Keyword Optimization in Sponsored Search via Feature Selection

Kiritchenko, Svetlana; Jiline, M.

Proceedings of the ECML PKDD 2008, Workshop on Newchallenges for feature selection in data mining and knowledgediscovery, September 15, 2008, Antwerp, Belgium, 2008

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*Keyword Optimization in Sponsored Search via Feature Selection **

Kiritchenko, S., and Jiline, M.
September 2008

* Proceedings of the ECML PKDD 2008, Workshop on New challenges for feature selection in data mining and knowledge discovery, 15 September 2008, Antwerp, Belgium

Canada

Keyword Optimization in Sponsored Search via Feature Selection

Svetlana Kiritchenko SVETLANA.KIRITCHENKO@NRC-CNRC.GC.CA
Institute for Information Technology
National Research Council Canada
Ottawa, Canada

Mikhail Jiline MZHILIN@EPIPHAN.COM
Epiphan Systems Inc.
Ottawa, Canada

Editor: Saeys et al.

Abstract

Sponsored search is a new application domain for the feature selection area of research. When a user searches for products or services using the Internet, most of the major search engines would return two sets of results: regular web pages and paid advertisements. An advertising company provides a set of keywords associated with an ad. If one of these keywords is present in a user's query, the ad is displayed, but the company is charged only if the user actually clicks on the ad. Ultimately, a company would like to advertise on the most effective keywords to attract only prospective customers. A set of keywords can be optimized based on historic performance. We propose to optimize advertising keywords with feature selection techniques applied to the set of all possible word combinations comprising past users' queries. Unlike previous work in this area, our approach not only recognizes the most profitable keywords, but also discovers more specific combinations of keywords and other relevant words.

1. Introduction

Dimensionality reduction has been a critical step in many academic and real-life applications ranging from text classification to DNA microarray analysis. This paper presents a novel domain, paid advertisement or sponsored search, that can also benefit from the feature selection paradigm. We applied feature selection techniques to the task of keyword optimization in sponsored search. Unlike most domains where feature selection is used only as a preprocessing step, in this application feature selection constitutes the base algorithm for keyword selection and optimization.

Sponsored search is a fast-growing, multi-billion dollar industry that emerged just a few years ago. Today, most of the major search engines (including Google, Yahoo!, and Microsoft) have mechanisms to complement normal search results (organic search) with paid advertisements (paid search) related to a user's query. This process is potentially beneficial to all parties: an advertising company, a search engine company, and a user. An advertising company is presented with an opportunity for a large-scale direct advertisement.

The company is accountable for creating short-text ads of its products or services supplied with a list of keywords. Each search advertising keyword can be a single word or a multi-word phrase possibly with a negative qualifier (meaning "do not match"). If a user's search query contains one of these keywords, the corresponding text ad will be shown to the user (impression) in a specially marked area for sponsored search results. If the user is interested in the ad, s/he clicks on it and is redirected to the advertiser's webpage, called a landing page. Unlike display (banner) advertising, where the advertiser is charged for each ad display (pay-per-impression), in the sponsored search model the advertiser pays only for actual clicks on its ads (pay-per-click). This model is also different from content advertising where an ad is chosen based on its content similarity to the webpage. A search engine company, while charging small fees for each ad click (usually less than $1), generates billions of dollars in net ad revenue due to the tremendous scale of the project. Finally, a web search user is presented with an additional set of highly relevant search results, frequently inaccessible otherwise.

This advertising model presents a number of challenges to the research community. Often, several advertisers are interested in the same keyword. They enter an auction and bid a maximum amount they are willing to pay for this keyword. They also specify the maximum daily budget. A search engine company has to decide which ads to display based on the bidding prices, the click-through rates and other parameters, with the ultimate goal of maximizing the profit. An advertising company also aims at maximizing its profit by selecting the most appropriate keywords, optimizing their bidding strategies, and creating attractive ad texts. They also have to design precise and detailed landing pages to persuade a user to a conversion, i.e. buying a product, making a reservation, registering, etc.

