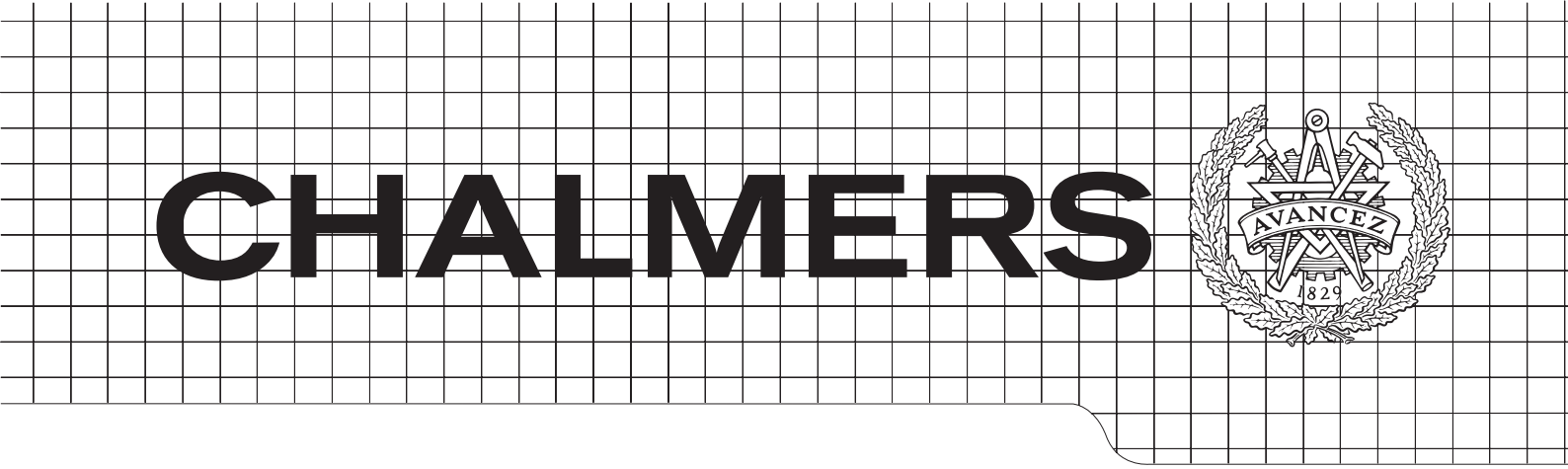

Automated analysis and planning of online marketing campaigns

Master's Thesis in Software Engineering

ERIK BRÄNNSTRÖM

Department of Software Engineering
Chalmers University of Technology
Göteborg, Sweden 2012
Master's Thesis 2012:X

Abstract

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

The Authors, Location 11/9/11

Contents

1

Introduction

As more and more people are using the Internet on a daily basis, the area of online marketing is expanding as a way for organizations to reach large audiences for a relatively small amount of money. A site that displays advertisement, the publisher, is often paid by the advertiser based on the number of times visitors see or click on the ads, but it may also be coupled with other requirements, such as that the visitor goes on to buy a product from the advertiser in a given time span. Because advertisers want to pay as little as possible while still getting good results and distributors often preferring to only show relevant information to their visitors to keep them coming back, the advertisement is typically tailored to the expected interests of the visiting user.

The material is created by the advertiser, its impact analyzed and based on these results, material is created, adapted or removed from circulation to better suit a new or existing target audience. This circular process requires a lot of manual labor since the analysis results must be understood and applied to the context of the material as well as the target group. There is a lot of data that needs to be correlated, a task that is well suited for computers.

On Duego, a social networking company founded in 2010, online advertisement is leveraged as a way to promote the site to new users. The process of managing the online marketing data follows the manual process described above. The goal of Duego and this thesis is to automate parts of this process using a software solution.

The system is required to analyze existing campaigns with regard to ads and target groups along with campaign metrics, for example the number of times an ad is shown. This data is then used as a basis for suggesting new ads. The data set that will be used in production consists of hundreds of different campaigns, each with a large number of ads. Along with the attributes of both campaigns and ads, this means that the complete advertisement data set consists of hundreds of thousands of possible attribute combinations that need to be analyzed.

A generalization of this problem is knowledge discovery in databases (KDD) and,

more specifically, data mining. Data mining is the process of delegating and automating the task of identifying knowledge that by some definition is useful in a large set of data to a computer, either fully or partially.

