the table 4.1

Entity	**Symbol**
a keyword	k
an unigram (a term in keyword)	u
the set of keywords	K
the set of negative keywords	K^-
the set of positive keywords	K^+

Table 4.1: The Table of Notations

4.1.2 Labeling Keywords with Click Potential and Bounce Rate Index

Google Adwords provide historical data for the advertiser, with that given historical our DSS firstly label keywords as negative and positive according to their *Click Potential (CP)* and *Modified Click Potential Index (MCPI)*. CP is formulated by Jansen and Clarke (2017) to observe the traffic rate on an ad. It is a predictor factor which impact the possibility of an ad attracting a user's attention and then generating clicks. Its formulation is shown as:

$$CP = RI * CTR \tag{4.1}$$

Where RI resembles the relative impressions and calculated as:

$$RI = I/BI * 100\% \tag{4.2}$$

Due to insufficiency of absolute metrics there is a need for relative metrics to observe the changes in any other attribute. Also relative metrics are becoming the rising major tools in marketing (Keiningham et al., 2015). Here relative impressions represent the impressions in relation to baseline measure. BI is the baseline impressions which is taken as mean of the all keywords' impressions in the dataset.

The second metric developed in this research is Modified Click Potential Index which is calculated by CTR and Bounce Rate. CTR is the traditional observable measure of user satisfaction but it has limitations while observing the user experience on landing page. However since bounce rate is the fraction of all click throughs which result in a bounce with in a certain limited time, it is an effective measure of user satisfaction (Sculley et al., 2009). Also it is a essential metric for advertisers to examine the conversion action because user who bounces from the landing page is unlikely to perform a conversion action(bou, 2007). Therefore high bounce rate means that the ad is serving a poor experience to users after the click action which results as a poor advertisement gain. As a result advertisers who has good reputations among users have higher click and lower bounce rate.

Although bounce rate is an intuitive and widely used metric, there isn't enough research about it in literature. So in this research we tried to investigate a new metric using the bounce rate to enhance the studies about this subject. Bounce rate has a strong correlation to CTR (Sculley et al., 2009). While CTR of an advertisement has an increasing trend, on the other side decreasing bounce rate characteristic has been observed. So we formulate Modified Click Potential Index as:

$$MCPI = (1 - BounceRate) * CTR * RI \tag{4.3}$$

MCPI of an ad measures the satisfaction of the user, between the period when the ad shown on SERP till the user left the landing page. Therefore it provides information about the relevance of the keyword that matches with the user's search query and also relevance of the ad it self to the user. Since our dataset is from a big brand advertiser, modified click potential index useful to detect the negative keywords in terms of big brand company advertising goal. They want to be more attracting to the users and do not want to misplace their significance among users so modified click potential index is a satisfying metric.

	September		October	
	positive	negative	positive	negative
CP	44	624	112	1051
MCPI	–	–	162	1001
Total Keywords	669		1163	

Table 4.2: Distribution of Keywords

Keywords are labeled as negative or positive after calculating the CP and MCPI of each of them. So we created two different datasets from September and October Keyword

Level data, one of them is labeled with respect to CP and the other to MCPI. The keywords below the mean of CP taged as negative and the ones above that tag as positive. Same process held with MCPI too. Distribution of the keywords are shown on Table 4.2.

4.1.3 Eliminating The Negative Keywords at Unigram Level

Three-fold cross validation is applied on the keyword level data after the labelling. Datasets are divided into two parts as train and test. Then a validation set generated from train data and split into two parts. One part of the validation set trained with the classifiers; Naïve Bayes, Decision Tree and Logistic Regression. Keyword level datasets transformed into unigram level datasets as represented in the Table 4.3 to perform the classification methods in the training part. Unigram level data provides to observe which unigrams in the whole campaign are in the keyword and which are not. The data also has the information of sign which is the label of the keywords as negative or positive. Classifiers are trained to find out which unigrams are mostly seen in the negative keywords. Detected negative unigrams eliminated from the campaign keyword set by taking out the keywords which have them.

	unigram 1	unigram2	unigram 3	$\cdots$	unigram m	label
keyword 1	0	1	0	$\cdots$	1	n
keyword 2	1	0	0	$\cdots$	0	p
keyword 3	1	1	0	$\cdots$	1	n
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
keyword n	0	0	1	$\cdots$	1	p

Table 4.3: Unigram Level Dataset Representation

Classifiers; Naïve Bayes, Decision Tree and Logistic Regression provide a unigram set with likelihood, gain and coefficient values respectively. Elbow method applied on these unigram sets. Unigrams above the elbow criterion e selected to be tested on the other part of the validation dataset. Keywords which have these unique unigrams are eliminated from this validation dataset. This process is repeated for t times so t unigram sets are generated.