In this work we address one of the research challenges of sponsored search, namely keyword selection. In the pay-per-click advertising model, the quality of a keyword is determined by its ability to bring revenue or, in other words, to attract buyers. An effective keyword would have a high percentage of conversions (purchases). The click-through rate (CTR)[1] is less important since keywords with low CTR increase the total cost insignificantly. Traditionally, search advertising keywords are selected heuristically. A good starting point can be the keywords found on an advertiser's website (Abhishek, 2007). This initial set of keywords can further be extended with semantically related phrases (Abhishek, 2007; Chen et al., 2008). However, it is hard to expect equally good performance from all those keywords. We can optimize a set of search advertising keywords by analyzing the historic data of the keyword performance. Search engine companies usually report some keyword statistics for a particular advertising campaign. In addition, many advertising companies collect logs of their website visits with information on visited pages, time spent at the website, a referring site (including a complete search query, not only the keyword matched), and user actions. With this information we can analyze the effectiveness of keywords as well as the effectiveness of all words and phrases constituting the users' queries. This would allow us not only to select high-quality keywords but also to improve some of the keywords with additional (possibly negated) words or phrases.

Overall, the objective of this study is to make search advertising keywords more specific and, as a result, more profitable, by extending them with (possibly negated) words from

1. The click-through rate for a keyword is defined as the number of clicks divided by the number of impressions generated by the keyword.

search queries. To achieve this objective, we propose to apply feature selection techniques on a set of all possible phrases generated from users' queries. The complete procedure consists of four steps. First, a set of all possible single and multi-word phrases is generated from available search queries. Second, a feature selection method is applied to sort the phrases by their effectiveness on historic data. Then, a number of top-quality phrases are selected to maximize the profit from the advertising campaign. Finally, the resulting list of phrases is converted into an improved set of keywords.

We show that the produced list of phrases has predictive power comparable to that of state-of-the-art classification techniques, while being much easier to interpret. In addition, this list of phrases can directly be converted to a new set of advertising keywords with improved performance. Even the most comprehensible traditional classification methods, such as decision trees and rules, are less flexible, time-consuming, and hard to interpret in this setting.

The rest of the paper is organized as follows. First, we describe the problem at hand and the available data in greater detail. Then, we report on previous work related to keyword selection in sponsored search. After that, we present our novel approach to keyword optimization, evaluate its performance on the available dataset and discuss the results. We conclude the paper with the directions for future work.

2. Problem and Data Description

In this section we define the notion of search advertising keyword employed in the current work and formalize the problem. In the following, we call ***term*** any sequence of non-space characters. Generally, terms represent natural language words like *video*, *frame*, etc., but can also represent numbers (*2.0*), models (*ES-388*), web addresses (*http://www.google.com*), and other combinations of word and non-word characters.

Definition: A ***search advertising keyword*** is a set of one or more positive terms and zero or more negative terms, i.e. "***term*** $\{1,n_{Pos}\}$ $\neg$ ***term*** $\{0,n_{Neg}\}$".

Examples of keywords include "*frame*", "*frame grabber*", "*frame grabber* $\neg$*image* $\neg$*processing*". We say that a query matches a keyword if it contains all positive terms of the keyword in any order and does not contain any of the negative terms of the keyword. Note that a query can include terms other than the keyword terms. For example, search query "*video frame grabber for linux*" matches keyword "*frame grabber* $\neg$*image* $\neg$*processing*".

Definition: given a set of initial keywords, an ***extended keyword*** is a search advertising keyword containing one of the initial keywords.

An extended keyword can coincide with an initial keyword or extend it with one or more positive and/or negative terms. For example, "*frame grabber* $\neg$*image* $\neg$*processing*" extends the keyword "*frame*". Longer keywords tend to be more specific and, thus, should better match an information need of a user. For example, an average-profit generic keyword can be merged with another (possibly negative) word matching fewer irrelevant queries and, as a result, leading to a higher percentage of conversions.

Objective: given a set of initial search advertising keywords produce a set of extended keywords that results in higher profit.