The research question for this thesis is how to use historical online marketing data to automate the creation of new advertisement campaigns. The answer to this question is presented as a procedure that is applicable to online advertisers. The following subsections will describe terminology from the fields of automation and marketing respectively. This is followed by a description of the problem this thesis addresses and finally a limitation of the scope of this paper.

1.1 Automation

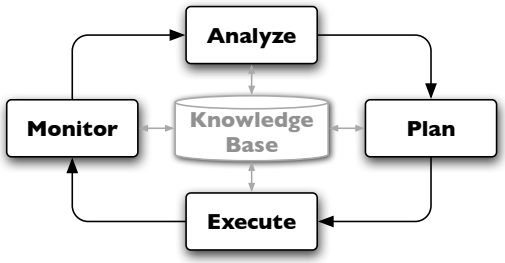

Figure 1.1: General MAPE control loop

Useful terminology is defined by IBM [2006] in the field of autonomic computing. Figure 1.1 shows an autonomic manager, which is a component that collects data from a system and, based on this data, performs actions with the purpose of improving the system. This control loop is divided into four subtasks called monitor (collect system information), analyze (correlate and model data), plan (design behavior required to reach goal) and execute (run the planned actions), sometimes referred to as MAPE. Each subtask can optionally interact with a knowledge base for storing and retrieving data. This will then be applied to online marketing.

1.2 Marketing

A number of marketing terms will be used throughout this paper. A *marketing campaign*, or simply a *campaign*, is comprised of one or more advertising messages, *ads*, that are directed to one defined audience, the *target* or *target group*. Any ad, and by extension campaign, have numeric measures of success called *metrics*. The most commonly used are *impressions*, which is the number of times an ad has been shown, and clicks. In this paper however, the word *action* will be used when referring to an user interaction, since one might be interested in other values than simple clicks, such as for example the number of users who go on to register at the site after clicking. Ads have a number of

properties, which depend on the media that is used. For example, textual ads in search engines typically have a title, a short text and a URL to which the user is redirected upon clicking the ad. A target is defined based on the options available of the advertisement type used. A graphical description of these terms is shown in figure 1.2.

Four such classes are identified for the purpose of this paper based on the way the ads are adapted based on the user. *Search* advertising uses the user's search terms, *social* advertising uses demographic and personal data, *contextual* advertising finds keywords on the page on which the ads are displayed or uses manual categorization and finally *non-contextual* advertising, which does no relevancy matching. The separation between these classes is not necessarily distinct and a single publisher can use targeting criteria from different classes.

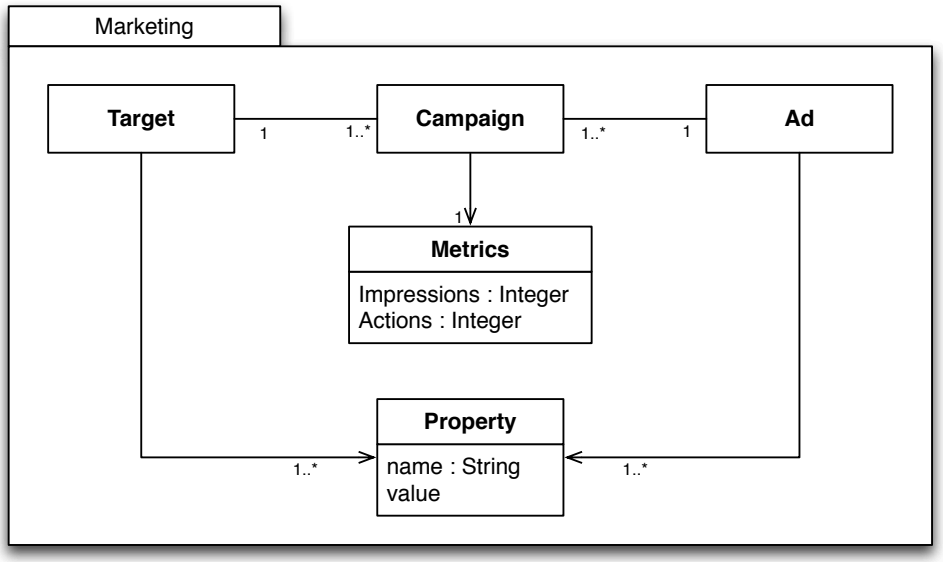

Figure 1.2: UML description of marketing terminology.