Each unigram set eliminated from the keywords. The keywords set which gives the maximum total click / total impression is selected. The unigram sets of these tested on the actual test set. The keywords which has these unigrams eliminated from the test set and the cost of the campaign with this new keyword set analyzed.

4.2 Classifier Based Methods at Unigram Level

Machine learning techniques applied on the unigram based data to found out the unigrams that have negative effects on the keywords. The unigram based data is the validation data that split from the train data demonstrated in Table 4.3.These techniques used are classifier based methods; naïve bayes, decision tree and logistic regression.

4.2.1 Naïve Bayes Classifier

A binary classifier which is naïve bayes learning method trained on the features extracted from K. Naive bayes classifier determined the most informative features according to their likelihood ratio. Features with value 1 represents that unigrams that exist in the keyword. Among these features the ones which has negative sign selected.Selected unigrams existence assumed to effect the campaign negatively and tested on the test data.

4.2.2 Decision Trees Classifier

An essence of decision tree classifier is built on the unigram based validation data to detect the unigrams that have negative effect on the campaign. We assume each feature as the root of the one level decision tree. One leaf represent its existence and the and the other one the opposite. Each edge ends with a sign denoting that unigram's keyword is negative or positive. An example of the tree structure is sketched on Figure 4.1.

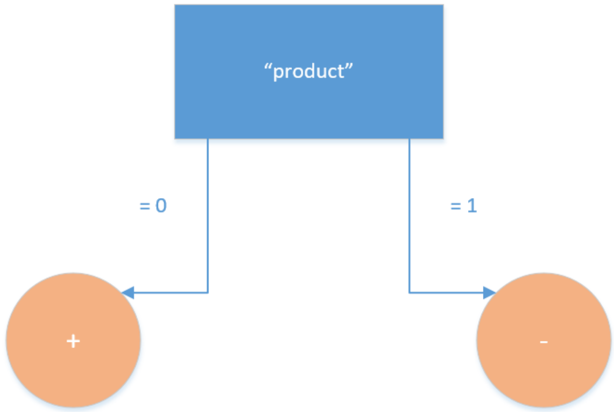

Figure 4.1: Decision Tree

We selected the unigrams, if their feature's value is 1 (which means it exists in the keyword) and their sign is negative (which means it is in a negative keyword). Then we calculated that unigram's information gain. This information gain indicates us that the existence of u in a keyword k has a negative effect on the campaign performance.

4.2.3 Logistic Regression Classifier

Logistic regression allows to establish a relationship between a binary output and a group of predictor variables which is so suitable for our data. We have a binary output of sign and features as predictor variables. Logistic regression of sign on unigrams estimate the parameter value β for each unigram. These estimated β values indicates the unigrams that are most likely exists in the negative keywords. Unigrams with large negative coefficient means that the existence of u in keyword k likely to negatively effect the campaign when it compared to the unigrams with smaller coefficient. Therefore a set of negative unigrams constitute to be eliminated from the campaign keyword set.

Chapter 5

Experimental Analysis and Discussions

5.1 Details of The Campaign Dataset

The datasets are acquired from P&G Turkey which is a leader company in FMGC sector and has online e-commerce service for its products. We have two similar datasets for two different time periods. One of them is from September 2013 and consists of 834 attributes. The other one is from October 2013 and has 1177 attributes. Here attributes are the keywords that was selected and used in past campaigns and features are the metrics related to them.