We optimize a keyword set based on historic data of users' search queries. The following analysis is based on 3-month data of an SME's advertising campaign at Google. The company operates in a video signal processing business. In the reported period, it advertised on 388 unique keywords ranging from single words to 5-word phrases. The dataset is constructed from the company's weblogs and contains all users' queries resulted in paid clicks along with the label on users' activities. The activity of our primary interest is a conversion (purchase). Even though we have information on immediate conversions linked to users' queries, it represents only a small portion of the company's Internet sales. Most purchases are delayed due to the nature of the company's business (business-to-business sales). Therefore, we consider all visits indicating some interest (engaged visits) as targets. People that spend at least a few minutes browsing the company's website, visit several pages and/or make a purchase are considered (potential) buyers. In particular, we define the Engaged Visit score as the time spent at the website multiplied by the number of pages visited. If the score ≥ 5 or a purchase was made, the visit is labeled as engaged.

The ultimate goal of the project is to globally optimize a set of advertising keywords. However, in the current study we focus on only local transformations of the keywords since this can be evaluated on the available data. More drastic changes in a keyword set, such as adding completely different keywords, would require an explicit evaluation through a new advertising campaign, a possibly larger campaign budget, and a waiting period. We will deal with this matter in future work.

3. Related Work

Sponsored search is a new research area with primary focus on auction mechanism design and bidding strategy optimization. Research studies investigate the best practices for a search engine company for selecting the most profitable advertisements (Mehta et al., 2005; Abrams and Ghosh, 2007) and the best bidding strategies for an advertising company for maximizing its profit (Kitts and Leblanc, 2004; Chakrabarty et al., 2007). Only a few papers concern the issue of keyword creation and optimization. Google's Adword Tool[2] help advertisers to extend their seed keywords by suggesting past frequent queries that contain one of the keywords. Semantically close phrases can be mined from advertiser data and search click logs (Bartz et al., 2006). In general, advertising keywords associated with the same landing page are closely related. The same is true for user search queries associated with the same clicked URL. Bartz et al. make use of these data with logistic regression and collaborative filtering techniques. The goal of the work by Abhishek is to produce an extensive list of less common and thus less expensive phrases semantically similar to seed keywords (Abhishek, 2007). Bidding on a large number of low-cost terms can potentially generate the same amount of traffic while costing less. The works mentioned above estimate semantic similarity of words through statistical co-occurrence. A more recent study by Chen et al. replaces statistical similarity with conceptual similarity fully exploiting the knowledge from a concept hierarchy (Chen et al., 2008).

2. https://adwords.google.com/select/KeywordToolExternal

The two most related papers to our work refer to the problem of keyword selection. The work by Rusmevichientong and Williamson selects n best keywords from a given list of keywords sorted by profit-to-cost ratio (Rusmevichientong and Williamson, 2006). The authors show that this strategy guarantees the conversion of the average expected profit to a near-optimal solution. Rutz and Bucklin build a binary logit model augmented with shrinkage procedures to select best keywords based on their estimated cost-per-conversion (Rutz and Bucklin, 2007). Their model suggests that the conversion rate of a keyword depends on many secondary factors like click-through rate, position in a paid search result listing, and semantic characteristics of a keyword. Different from these two studies that focus on keywords, the current work analyzes actual search queries. Having a larger context of the search, our approach is able to prioritize the original keywords along with the keywords extended with highly predictive words.

4. Optimizing Keywords via Feature Selection

In the current study, the search advertising keyword optimization task is addressed through direct optimization of the feature set. We propose the following procedure:

1. generate all possible single and multi-word phrases from available search queries;
2. apply a feature selection method to sort the phrases by their past performance;
3. select the number of top-quality phrases to maximize the profit from the advertising campaign;
4. convert the list of selected phrases into an improved set of keywords.

While single words frequently have a broad meaning, multi-word phrases are more specific and, thus, can be more discriminative as advertising keywords. For that reason, in the first step, we pull together all possible combinations of words appearing in a search query to form the feature set. The order of words and their proximity in a query are not taken into account. For example, for a search query "*a c b*" the following 7 combinations are generated: "*a*", "*c*", "*b*", "*a c*", "*a b*", "*b c*", "*a b c*". All combinations that appear less than 5 times in the training set are removed resulting in 12,721 features.

This idea of enumerating all word combinations that appear in training examples, infeasible in standard text classification and even in sentence classification, is realistic in our setting. As opposed to textual documents (e.g. articles, web pages, emails, etc.) having thousands of different words or sentences having tens of words, most search queries are short containing 1-5 single words (see Figure 1 for the query length distribution in the training data). While the number of word combinations grows exponentially with each new word, the vocabulary is restricted to the company's area of expertise. As a result, the feature sets can be handled effectively on most contemporary machines. If needed, special data structures, such as suffix trees, can be employed. If the number of word combinations becomes prohibitively large (in the case of a large-scale advertising campaign), the length of phrases can be restricted to 3-5 while still getting most of the benefits of the presented approach.