As the word *user* is typically used to refer to a person interacting with either the web site that is being marketed or the web site on which the advertisement is shown, we will use the word *operator* when discussing a person interacting with the system that is described in this paper to avoid confusion.

1.3 Automated marketing

By merging the fields of automation and online marketing we put the automation terminology in context. Monitoring is the collection of metrics for online advertisement. To optimize these metrics, the gathered information is analyzed and this analysis forms the basis for the planning of future campaigns. Once the plans are completed, the new campaign can be launched, with new metrics being gathered and so on.

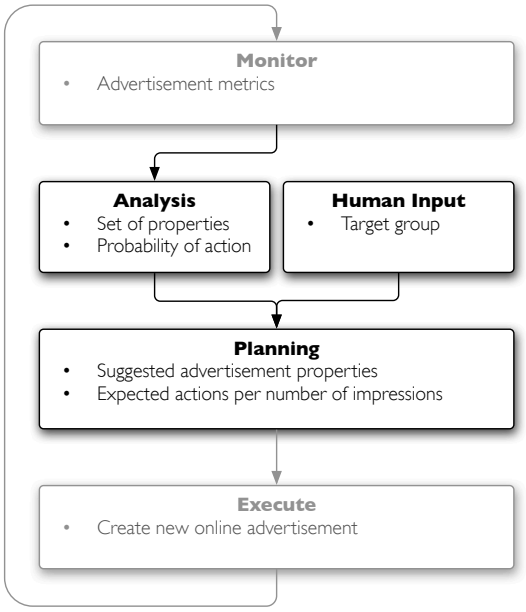

Figure 1.3: Online marketing automation system in control loop

Monitoring is commonly automated already, either using custom software or services such as Google Analytics, where as the other parts of the process are performed manually. It is infeasible to fully automate the whole process, due to for example the creative side of advertisement and the complex factors that decide which groups an upcoming campaign should target. There are however certain areas that can be automated and this paper will focus on automated analysis and planning of online marketing campaigns, shown in Figure 1.3.

A system to automate this process requires monitored data as input, which in this context equals historical data of campaigns and their metrics as mentioned previously. This data is then mined to identify subsets of campaign properties that are associated with a probability that an impression of that an ad with these properties leads to an action being taken by the user. Human input is required to specify which target group the next campaign will be aimed at, and based on this the system will generate suggested ad properties that optimize the number of expected actions taken per impression.

1.4 Scope

Using the MAPE framework in Figure 1.1, only the analysis and planning tasks are considered part of this paper, where as monitoring and execution are out of scope. The latter two are however relevant in the verification step, but will not be covered as a research topic. In this context, this means for example that the feature of integrating this system with marketing services to automatically add new ads and campaigns will

not be a part of the final system.

Furthermore, the data set will likely include attributes whose values are free text and images. Text mining and image recognition are beyond the scope of this project. Instead these attributes will be manually categorized, so that the value space is discrete and finite.

1.5 Foundations

A number of high-level descriptions of frameworks for knowledge discovery in databases exist [Fayyad et al., 1996, Frawley et al., 1992] and they exhibit a number of commonalities. These include the importance of having a knowledgeable human operator guiding the process in terms of supplying domain knowledge to the system formulating the goal of the knowledge discovery; feeding discovered knowledge back into the system; and the identification and application of a discovery method, or more specifically the data mining algorithms.

In data mining, the input to a system can be described using the terms *concepts*, *instances* and *attributes*, where concept is the actual result of the mining, i.e. what we want to be learned; an instance is one single example of data to be mined and can be compared to a row in a database; and attribute is a property of an instance, which in the database analogy is a column [Witten et al., 2011].

Instances are the smallest component of data input to a mining system, but its low-level description of knowledge may not be what is most useful. Chen et al. [1996] give an overview of the field of data mining where they mention the aspect of multi-level data mining, which states that correlations may not commonly exist on the lowest level of granularity, but instead by forming groups of related items. An example given would be that a specific brand of milk does not necessarily imply the purchase of a specific brand of bread, however purchasing milk of any kind may still be correlated to the purchase of bread irrespective of brand.

In a highly influential paper, Quinlan [1986] decribe how decision trees can be created from a training set and how well it handles the problem of unknown attributes values and noisy data. A related paper, Quinlan [1987a], deal with how generated decision trees can be simplified in order to more easily be applied. Four different methods are evaluated, one of which is the reformulation of the tree as a set of production rules. This specific topic is further analyzed in Quinlan [1987b] where such production rules are shown to be more compact and also in many cases improve the classification of unseen data. An added positive effect is that production rules from separate classifications can be merged more efficiently than their original decision trees.

