Evaluation Metrics for Persuasive NLP with Google AdWords

Marco Guerini, Carlo Strapparava, Oliviero Stock

FBK-Irst
Via Sommarive 18, Povo, I-38100 Trento
{guerini, strappa, stock}@fbk.eu

Abstract

Evaluating systems and theories about persuasion represents a bottleneck for both theoretical and applied fields: experiments are usually expensive and time consuming. Still, measuring the persuasive impact of a message is of paramount importance. In this paper we present a new "cheap and fast" methodology for measuring the persuasiveness of communication. This methodology allows conducting experiments with thousands of subjects for a few dollars in a few hours, by tweaking and using existing commercial tools for advertising on the web, such as Google AdWords. The central idea is to use AdWords features for defining message persuasiveness metrics. Along with a description of our approach we provide some pilot experiments, conducted both with text and image based ads, that confirm the effectiveness of our ideas. We also discuss the possible application of research on persuasive systems to Google AdWords in order to add more flexibility in the wearing out of persuasive messages.

1. Introduction

Evaluating systems and theories about persuasion is becoming more compelling as the field of automated message generation grows, e.g. (Fogg, 2009; Guerini et al., 2008). Measuring the persuasive impact of a message is of paramount importance in this context. Evaluation experiments represent a bottleneck for the field: they are expensive and time consuming, and recruiting a high number of human participants is usually very difficult.

In this paper we present a new "cheap and fast" methodology to overcome this bottleneck. This methodology allows conducting experiments with thousands of subjects for a few dollars in a few hours, by tweaking and using existing commercial tools for advertising on the web, such as Google AdWords.

Approaches to NLP that rely on the use of web tools have recently emerged. For example Amazon Mechanical Turk has been used for collecting annotated data useful in many NLP tasks (Snow et al., 2008). Another example is re-CAPTCHA, a free CAPTCHA service that helps to digitize books, newspapers and old time radio shows.

The paper is structured as follows: section 2 presents the main AdWords features, while section 3 describes how these features can be used for defining message persuasiveness metrics. Section 4 describes some pilot experiments to test the feasibility of our approach. Section 5 discusses the possible application of research on persuasive systems to Google AdWords in order to add more flexibility in the wearing out of persuasive messages. Section 6 presents conclusions and further ideas of our approach that can be implemented with other Google tools.

2. AdWords features

Google AdWords is Google advertising program. The main idea is to let advertisers display their ads only to relevant audiences by means of keyword-based contextualization on the Google network. Google network is divided into:

- **Search network**: Includes Google search pages, search sites and properties that display search result pages, such as Froogle and Earthlink.
- **Content network**: Includes news pages, topic-specific websites, blogs and other properties - such as Google Mail and The New York Times.

When a user enters a query - like "cruise" - in the Google search network, Google displays a variety of relevant pages, along with ads that link to cruise trip businesses. In order to be displayed, these ads were explicitly associated with relevant keywords selected by the advertiser. An example is given in Figure 1 (search results are blurred to highlight the ads).

Every advertiser has an AdWords account that is structured like a pyramid. Each level has its own components:

- At the top level there is the account: unique email address, password, etc.
- Campaign: start and end dates, daily budget, target languages and locations, etc.
- Ad group: ads, keyword and/or placement list and CPC (cost-per-click) or CPM (cost-per-thousand impressions) bids.

In this paper we focus on ad groups and CPC bids. Each grouping gathers similar keywords together - such as by a common theme - around an ad group. For example, if the campaign goal is to sell coffee beans, ad groups might include the following keywords:

Gourmet Coffee Beans	**Organic Coffee Beans**	**French Roast Beans**
Keywords:	Keywords:	Keywords:
Speciality coffee	Organic coffee beans	Decaf French roast coffee
Gourmet coffee	Decaf organic coffee	French roast coffee beans
Gourmet coffee beans	Natural coffee	French coffee beans

Table 1: Example of keywords for different ad groups

For each ad group, the advertiser sets a CPC bid. The CPC bid refers to the amount the advertiser is willing to pay for a click on his ad. The cost of the actual click is instead based