In the second step, a filter feature selection algorithm is applied to sort the word combinations by their importance in class discrimination. In the results section, we demonstrate

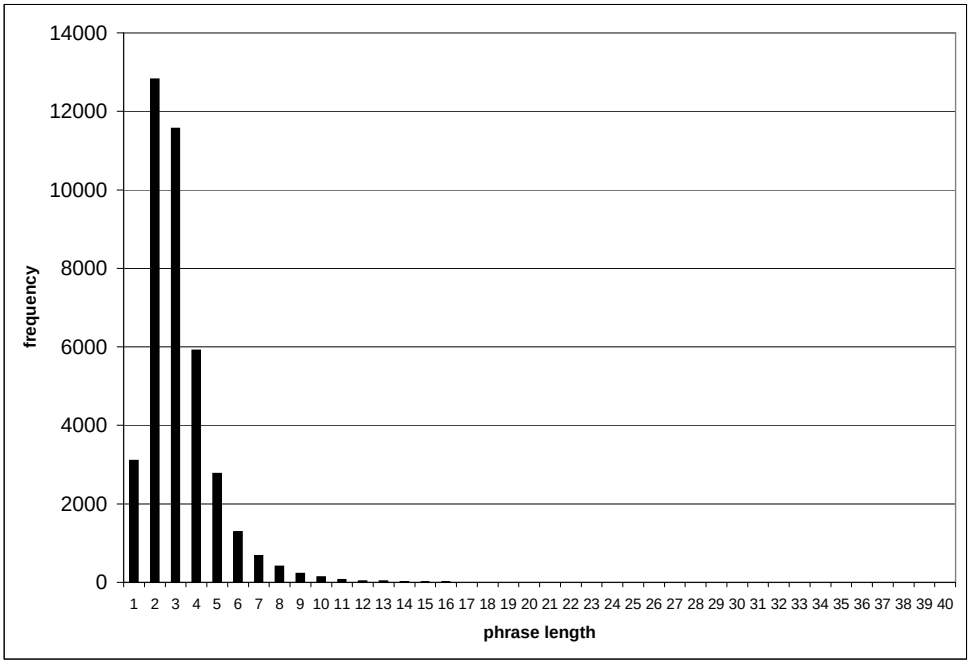

Figure 1: Query length distribution.

the potential of the approach with the widely accepted feature selection techniques: information gain, chi-square statistics, odds ratio, and symmetrical uncertainty (Table 1). We also include a simple feature selection strategy that seems natural for this task: selection by the precision on the positive class.

In the third step, we select n top-scoring phrases to form a new set of keywords. The number n is chosen based on the revenue-costs analysis described in detail in Section 6.

Finally, an ordered list of phrases is converted into a set of extended keywords. There are a number of ways to do it. We have chosen the following straight-forward approach:

1. consider highly-ranked negative phrases (phrases associated with the negative class) as low-ranked positive phrases, i.e. re-rank all negative phrases in the reverse order placing them after all positive phrases;

2. set the threshold at top n phrases;

3. replace the original set of keywords with extended keywords generated from the list as follows:

   (a) if a phrase above the threshold represents an original keyword or contains an original keyword, include the phrase into the new set;

Table 1: Feature selection metrics. The functions specify the relevancy of term t_k to category $c_i \in C$ in the probabilistic form. $|D|$ denotes the total number of examples, $\bar{t_k}$ represents the absence of term t_k, and $\bar{c_i}$ represents all categories in C other than c_i.