Cased-based reasoning is a methodology that is similair in many ways to the automation framework described previously, but it is more closely related to human cognition. The basic concept as described in Watson [1999] is that in order to solve a new problem, the first step is to identify (retrieve) existing problems that are similar. Select a solution to one of the retrieved problem and apply it to the current problem (reuse). Adapt (revise) the existing solution to better match the problem at hand, if necessary. Finally,

store (retain) the problem and its solution if it was successful.

2

Related works

Richardson et al. [2007] describe a model for predicting the click-through rates of ads given a set of properties in the context of search engine advertisement. Though it is a relevant subject, the paper is limited to the specific class of search advertising, and is therefore not directly applicable to this paper.

In online advertisement there have been problems of so called click-spam or click-fraud, meaning that the number of clicks on ads is increased in a fraudulent manner, for example using bots. Dave et al. [2012] provides a methodology to measure click-spam in their networks as well as a study of the severity of the problem on different classes of networks. Their results show that established search advertising on for example Google and Bing is fairly accurate in their filtering of click-spam, where as the problem is greater for contextual and social advertisement, and severe for mobile advertisement. A related paper by Zhang et al. [2011] describes a methodology for advertisers to evaluate the quality of click traffic and use this to assess the difference in quality between bulk traffic vendors and established pay-per-click networks.

Research has also shown that using targeted advertisement is effective given that the ad is not experienced as being obtrusive, though obstrusive ads that are not targeted also increase performance [Goldfarb and Tucker, 2011]. This promotes the idea of more effective targeting but may also be important to understand in how the solution presented in this thesis is applied.

Automation in marketing is not a completely novel idea. Joshi and Motwani [2006] present a technique for automatically finding related keywords for broader targeting of ads by creating a graph of search terms based on the domain knowledge contained in search engines. Continuation of that work is that of Thomaidou and Vazirgiannis [2011], where keywords are also extracted from the landing page to further improve the results. Both approaches are however only relevant for contextual and search engine advertisement.

3

Method

The solution described in this paper is the result of thourough analysis of existing data sets as well as experiments. The method used to reach this solution is described in this chapter.

4

Solution

The solution has been divided into one section each for the two subtasks identified in the problem description. Analysis covers the problem of creating an appropriate model of the available data that is used in the next stage. Planning, in turn, is the utilization of that model with the purpose of outputting useful information to the operator. Finally, this is followed by the steps taken to evaluate the results.

4.1 Analysis

The first step in the process is analysis, and the first substep is how to deal with the input data. The data contains a set of metrics which are not best used in their raw form. The analysis step therefore calculates a confidence interval on the mean of the action rate, since it is more helpful in later processing steps than the more commonly used average action rate which do not take the size of the sample space into account.

The reason for specifying a confidence interval is to account for the inherit variance of values such as the action rate. Even though a campaign has been shown millions of times, there is still some degree of uncertainty to what the action rate would be if it was shown another million times. A $100(1-\alpha)\%$ confidence interval is defined as the interval of values that a random variable X will take with a probability of $1-\alpha$, or formally

$$P[L_1 \leq X \leq L_2] = 1 - \alpha \tag{4.1}$$

The standard function for calculating the values of the interval limits L_1 and L_2 is shown in equation 4.2.

$$L = \bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n} \tag{4.2}$$

The confidence interval is calculated based on the assumption of a normal distribution of the action rate. While the underlying distribution of each ad impression might not be

normal, the Central Limit Theorem resolves that problem. The theorem defines that for a sample $X_1, X_2, ..., X_n$ where n is large (empirically shown to work for n as low as 25), $\bar{X}$ is approximately normal with mean μ and variance σ^2/n [Milton and Arnold, 2002].

At first look there is a problem with the use of a normal distribution. Because the variance is unknown and therefore needs to be estimated, this means that Student's t-distribution is more appropriate and that equation 4.3 should be used. However, because of the fact that $t_{\alpha/2} \to z_{\alpha/2}$ as the degrees of freedom approaches infinity (at 2000 degrees of freedom, $t_{0.025} \sim z_{0.025} \sim 1.96$ when rounded to two decimals), they are for all intents and purposes equal since we are typically working with impressions that quickly generate high degrees of freedom.

$$L = \bar{X} \pm t_{\alpha/2} S/\sqrt{n} \tag{4.3}$$

The distribution was also empirically tested using plotting to see that it follows the well-known bell curve of normal distributions as well as using the empirical rule, which states that for a normal distribution, about 68/95/99.7 percent of the values are respectively within 1/2/3 standard deviations of the mean.