We preprocessed both datasets and eliminate some of the attributes and features. Since *Phrase Match* is the one that mostly preferred in SEA, we eliminated the attributes which has other match types. Then we discarded the attributes with missing values. Also we removes Account Name, Customer ID and Campaign from the features. As a result we have September dataset with 669 Attributes, 9 features and October dataset with 1164 attributes, 12 features. Features of September dataset includes; *Clicks, Impression, CTR, Average CPC, Average CPM, Cost, Average Position, Quality Score* and *Search Impression Share*. October dataset has the same features as September dataset but also has a few more. These are *Bounce Rate, Page/Visit* and *Percentage of New Visits*.

September	**October**
Clicks	Clicks
Impression	Impression
CTR	CTR
Average CPC	Average CPC
Average CPM	Average CPM
Cost	Cost
Average Position	Average Position
Quality Score	Quality Score
Search Impression Share	Search Impression Share
	Bounce Rate
	Page/Visit
	Percentage of New Visits

Table 5.1: Features of September and October Campaign Data

5.2 Experimental Design

We labeled the data by CP and MCPI to negative and positive. Afterwards to apply the data mining approaches we split the data in to three sets and use three fold cross validation to evaluate. 2/3 of the all data used as training set and 1/3 for testing. The training data split in to three sets for validation. Decision Tree, Naive Bayes and Logistic Regression algorithms trained on 2/3 of the validation data and then tested on the 1/3 of it. Lastly the unigram set which gives the lowest total cost tested on the test set.

Decision Tree, Naive Bayes and Logistic Regression algorithms that applied on the 2/3 of the validation data provides us gain, likelihood and coefficient values of the unigrams. Elbow method used on these values to select the unigrams that going to eliminate from keywords. Elbow criterion n set as 30, 40 and 60. Then set of unigrams composed by selecting randomly k of the unigrams that are above the elbow criterion for each n. Also k is set as 10, 20, 30. So we gathered 9 cases to compare. Each n-t combination is iterated for 100 times to implement a GRASP algorithm. As a result 100 unigram sets generated. These sets eliminated from the keywords in the validation test set. For each unigram set total cost and total impression of the keywords calculated. The unigram set which gives the lowest total click/ total impression is selected and these unigrams eliminated from the test set.