Feature selection metric	**Formula**
information gain	$H(C) - H(C\|A)$, where $H(C) = -\sum_i P(c_i) log_2 P(c_i)$, $H(C\|A) = -\sum_{A\in\{t_k,\bar{t_k}\}} P(A) \sum_i P(c_i\|A) log_2 P(c_i\|A)$
symmetrical uncertainty	$2 \times \left[\frac{H(C)-H(C\|A)}{H(C)+H(A)}\right]$
chi-square statistics	$\frac{\|D\|\cdot(P(t_k,c_i)\cdot P(\bar{t_k},\bar{c_i}) - P(t_k,\bar{c_i})\cdot P(\bar{t_k},c_i))^2}{P(t_k)\cdot P(\bar{t_k})\cdot P(c_i)\cdot P(\bar{c_i})}$
odds ratio	$\frac{P(t_k\|c_i)\cdot(1-P(t_k\|\bar{c_i}))}{(1-P(t_k\|c_i))\cdot P(t_k\|\bar{c_i})}$
precision on the positive class	$P(c_i\|t_k)$

(b) if a phrase above the threshold does not contain an original keyword, include all original keywords with the added phrase (if it is not already part of the keyword) to the new set;

(c) if a phrase below the threshold contains one of the phrases above the threshold, the extra part of the low-ranked phrase is added to the high-ranked phrase with negation.

Here we give some examples:

(3a) The phrase "*video converter*" is one of the original keywords in the dataset. Since it is highly ranked by information gain, it is included in the new set of keywords.

(3b) The word combination "*digital video*" does not contain any of the original keywords, but is also ranked high; therefore, all original keywords "k_1", "k_2", ..., "*video converter*", ..., "k_m" with the added phrase "*digital video*", i.e. "k_1 *digital video*", "k_2 *digital video*", ..., "*digital video converter*", ..., "k_m *digital video*", are included in the new set.

(3c) If there is a high-ranked phrase "*frame grabber*", but a low-ranked extended phrase "*frame grabber image processing*", the phrase "*frame grabber*" is modified as "*frame grabber* ¬*image* ¬*processing*". This improved phrase should prevent the advertising company from paying for the useless traffic from users looking for frame grabbers with image processing capabilities.

Unlike previous work on keyword selection in sponsored search, the new approach allows us not only to recognize highly predictive keywords, but also to discover better keywords by adding phrases to the original keywords. Presently, if a high-ranked feature does not contain an original keyword, we would not add it as a separate keyword, since we do not have full evidence of its past performance. For example, the word "*frame*" is one of the top features selected by information gain. That makes it a highly predictive word, but only in combination with the original keywords. Without the context of the original keywords, the word "*frame*" is generic and would probably generate lots of worthless traffic. That's why we add such highly predictive phrases to the existing keywords. On the other hand, the presented approach suggests new words and phrases that can potentially be useful on their own or in combination with other words, though their performance will have to be evaluated in a separate campaign. We plan to address this matter in the future work.

5. Results

A set of experiments demonstrating the effectiveness of the presented approach was conducted on the available data. The 3-month data were split into a training (first two months) and a test (the last month) set. For training, only visits with non-empty search query (both paid and organic) not mentioning the company's name or its product names were included. For test, the data were further restricted to paid referrals from Google. There are 39,127 (9% positive) training and 14,566 (10% positive) test examples. The training data contains 12,707 single words and 3,046,171 word combinations. This set is reduced to 12,721 features by keeping the phrases that appear at least 5 times in the training data.

The quality of an advertising keyword set is determined by the profit made through the advertising campaign. Still, we first report the effectiveness of the algorithm in terms of traditional machine learning evaluation measures. The economic aspect of the problem is discussed in Section 6. The algorithm produces a ranked list of sets of extended keywords. In machine learning, the ranking quality is conventionally measured with the Area under the ROC Curve (AUC). We say that a set of keywords matches a search query if at least one of the keywords from the set matches the query. A matched query from an engaged visit counts as true positive, a matched query from a non-engaged visit counts as false positive. By monotonically increasing the threshold n of the top selected features (cf. Sec. 4), different sets of extended keywords are evaluated and a ROC curve is plotted.

In the first set of experiments we compare the performance of the proposed approach with different feature selection methods, namely information gain, chi-square statistics, odds ratio, symmetrical uncertainty, and precision on the positive class. Table 2 reports the results. All methods show similar performance with symmetrical uncertainty being the winner by a slight margin. The simplest method of precision on the positive class is a little inferior to other techniques.