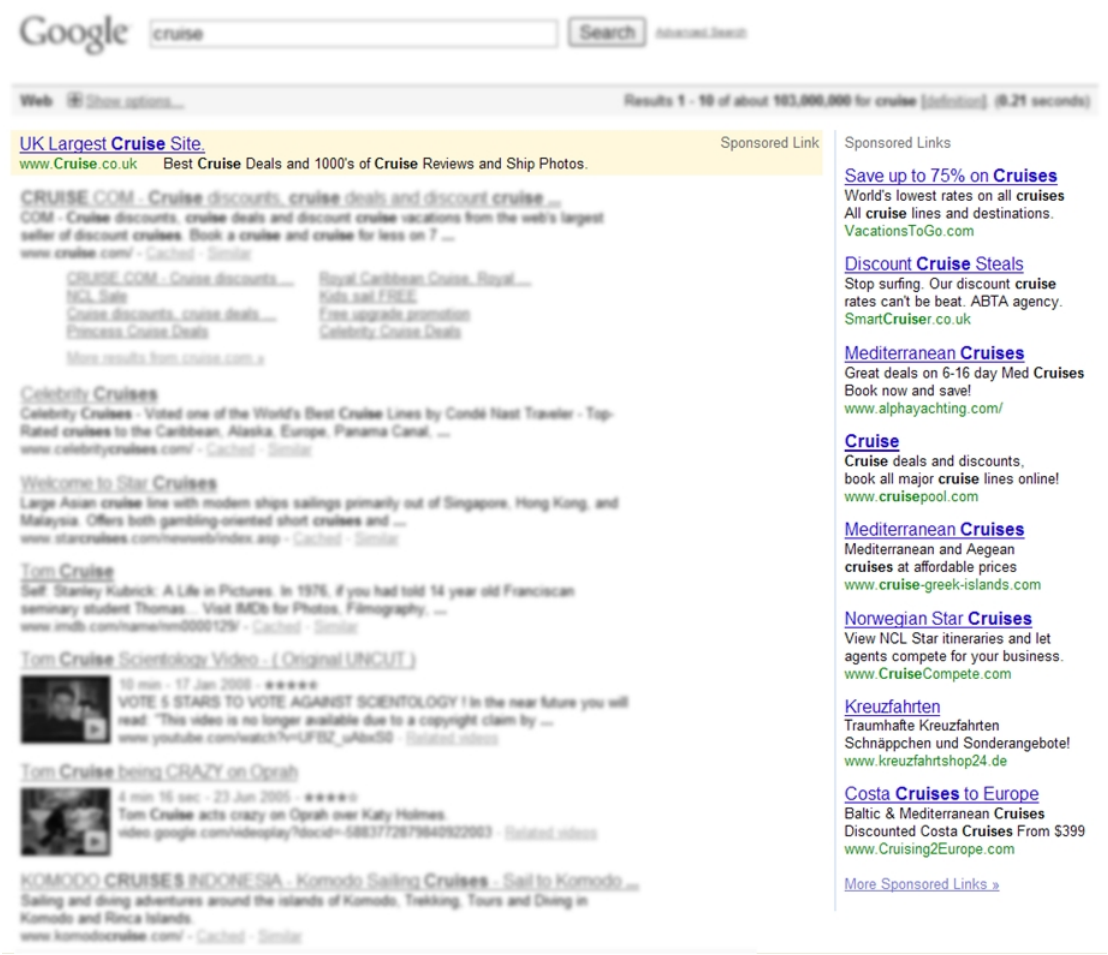

Figure 1: Search result page for "cruise"

on its quality score (a complex measure out of the scope of the present paper).

For every ad group there might be multiple ads to be served (displayed), for example the "Gourmet coffee beans" group, in Table 1, can have ads (A) "*Tasty Gourmet coffee beans*", (B) "*Cheap Gourmet coffee beans*", etc., competing with each other. Ads from the same advertiser cannot appear at the same time on the same search, i.e. only one ad per time is displayed. It is possible to choose between two different ad serving options:

1. **Optimise**: Over time, the system determines which ad in the group is performing better, based on historic click-through rates (CTRs) and Quality Scores. Based on this data, the higher performing ads will be displayed more often.
2. **Rotate**: This option will serve all the ads in a group more evenly on a rotating basis, regardless of their performance.