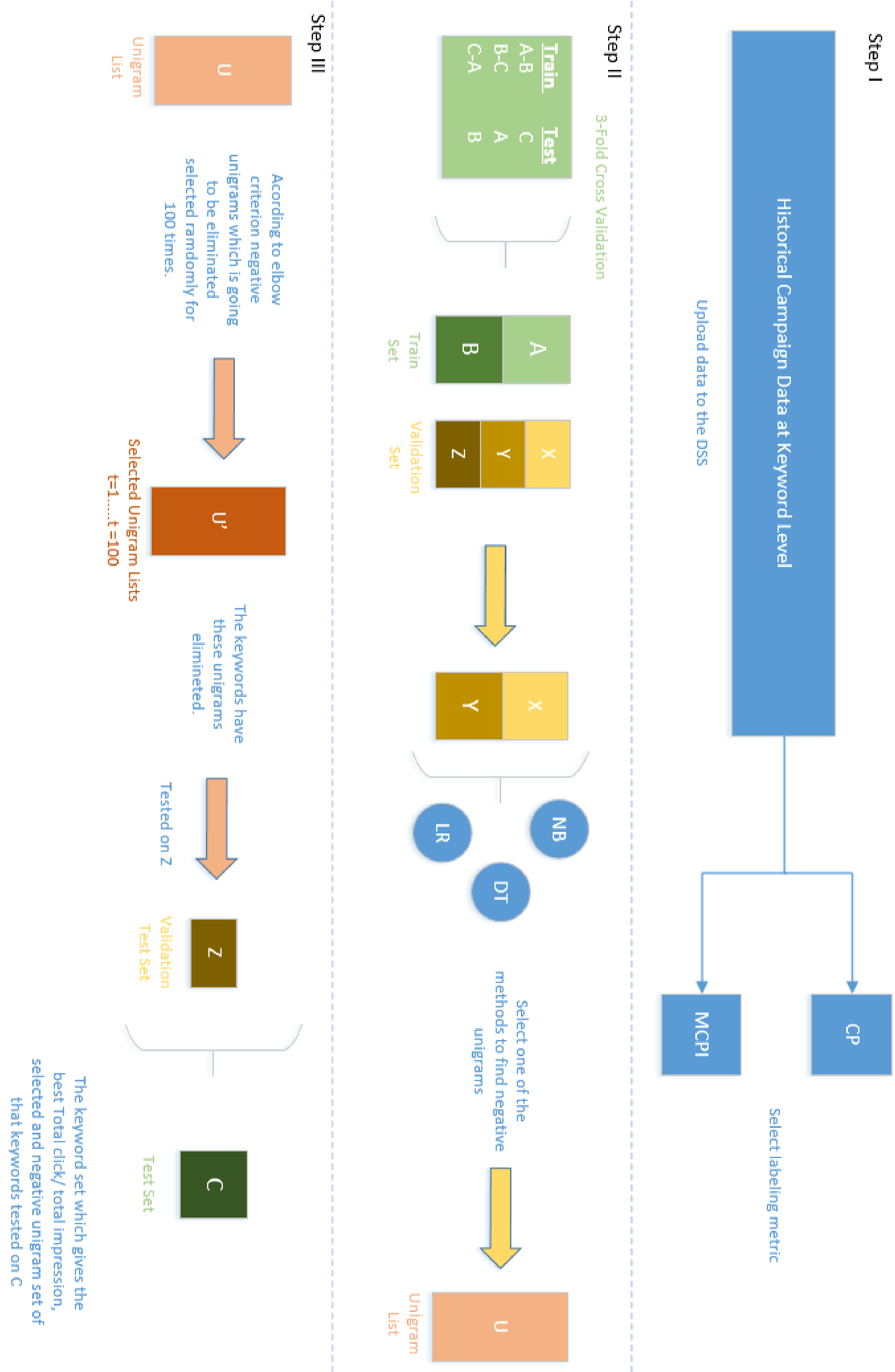

Figure 5.1: Workflow of the DSS

5.3 Results and Discussions

The results of our approach can be observed in Tables 5.3, 5.4, 5.5. Tables are categorized according to the data mining technique used on three different data which gathered from labelling. September and October is the data that labeled by CP. October MCPI and GRASP is the one that labeled by MCPI. We can not get MCPI results for September Data because it has not got Bounce Rate value to calculate MCPI

Each table shows the Total Click/ Total Impression for the resulted keywords. In addition values for each *n x k* is provided. Also the corresponding values of the original data that used in the research can be found in Table 5.2 to compare the results.

We expected to have higher total click / total impression rates but our experiment resulted in the opposite way. Total click / total impression rates are lower then the original values gathered from the original data in each algorithm (decision tree, naive bayes, logistic regression) for each *n x k* combination. Also MCPI labelling did not change this trend. The values gathered from the MCPI is lower than the original data. This means that eliminating the keywords did not accurately effect the click and impression rate as we expected. We attract less people with the new keyword set so they did not click to the ad. Which means the new keyword set is not efficiently eliminating the negative unigrams to detect the negative keywords.

When we compare MCPI and CP on October Data in all algorithms that we applied, the results from MCPI labelling gave better results that CPI labelling. Total click / total impression rates are higher than CPI labelling in all *n x k* combinations.

Among all algorithms Naive Bayes performed better on October MCPI data in all *n x k* combinations. It gave higher results than decision tree and logistic regression. There is a improvement in naive bayes when it is compared with the other ones.

However we can not observe a trend in CP labelling as in MCPI. In both September and October Datas in all *n x k* there isn't any continuity among the results. All of them is below the original results. Decision tree *30 x 10* has the best value, it is slightly below the original values.

	September	October
Total Click	2606	5074
Total Impression	130447	500191
Click/ Impression	0.019977	0.010144

Table 5.2: Attribute Values of the Data

Gain	September	October	October MCPI
30x10	0.009103	0.007453	0.008897
30x20	0.006516	0.008542	0.009066
30x30	0.007482	0.00696	0.007921
40x10	0.008935	0.007212	0.007477
40x20	0.007574	0.007783	0.008728
40x30	0.007425	0.008404	0.008092
60x10	0.007199	0.008215	0.008756
60x20	0.006841	0.007974	0.007043
60x30	0.006093	0.006977	0.007403