In the second set of experiments we compare the performance of the new approach and the state-of-the-art classification algorithms: Support Vector Machines (SVM) (Chang and Lin, 2001), Naive Bayes (Witten and Frank, 2005), C4.5 Decision Trees (Ruggieri, 2004), and JRip, a Weka version of the well-known Ripper rule learning algorithm (Witten and Frank, 2005). The classification algorithms learn predictive models discriminating engaged and non-engaged visits on the bag-of-words representation of search queries. The feature set

Table 2: Comparison of the feature selection based approach and traditional classification algorithms on the search advertising keyword optimization task. Reported is the Area under the ROC curve (AUC) with 95% confidence intervals estimated as in Hanley and McNeil (1982).

Classification algorithm	**AUC**
Feature selection based	
information gain	0.64246 ± 0.01614
chi-square statistics	0.64332 ± 0.01614
odds ratio	0.6399 ± 0.01615
symmetrical uncertainty	**0.64407±0.01614**
precision on the positive class	0.63782 ± 0.01616
SVM	0.64879 ± 0.01612
Naive Bayes	**0.65810±0.01607**
C4.5 Decision Trees	0.58273 ± 0.01622
JRip	0.62177 ± 0.01621

consists of 1771 single words appearing at least 5 times in the training set. No stemming and no stop word removal are performed. For SVM and Nave Bayes, standard feature selection is applied.[3] Figure 2 shows the corresponding ROC curves, and Table 2 reports the area under the ROC curves.

The experiments demonstrate that the performance of the feature selection approach is comparable to that of the learning techniques. At the same time, the application of traditional classification algorithms on this task faces several practical issues. The first issue concerns the class imbalance: there are about 10 times fewer engaged visits than non-engaged. To compensate for that, cost-sensitive learning (via re-weighting of the training examples) is applied to balance the class distributions. The second issue relates to the size of the dataset. Some of the classification algorithms require an extensive amount of memory and/or CPU time on data of such scale. The third and most important issue concerns the effectiveness of the approach. Even the most interpretable classifiers, decision trees and rules, appear to be inadequate in this setting. C4.5 builds a very large and branchy decision tree, consisting of 4733 nodes. It is very hard to analyze and convert to a set of practical rules. JRip, on the contrary, produces 9 rules with 3 - 11 predicates each, where most of the predicates test the absence of a word. These rules cover only 18 single words from the given 1771 features. Overall, we were unable to make significant decisions on how to optimize the set of keywords using the classical machine learning algorithms.

3. For SVM and Nave Bayes, reported are the best results achieved on the test set by varying the feature selection method (information gain, chi-square statistics, or symmetrical uncertainty) and the number of best features. The decision tree and rule learning algorithms have an embedded ability to select features and, thus, are less sensible to prior feature selection.

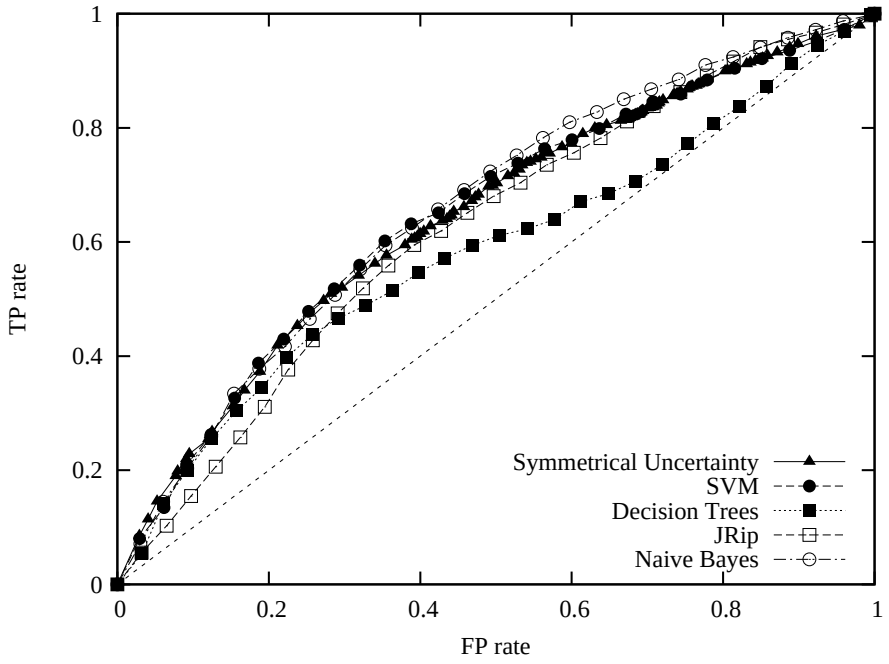

Figure 2: ROC curves for the classification algorithms and the symmetrical uncertainty feature selection strategy.