After this processing, the data is added to the knowledge base for future use. The knowledge base will store all data relating to the marketing campaigns to increase the amount of available data to base decisions on.

Following this, additional data sets are created to support the planning phase. The idea is that ad properties are dealt with not only in combination but also as stand-alone entities to gain further insight into what makes ad work. Each ad property in every campaign instance from the data set is stored separately along with the campaign's metrics (i.e. impressions and actions). If an identical ad property value already exists, the metrics are aggregated so they reflect all occurrences of that property value.

Image	Text	Impressions	Clicks
Image A	Text C	1000	5
Image B	Text C	1200	19
Image A	Text D	700	4

Figure 4.1: Data set example

Figure 4.1 shows an example of how the original data set could look like. The campaigns do not include any form of targeting and thus only consist of the ad properties image and text, as well as the metrics gathered, in this case impressions and clicks.

The second step of aggregating metrics for individual properties requires both impressions and clicks to be summed for each value of every ad property. For example, the data set for images would have one instance for Image A which has 1700 impressions and 9 clicks. The end result is shown in figure 4.2.

Image	Impressions	Clicks
Image A	1700	9
Image B	1200	19

Text	Impressions	Clicks
Text C	2200	24
Text D	700	4

Figure 4.2: Aggregating metrics in data set for each property value

For advertisers who also use targeting which is likely to impact the choice of ad properties, these additional data sets must be related to their respective targets and the aggregation of metrics should take this into account. For example, if the advertiser runs campaigns with images that target male and female users separetly, there will be one data set for images used on males and one data set for images used on females. An image used to target both sexes will thus appear in both of these data sets, but the metrics may be different.

4.2 Planning

The goal of the planning step is to formulate suggestions for upcoming campaigns. This is done by selecting values for each ad property that has an average action rate that lies above the mean of the property. Using the example in figure 4.2, the mean is $(9 + 19)/(1700 + 1200) \approx .0097$ for images and $(24 + 4)/(2200 + 700) \approx .0097$ for texts, so Image B and Text C are selected for combination.

If the advertiser is running campaigns with different target groups, the basis for the planning must only be properties that are used with the specific target groups decided upon by the operator.

For every combination, an estimate of the real action rate is created. In the case of the above example, there is already an existing campaign in the knowledge base that is identical, so the actual action rate of that ad (or the average of all identical campaigns, if there are more than one) is used as an estimate.

An interesting and likely problem arises when the combination has never been used before. Once again using the example data set, the combination Image B and Text D does not exist previously so a method for estimating the action rate for such campaigns must be defined.

Existing variables to base the estimate on are the metrics for each property, but they are not enough by themselves since one property may carry a stronger influence on the actual result than another. The Pearson coefficient of correlation is such a measure and

it is defined as

$$\rho = \frac{Cov(X,Y)}{\sqrt{(VarX)(VarY)}} \tag{4.4}$$

Because the covariance and variance are unknown they have to be estimated based on the sample.

$$\hat{\rho} = \frac{\widehat{Cov(X,Y)}}{\sqrt{(\widehat{VarX})(\widehat{VarY})}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{(\sum_{i=1}^{n}(X_i - \bar{X})^2)(\sum_{i=1}^{n}(Y_i - \bar{Y})^2)}} \tag{4.5}$$

Because $-1 \leq \hat{\rho} \leq 1$, where a value of 1 or -1 means there is an exact linear correlation between the two random variables and 0 means there is no linear correlation at all, values for the correlation of X_1 and X_2in relation Y_0 are comparable. This is useful to get an understanding of which variable most strongly influences the value of the dependent variable and also whether that influence is positive or negative.

Another possible solution to identifying the correlation is to use multiple linear regression. The general linear model is defined as

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{4.6}$$

where μ is the mean of Y given values for its dependent variables [Milton and Arnold, 2002]. The β-values define the change in μ for each unit of change in the related x, and could as such be used as some form of correlation measure. This approach would be useful if the linear regression model could provide good estimates for the dependent variable's value, but empirical analysis on this project showed a much too great variance for the model to be useful for this purpose. As such, using linear regression cannot provide any additional value than the simpler Pearson coeffiecient.