Table 5.3: Results for Decision Tree

Likelihood	September	October	October MCPI
30x10	0.007606	0.007647	0.008927
30x20	0.006719	0.006882	0.008189
30x30	0.00698	0.00727	0.008935
40x10	0.008185	0.007351	0.009051
40x20	0.007882	0.006908	0.007798
40x30	0.006771	0.007239	0.009567
60x10	0.007781	0.007668	0.008674
60x20	0.00632	0.007282	0.008827
60x30	0.007908	0.007205	0.007749

Table 5.4: Results for Naive Bayes

Coefficient	September	October	October MCPI
30x10	0.008598	0.007803	0.008766
30x20	0.007951	0.007379	0.007818
30x30	0.007789	0.007624	0.007289
40x10	0.007193	0.00765	0.008946
40x20	0.006761	0.00785	0.008777
40x30	0.005457	0.007793	0.007985
60x10	0.007148	0.007296	0.009174
60x20	0.00769	0.008643	0.008292
60x30	0.007075	0.006886	0.009168

Table 5.5: Results for Logistic Regression

Chapter 6

Conclusion

Search Engine Advertising is growing rapidly since 2010. It has the largest market share on internet advertising platform. It is going to be the major media for advertisers to reach their potential customers. The process is continuously evolving and getting complex. Therefore campaign managers has to examine the process closely. Domain experts has to analyze campaign reports periodically to maximize their revenue and to target the right customers. Therefore search engine advertisers aim to find the best keyword set for their campaign according to their own limitations and expectations from the search engine advertising.

Our contribution is to assist them with our overall campaign index to find the sufficient keyword set for their SEA campaigns. Current metrics for SEA is not adequate to evaluate traffic and conversion according to the other important attributes. Also there is limited research in literature on metrics to advise a solution to companies' different specifications from SEA campaigns. Thus we propose an application for the overall campaign assessment problem.

In our research first we investigate and analyze keywords according to CP and MCPI to label keywords as negative or positive. Afterwards we evaluate them on unigram level to identify the unigrams. Naive Bayes, Decision Trees, Logistic regression classifiers' core concepts used to to state that a set of unigrams are negatively or positively affecting the campaign. The unigrams which has negative effects eliminated from the keyword set.

The results gathered from three different data sets compared. Total click / total impression decreased in all datasets in three classifiers. As a result MCPI labelling slightly better than CP labelling but our results are below the orginal values.

Support Vector Machines and other machine learning techniques can be applied on the data as a future research. Also a balanced data can be gathered from the raw data and then the labelling techniques can be applied on this balanced data.

Random Sampling which is done on the data is a limitation of our research but by stratified sampling our results can be better. As a future work stratified sampling can be applied.

Bibliography

(2007). bounce rate. http://www.marketingprofs.com/7/bounce-rate-sexiest-web-metric-ever-kaushik.asp.

(2017). adrank adwords. https://support.google.com/adwords/answer/1752122?hl=en.

(2017). auction. https://www.wordstream.com/articles/what-is-google-adwords.

(2017). quality score adwords. https://support.google.com/adwords/answer/156066QSvAR.

Abhishek, V. and Hosanagar, K. (2007). Keyword generation for search engine advertising using semantic similarity between terms. In *Proceedings of the ninth international conference on Electronic commerce*, pages 89–94. ACM.

Alchalabi, A. E., Elsharnoby, M., and Khawaldeh, S. (2016). Rightscope: Detecting search campaingns positive and negative queries. In *Machine Learning and Cybernetics (ICMLC), 2016 International Conference on*, pages 290–295. IEEE.

Bucklin, R. E., Rutz, O. J., and Trusov, M. (2009). Metrics for the new internet marketing communications mix. In *Review of Marketing Research*, pages 175–192. Emerald Group Publishing Limited.

Bulut, A. (2015). Lean marketing: Know who not to advertise to! *Electronic Commerce Research and Applications*, 14(6):631–640.

Chen, Y., Xue, G.-R., and Yu, Y. (2008). Advertising keyword suggestion based on concept hierarchy. In *Proceedings of the 2008 international conference on web search and data mining*, pages 251–260. ACM.

Çoğalmış, K. (2016). *AdScope: intelligent scoping of paid search campaigns using relevance feedback*. PhD thesis.

Cui, Q., Bai, F.-S., Gao, B., and Liu, T.-Y. (2015). Global optimization for advertisement selection in sponsored search. *Journal of Computer Science and Technology*, 30(2):295–310.