There are many AdWords measurements for identifying the performance of each single ad (its "persuasiveness" from our point of view):

- **CTR**, ClickThrough Rate: measures the number of clicks divided by the number of impressions that the ads have received (number of impressions is the number of times an ad has been displayed in the Google Network).
- **Conversion Rate**: how many user clicks turned into actual conversions for the advertiser. Conversion rate equals the number of conversions divided by the number of ad clicks.
- **ROI**: if someone clicks on an ad, and buys something on your site, that click is a conversion from a site visit to a sale. Other conversions can be page views or signups. By assigning a value to a conversion the resulting conversions represents a return on investment, or ROI.
- **Google Analytics Tool**: Google Analytics is a web analytics tool that gives insights into website traffic, such as: number of visited pages, time spent on the site, location of visitors, etc.

So far, we have been talking about text ads - Google's most traditional and popular ad format. In addition there is also the possibility of creating the following types of ads:

- Image (and animated) ads
- Video ads

- Local business ads
- Mobile ads

The above formats allow for a greater possibility of investigating the persuasive impact of messages (in addition to text-based).

3. Evaluation and Targeting

Evaluation of the effectiveness of persuasive systems is very expensive and time consuming, as the STOP experiment showed (Reiter et al., 2003): experiment design, subjects recruitment, make them take part to the experiment, dispense questionnaire, data collection, data analysis, etc.
AdWords can be used to design and develop various metrics for fast and semi-automated evaluation experiments. Note that this is an uncommon use of the tool, which is built to automatically optimize and maximize the performance of the campaign, rather than testing scientific hypotheses that potentially require also keeping poorly performing conditions around.
Let us hypothesize that we designed an experiment with 3 conditions. First we create an ad group with 3 competing messages (one message for each condition). Then we choose the serving method (in our opinion the "rotate" option is better that "optimize" as it guarantees subject randomness and is more transparent) and the context (language, network, etc.). Then we need only activate the ads and wait. As soon as data are collected we can evaluate the conditions:

- **Basic Metrics**: a higher CTR score indicates which message is best performing. *It indicates which message has the highest initial impact.*
- **Google Analytics Metrics**: measures how much the messages kept subjects on the site and how many pages have been viewed. *It indicates interest/attitude generated in the subjects.*
- **Conversion Metrics**: measures how much the messages converted subjects to the final goal. *It indicates complete success of the persuasive message.*
- **ROI Metrics**: by creating specific ROI values for every action the user perform on the landing page. The more relevant (from a persuasive point of view) the action the user performs, the higher the value we must assign to that action.

In our view combined measurement are better: for example, there could be cases of messages with a lower CTR but a higher conversion rate.
AdWords allows for very complex targeting options that can help in many different evaluation scenarios:

- Language (see how a message's impact can vary in different languages)
- Location (see how a message's impact can vary in different cultures sharing the same language)
- Keyword matching (see how a message's impact can vary with users having different interests)
- Placements (see how a message's impact can vary among people having different values - e.g. the same message displayed on Democrats or Republican sites).

4. Pilot Experiments

We conducted some preliminary studies to test the feasibility of our approach, focusing on the use of irony in advertisements. One of the authors provided marketing consulting to a publishing company, so we had the opportunity to run our experiments within a real promotion campaign. The campaign involved the promotion of (i) a book and (ii) the corresponding book series. Ads for the book were text based, ads for the book series were image based.
The book has the ironic title "Diversamente Occupati" (tr. "differently employed") that mocks the Italian phrase "diversamente abile" (translation: "differently able"), a politically correct word for handicapped persons. In this case the invented word "Diversamente Occupati" ironically refers to temporary employed (i.e., "differently employed") people, generally considered a less than ideal state of employment. The book series is called "Resistenza Umana" (translation: "Human Resistance") and deals with ironical manuals to "survive office stress".
All the experiments we conducted had a between-subject design with two conditions:

- a control condition, with a neutral message
- an experimental condition with an ironical message

The landing site was the same for all the conditions (the site of the book series), but the landing pages were different: the book web-page for text ads, the homepage for image ads.
The metrics we considered were: CTR and Google Analytics metrics (Conversion metrics were not available yet at the time of the experiments).