On a side note, the relatively poor performance of all the methods suggests the high complexity of the task at hand. This can be explained by the noisy nature of the data. All original keywords are carefully selected by humans, so there are no obviously bad keywords. At the same time, there are no perfect keywords either. In general, search queries are short, fairly generic and represent only a high level of a user's needs, therefore the same query can indicate an engaged or non-engaged visit. On the other hand, all algorithms demonstrate better than random performance proving that at least some learning is possible on this task.

6. Economic Considerations

The feature selection paradigm allows us to compile a list of phrases with a predictive performance comparable to the performance of state-of-the-art classification algorithms. At the same time, this list of phrases is easy to analyze, interpret and convert to a set of advertising keywords.

The question remaining is how to select the number of top phrases to maximize the company's profit from the advertising campaign. The profit is defined as follows:

$$profit = RPC * number\ of\ conversions - CPC * number\ of\ clicks,$$

where RPC denotes average revenue-per-conversion and CPC is average cost-per-click.

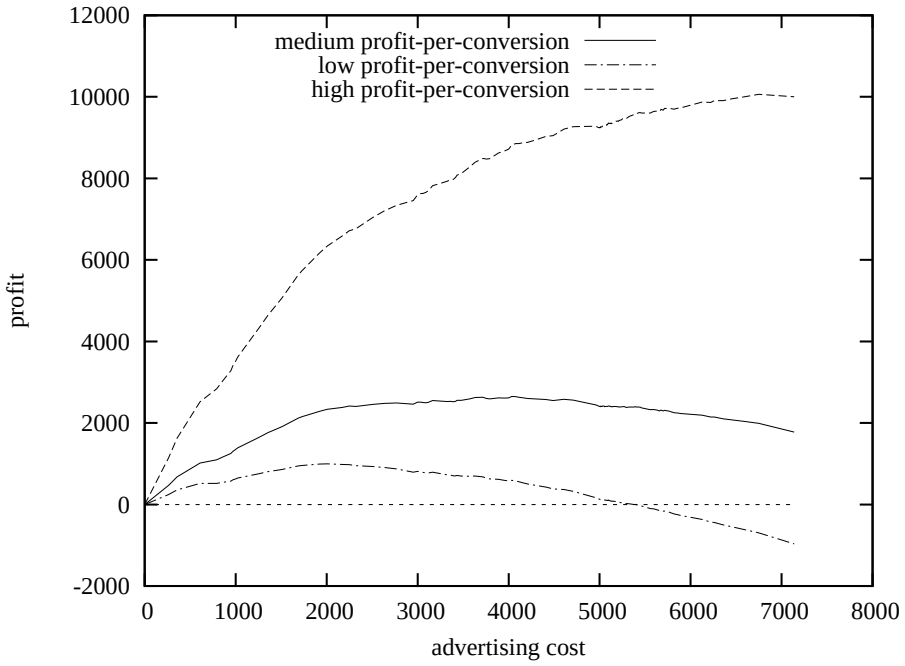

Figure 3: Profit from an advertising campaign.

We estimate the total number of conversions using the engaged-to-conversion rate (ECR):

$$number\ of\ conversions = number\ of\ engaged\ visits\ *\ ECR,$$

$$\text{where } ECR = \frac{total\ number\ of\ purchases}{total\ number\ of\ engaged\ visits}.$$

On the whole, the profit depends on the company's prices, costs to manufacture the products, and costs to advertise. When all these parameters are taken into account, a profit curve can be generated plotting the profit versus the costs of the campaign for different numbers of selected features. Figure 3 shows three situations an advertising company can face. In all three situations the average cost-per-click is set to \$0.50, the engaged-to-conversion rate is 5%, and the analysis is based on the ROC curve for the information gain feature selection approach. The difference is in the revenue-per-conversion figures.