The variables available for basing the estimation for the combination of two different properties X and Y are thus the low and high values of the action rate confidence interval, refered to as x_{low}, x_{high}, y_{low} and y_{high} respectively, as well as the correlation coefficient for each attribute, ρ_x and ρ_y. The idea behind the estimation function is to sum the average action rate of each property, which is weighted depending on how much the property tends to influence the result, and normalize that sum. Using the previously mentioned variables, the estimation function is described as the following inequality.

$$\frac{\rho_x x_{low} + \rho_y y_{low}}{\rho_x + \rho_y} \leq z \leq \frac{\rho_x x_{high} + \rho_y y_{high}}{\rho_x + \rho_y} \tag{4.7}$$

Finally, the estimates are used to rank suggested campaigns based on the likelihood of resulting in high action rates. This is to provide decision support for domain experts who handle the final execution of marketing campaigns.

4.3 Evaluation

For the system to be useful to the advertiser, the final estimates for suggested campaigns must be accurate enough so they can be trusted by the domain experts making decisions on which campaigns to run. As the combining of possibly unrelated ad properties may give rise to more uncertainty than what existed when those same properties were considered separatly, the estimation function needs to be evaluated.

The system is tested by using using the available infrastructure for managing marketing campaigns at Duego. The test consists of selecting ten suggested campaigns to run live. After each campaign has gathered at least half a million impressions, the final mean action rate is calculated and compared to the original estimate. If a 90% confidence interval on the action rates used as the basis for estimation is chosen, and assuming this confidence interval is applicable to the estimates as well, there is a 10% chance that the mean action rate from the test is not in the interval. The results are considered to be validated if no more than two of the ten campaigns fall outside of the estimated intervals.

5

Discussion

5.1 Risk factors

Because the output of the system is dependent on historical data, an assumption has been made that older data is still representative of the current state of the marketing. This is definitely not true for campaigns that are adapted for Christmas, Valentine's Day or other special occasions. The assumption is that the amount of such time-dependent data is so small that it will not impact the final results. If however there is reason to believe that this set of data would influence the output, the recommended approach is to remove it from the knowledge base.

5.2 Future work

This paper uses a very simple similarity measure for the targeting because it is not a central part of the thesis. For advertisers with many campaigns with differing targeting it would probably be interesting to develop a more advanced form of measure. Certain attributes may not have a binary relation (related or not related) but rather they may themselves have a similarity score. An example of this could be age, where the similarity score in this paper would be zero for two targets of age 19 and 20 respectively, though in reality it seems likely that these groups would be attracted to the same properties in an ad.

6

Conclusion

Empty.

Bibliography

Ming-Syan Chen, Jiawei Han, and Philip S. Yu. Data Mining: An Overview from a Database Perspective. In *IEEE Transactions on Knowledge and Data Engineering*, volume 8, pages 866–883, 1996.

V. Dave, S. Guha, and Y. Zhang. Measuring and fingerprinting click-spam in ad networks. *Proceedings of the Special Interest Group on Data Communication (SIGCOMM)*, 2012.

Usama Fayyad, Gregory Piatetsky-shapiro, and Padhraic Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, August 2-4, 1996*, pages 82–88. AAAI Press, 1996.

William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. Knowledge discovery in databases: An overview. *AI Magazine*, 13(3):57–70, 1992.

A. Goldfarb and C. Tucker. Online display advertising: Targeting and obtrusiveness. 2011.

IBM. An architectural blueprint for autonomic computing. *Quality*, 36(June):34, 2006.

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

J.S. Milton and J.C. Arnold. *Introduction to probability and statistics: principles and applications for engineering and the computing sciences*. McGraw-Hill, Inc., 4th edition, 2002.

J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986. ISSN 0885-6125.

J.R. Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987a.

J.R. Quinlan. Generating production rules from decision trees. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, volume 30107, pages 304–307. Citeseer, 1987b.

M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530. ACM, 2007.

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

I. Watson. Case-based reasoning is a methodology not a technology. *Knowledge-Based Systems*, 12(5):303–308, 1999.

Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 3rd edition, 2011.

Q. Zhang, T. Ristenpart, S. Savage, and G.M. Voelker. Got traffic?: an evaluation of click traffic providers. In *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality*, pages 19–26. ACM, 2011.