4.1. Experiments in the Search Network with text ads

The first two experiments were run in the Search Network in Italian, using the keyword search "precari" (translation: "temporary employed") in broad match mode to display ads to a relevant audience. The ads have the same bodytext, differing only in the headline (see Figure 2).

First Experiment
Duration: 24 h
Total Cost: $\sim$1.2 euros (CPC 0,04 euros)
Number of subjects: $\sim$3000
Conditions: 1A "Sei Precario?" (translation: "Are you temporarily employed?", the control condition) 1B "Sei Diversamenteoccupato?" (translation: "Are you differently employed?", the experimental condition)
Results: CTR (1A) = 1,7% CTR (1B) = 0,5%

Discussion: the test is significant (χ^2 = 9,03; 1 degree of freedom; p <0,01), condition 1A performs better than 1B - three times the clicks -. Should we conclude that irony performs worse than non-ironic messages, or are the results biased by the fact that the word "Precario" in condition 1A is automatically highlighted by Google (since it matches the query search keyword)? To understand this we ran a

Figure 2: Ads used in experiment 1 and 2

Condition	Clicks	No Clicks	Total
1A	24	1410	1434
1B	7	1364	1371
Total	31	2774	2805

Table 2: First experiment results

second experiment.

Second Experiment
Duration: 24 h
Total Cost: ∼1.5 euros (CPC 0,04 euros)
Number of subjects: ∼4000
Conditions: 2A '"Sei Disoccupato?" (tr. "Are you unemployed?", the control condition) 2B "Sei DiversamenteOccupato?"(the experimental condition)
Results: CTR (2A) = 1,2% CTR (2B) = 0,8%

Condition	Clicks	No Clicks	Total
2A	23	1990	2013
2B	14	1889	1903
Total	37	3879	3916

Table 3: Second experiment results

Discussion: there is no statistically significant difference between conditions 2A and 2B (χ^2 = 1,83; 1 degree of freedom; p = 0,17). It is reasonable to conclude that in the previous experiment the difference was given by the term highlighted by Google. What about irony? At first glance it does not seem to boost the performance of the message (CTR comparison between 2A and 2B), but using the Google analytics tool we found that condition 2B outperformed the other conditions (1A, 1B, 2A) in terms of permanence of the users on the landing site (∼3 min vs ∼1 min). We can conclude that having capitalized letter "O" in "DiversamenteOccupati" helped the user to spell the word and get the irony. Irony itself then induced more interest in the subjects that spent more time browsing the website.

4.2. Experiments in the Content Network with image ads

We conducted two other experiments on the Content Network with image ads to understand what impact the use of pictures in the message, and in particular context (i.e. websites), has on persuasiveness. Below the two conditions are presented (format 300x250). C1, the control condition, contains only a sentence, while C2, the experimental condition, contains the same sentence plus an ironic image. The translation of the sentence is "Does the office wear you out?":

Figure 3: Ads used in experiment 3 and 4

Third experiment
Duration: 12 h
Total Cost: ∼7 euros (CPC 0,6 euros)
Context: a single site of job offers on the Content Network, Italian language
Number of subjects: ∼4200
Results: CTR (C1) = 0,5% CTR (C2) = 0,1%

Discussion: the test is significant (χ^2 = 5,87; 1 degree of freedom; p <0,01), condition C1 performs better than C2 - five times the clicks -. Should we conclude that the image has a negative impact on the message? To understand this we conducted a fourth experiment.