Edelman, B., Ostrovsky, M., and Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *The American economic review*, 97(1):242–259.

Fain, D. C. and Pedersen, J. O. (2006). Sponsored search: A brief history. *Bulletin of the American Society for Information Science and Technology*, 32(2):12–13.

Ghose, A. and Yang, S. (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55(10):1605–1622.

Hillard, D., Schroedl, S., Manavoglu, E., Raghavan, H., and Leggetter, C. (2010). Improving ad relevance in sponsored search. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 361–370. ACM.

Hubinette, C. F. (2012). Identifying negative keywords associated with advertisements. US Patent 8,229,942.

Jansen, B. J. and Clarke, T. B. (2017). Conversion potential: A metric for evaluating search engine advertising performance. *Journal of Research in Interactive Marketing*, 11(2):142–159.

Jansen, B. J. and Mullen, T. (2008). Sponsored search: an overview of the concept, history, and technology. *International Journal of Electronic Business*, 6(2):114–131.

Jansen, B. J. and Spink, A. (2008). Investigating customer click through behaviour with integrated sponsored and nonsponsored results. *International Journal of Internet Marketing and Advertising*, 5(1-2):74–94.

Joshi, A. and Motwani, R. (2006). Keyword generation for search engine advertising. In *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on*, pages 490–496. IEEE.

Keiningham, T. L., Buoye, A., and Ball, J. (2015). Competitive context is everything: Moving from absolute to relative metrics. *Global Economics and Management Review*, 20(2):18–25.

King, M. A., Abrahams, A. S., and Ragsdale, C. T. (2015). Ensemble learning methods for pay-per-click campaign management. *Expert Systems with Applications*, 42(10):4818–4829.

Kiritchenko, S. and Jiline, M. (2008). Keyword optimization in sponsored search via feature selection. In *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, pages 122–134.

Lee, K. (2009). *The Truth About Pay-Per-Click Search Advertising*. Que.

Miller, P. (2016). *Optimizing AdWords: A Guide to Using, Mastering, and Maximizing Google AdWords*. Routledge Taylor and Francis Group.

Mukherjee, P. and Jansen, B. J. (2014). Performance analysis of keyword advertising campaign using gender-brand effect of search queries. *Electronic Commerce Research and Applications*, 13(2):139–149.

Nabout, N. A., Lilienthal, M., and Skiera, B. (2014). Empirical generalizations in search engine advertising. *Journal of Retailing*, 90(2):206–216.

Rangaswamy, A., Giles, C. L., and Seres, S. (2009). A strategic perspective on search engines: Thought candies for practitioners and researchers. *Journal of Interactive Marketing*, 23(1):49–60.

Rosso, M. A. and Jansen, B. J. (2010). Brand names as keywords in sponsored search advertising. *CAIS*, 27:6.

Rutz, O. J., Trusov, M., and Bucklin, R. E. (2011). Modeling indirect effects of paid search advertising: which keywords lead to more future visits? *Marketing Science*, 30(4):646–665.

Sculley, D., Malkin, R. G., Basu, S., and Bayardo, R. J. (2009). Predicting bounce rates in sponsored search advertisements. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1325–1334. ACM.

Silverman, D. (2018). Iab internet advertising revenue report. *Interactive Advertising Bureau. New York*.

Thomaidou, S. (2014). *Automated Creation and Optimization of Online Advertising Campaigns*. PhD thesis, Ph. D. thesis, Department of Informatics, Athens University of Economics and Business.

Thomaidou, S., Liakopoulos, K., and Vazirgiannis, M. (2014). Toward an integrated framework for automated development and optimization of online advertising campaigns. *Intelligent Data Analysis*, 18(6):1199–1227.

Thomaidou, S. and Vazirgiannis, M. (2011). Multiword keyword recommendation system for online advertising. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 423–427. IEEE.

Varian, H. R. (2007). Position auctions. *international Journal of industrial Organization*, 25(6):1163–1178.

Yang, S. and Ghose, A. (2010). Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Science*, 29(4):602–623.

Yuan, Y., Zeng, D., Zhao, H., and Li, L. (2015). Analyzing positioning strategies in sponsored search auctions under ctr-based quality scoring. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(4):688–701.