In the first situation, with medium revenue-per-conversion (\$130), the profit grows with the number of selected keywords until it reaches its maximum of \$2,650. This point corresponds to the best number of keywords. In the second situation, with low revenue-per-conversion (\$90), we have a similar shape of the profit curve with the latter curve going below zero. That means that selecting too many keywords would result in the company's loss. Moreover, randomly selecting a set of keywords of any size (a straight line between the start and end points of this curve) always leads to the company's loss. At the same time, our feature selection strategy allows the company to generate profit up to \$1,000. Finally, when revenue-per-conversion is high (\$250), the profit curve is monotonically increasing,

reaching its maximum at the end point. That indicates that profit-per-conversion is so large comparing to the advertising costs that the best strategy would be to keep all original keywords without modification. This type of analysis can help a company to promptly react to the changing market and optimize its advertising campaigns adapting to the new conditions if necessary. Similar reasoning can be performed directly on the ROC curve: the tangent point of the line with slope angle

$$\alpha = \frac{\textit{total non-engaged visits}}{\textit{total engaged visits}} \times \frac{CPC}{RPC \cdot ECR - CPC}$$

maximizes the profit.

7. Conclusion and Future Work

In this work we present a novel problem for machine learning - keyword optimization in sponsored search. The task is to modify a list of keywords used in a pay-per-click advertising campaign to maximize the advertiser's profit. We propose to address this task with a strategy based on the feature selection paradigm. This strategy analyzes the past performance of individual words and phrases comprising the original search queries and selects the most promising keywords possibly extended with highly predictive (positive and negative) words. The proposed technique compares favorably with traditional classification algorithms in terms of both effectiveness and efficiency.

The current work presents a proof of concept for using feature selection techniques in the context of keyword optimization. The task has been simplified by ignoring some of the potentially critical information on individual costs of keywords, their placement in the sponsored search result listing, maximum daily budget, product-specific campaigns, etc. The value of this information will be investigated in the future. Also, other feature selection techniques have to be analyzed and possibly new ones have to be designed specifically for the task. Finally, the approach will be extended to generate keywords not containing the original ones. Based on the current method, new words and phrases can be proposed as potential keywords, yet their performance has to be evaluated in a separate advertising campaign.

Acknowledgements

The authors would like to thank Peter Turney, Joel Martin, and anonymous reviewers for their thoughtful comments and suggestions.

References

V. Abhishek. Keyword generation for search engine advertising using semantic similarity between terms. In *Proc. of the Workshop on Sponsored Search Auctions*, 2007.

Z. Abrams and A. Ghosh. Auctions with revenue guarantees for sponsored search. In *Proc. of the Workshop on Sponsored Search Auctions*, 2007.

K. Bartz, V. Murthi, and S. Sebastian. Logistic regression and collaborative filtering for sponsored search term recommendation. In *Proc. of the Workshop on Sponsored Search Auctions*, 2006.

D. Chakrabarty, Y. Zhou, and R. Lukose. Budget constrained bidding in keyword auctions and online knapsack problems. In *Proc. of the Workshop on Sponsored Search Auctions*, 2007.

C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm, 2001.

Y. Chen, G.-R. Xue, and Y. Yu. Advertising keyword suggestion based on concept hierarchy. In *Proc. of the International Conference on Web Search and Web Data Mining*, pages 251–260, 2008.

J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

B. Kitts and B. Leblanc. Optimal bidding on keyword auctions. *Electronic Markets*, 14(3): 186–201, 2004.

A. Mehta, A. Saberi, U. Vazirani, and V. Vazirani. Adwords and generalized on-line matching. In *Proc. of the Annual IEEE Symposium on Foundations of Computer Science*, pages 264–273, 2005.

S. Ruggieri. YaDT: Yet another decision tree builder. In *Proc. of the International Conference on Tools with Artificial Intelligence*, pages 260–265, 2004.

P. Rusmevichientong and D. Williamson. An adaptive algorithm for selecting profitable keywords for search-based advertising services. In *Proc. of the ACM Conference on Electronic Commerce*, pages 260–269, 2006.

O. Rutz and R. Bucklin. A model of individual keyword performance in paid search advertising. SSRN: http://ssrn.com/abstract=1024765, 2007.

I. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.