Condition	Clicks	No Clicks	Total
C1	10	2016	2026
C2	2	2142	2144
Total	12	4158	4170

Table 4: Third experiment results

Fourth experiment
Duration: 1 week
Total Cost: ∼2 euros (CPC 0,2 euros)
Context: on the whole Content Network, Italian language, specific keywords used to individuate relevant websites
Number of subjects: ∼12000
Results: CTR (C1) = 0,16% CTR (C2) = 0,08%

Condition	Clicks	No Clicks	Total
C1	14	8493	8507
C2	3	3652	3655
Total	17	12145	12162

Table 5: Fourth experiment results

Discussion: there was not a statistically significant difference between conditions C1 and C2 (χ^2 = 1,25; 1 degree of freedom; p =0,26). We can reasonably conclude that in the previous experiment the image did not have a negative impact by itself, rather it had a "positive" impact in cutting uninterested users, by helping them in "disambiguating" the message (the users were looking for a job, not for amusement).

5. Tailoring

Research on persuasive systems (Guerini et al., 2008; Strapparava et al., 2007) can help in adding more flexibility to AdWords. Some of the present shortcomings are:

- Ad tailoring is performed manually on keywords query matching (as suggested by Google guidelines). This leads to very uniform messages. We also have syntactic, rather than semantic matching. Words like "bay" or "ship" are highly correlated to "cruise", but there is no way to take advantage of this.
- AdWords does not let us consider user profile/preferences in targeting the message. Google could let advertisers "use" this option (all users are tracked when querying), e.g., when a user types in "cruise" and he likes classical movies, then display the ad: "Tomorrow is another bay".

The core of the problem is double-faced: on one side Google matching is fairly limited (only keywords, no user model), on the other side exploiting user's profiles would bring a complexity explosion that could be barely handled by human copywriters. The example above, "Tomorrow is another bay", has been presented since it was automatically generated by a system that got in input "cruise"+"movie", mocking the famous quote "Tomorrow is another day" from *Gone with the wind* (Strapparava et al., 2007).

6. Conclusions and Future Work

AdWords gives us an appropriate context for evaluating persuasive messages. The advantages are fast building and evaluation of experiments. By using keywords with a low Cost Per Click (not relevant for business and with low competition) it is also possible to run large scale experiments with a cost of only a few dollars (advertisers pay only for clicks, not for impressions).
AdWords proved to be very accurate, flexible and fast, far exceeding our expectations (e.g., the possibility of accounting for the effect of capitalizing a single letter). This accuracy, flexibility and quickness call for careful design of the experiments. In the future we would also like to test:

1. Campaign-tracking variables, for longer messages. Campaign-tracking variables are labels attached to the hyperlinks. They enable us to uniquely identify all hyperlinks and the activity generated by those links (for example links included in an email campaign).
2. Google Analytics tools, for evaluation in different stages of persuasion. For example, competing ads can have the same text but different landing pages. Once users click on the message they are redirected to different pages corresponding to different experimental conditions. In this case the experiment begins once the subjects are in a later stage of the persuasion process (i.e. they already showed an interest in the topic, by clicking on the message).
3. Google's API, to automatically interface with AdWords. We can envisage a system generating messages for AdWords that can automatically learn and adapt to user tastes based on AdWords feedback.

7. Acknowledgments

We would like to thank Fabio Pianesi for his useful suggestions about experiments evaluation.

8. References

B. J. Fogg. 2009. Creating persuasive technologies: an eight-step design process. In *Proceedings of the 4th International Conference on Persuasive Technology*.

M. Guerini, O. Stock, and C. Strapparava. 2008. Valentino: A tool for valence shifting of natural language texts. In *Proceedings of LREC 2008*, Marrakech, Morocco.

E. Reiter, S. Sripada, and R. Robertson. 2003. Acquiring correct knowledge for natural language generation. *Journal of Artificial Intelligence Research*, 18:491–516.

R. Snow, B. OConnor, D. Jurafsky, and A. Ng. 2008. Cheap and fast but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii.

C. Strapparava, A. Valitutti, and O. Stock. 2007. Affective text variation and animation for dynamic advertisement. In *Proccedings of 2nd International Conference on Affective Computing and Intelligent Interaction (ACII2007)*, Lisboa, Portugal.