Expert Systems with Applications 42 (2015) 4818–4829

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Ensemble learning methods for pay-per-click campaign management

Michael A. King *, Alan S. Abrahams, Cliff T. Ragsdale

Pamplin College of Business, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

ARTICLE INFO

Article history:
Available online 4 February 2015

Keywords:
Sponsored search
Pay-per-click advertising
Classification
Ensemble modeling

ABSTRACT

Sponsored search advertising has become a successful channel for advertisers as well as a profitable business model for the leading commercial search engines. There is an extensive sponsored search research stream regarding the classification and prediction of performance metrics such as clickthrough rate, impression rate, average results page position and conversion rate. However, there is limited research on the application of advanced data mining techniques, such as ensemble learning, to pay per click campaign classification. This research presents an in-depth analysis of sponsored search advertising campaigns by comparing the classification results from four base classification models (Naïve Bayes, logistic regression, decision trees, and Support Vector Machines) with four popular ensemble learning techniques (Voting, Boot Strap Aggregation, Stacked Generalization, and MetaCost). The goal of our research is to determine whether ensemble learning techniques can predict profitable pay-per-click campaigns and hence increase the profitability of the overall portfolio of campaigns when compared to standard classifiers. We found that the ensemble learning methods were superior classifiers based on a profit per campaign evaluation criterion. This paper extends the research on applied ensemble methods with respect to sponsored search advertising.

1. Introduction

Search engines are an indispensable tool for interacting with the World Wide Web (WWW) and have long provided user value, from the earliest universal resource locator (URL) directories to present day highly optimized query results. Companies competing in the search engine industry have slowly monetized their early search innovations and have created sustainable business models by providing the business community with an advertising channel called sponsored search. Internet advertising is nearing a $42.78 billion dollar industry as reported by the Interactive Advertising Bureau (IAB Internet Revenue Report, 2013). The three search engine industry leaders, Google, Yahoo!, and Bing, who hold a combined market share of over 96% (comScore, 2014), each offer a competitive sponsored search platform. Search engine providers are acutely aware of user search behavior and the associated marketing value of the page location of search results (Haans, Raassens, & Hout, 2013). Search engines allow any individual or company to submit a URL for advertising purposes so it can be indexed and then made available for retrieval. Search engines call this submission process organic search and provide the service free. However, the probability of a search engine listing a specific URL for an advertiser's landing page in the top display section is quite low, even with a search engine optimized landing page (Jansen & Mullen, 2008). The statistics in Exhibit 1 are often quoted as support for sponsored search.

In contrast to organic search, sponsored search advertising is more complex, but offers the potential of high return on investment (Moran & Hunt, 2009). In sponsored search, advertisers first bid on keywords offered by the search engines and after a keyword is acquired, their advertisements associated with the keyword are displayed using proprietary ranking algorithms. Ad rankings typically take into account relevance to users' search and keyword bid amount offered by advertiser (Fain & Pedersen, 2005; Jansen & Mullen, 2008). The three predominant advertisement billing schemes used by search engines are pay-per-impression (PPM), pay-per-click (PPC), and pay-per-action (PPA) (Mahdian & Tomak, 2008; Moran & Hunt, 2009). When using a PPM billing scheme, advertisers are charged each time their ad is displayed, regardless of whether the user clicks on the ad. Under PPC billing, the advertiser is charged only when their ad or URL is clicked on, and with PPA billing, the advertiser pays only when a user action such as a sign-up or purchase occurs.

The research conducted for this article analyzed a large data set of PPC advertisements placed on Google. Our data set was provided

* Corresponding author at: Department of Business Information Technology, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. Tel.: +1 540 231 6596.

E-mail addresses: michael.king@vt.edu (M.A. King), abra@vt.edu (A.S. Abrahams), cliff.ragsdale@vt.edu (C.T. Ragsdale).

http://dx.doi.org/10.1016/j.eswa.2015.01.047
0957-4174/

93% of directed traffic to websites is referred by search engines.	Forrester Research, 2006
99% of Internet searchers do not look past the first 30 search results.	Forrester Research, 2006
97% of Internet searchers do not look past the top three results.	Forrester Research, 2006
65% of online revenue is generated by holders of the top three results.	Forrester Research, 2006
Approximately 131 billion searches performed each month, globally.	comScore, 2010

Exhibit 1. Search engine usage statistics.

by an industry leading Internet marketing company (NAICS 518210) specializing in PPC campaign services, that managed a campaign portfolio containing 8499 PPC campaigns for a multi-billion dollar home security provider (NAICS 561612). The Internet marketing company sells its online marketing services to its clients and assumes all PPC related costs, while the advertising clients pay on a PPA basis when purchases are made.

The objective for this study was to construct a set of classification models, using a combination of data and text mining techniques and ensemble learning techniques, that are capable of classifying a *new* PPC advertisement campaign as either sufficiently profitable or not. Profit is based on whether clicks per acquisition are lower than the breakeven threshold for the advertised item with an overall objective of maximizing total profit of the full portfolio of initiated campaigns. An "acquisition" refers to the event of the advertiser successfully selling a multi-year home security contract to the user who clicked on the advertisement.

The remainder of this paper is organized as follows. Section 2 provides a literature review of related work. Section 3 discusses our specific research questions and contributions. Methodology and research design implementations are described in Section 4. Section 5 provides a detailed discussion of the research results while Section 6 presents managerial implications, research limitations, and conclusion.

2. Related work

This section describes related work in the fields of sponsored search, advertisement content modeling, and data mining.

2.1. Sponsored search and search engine marketing

Advertisers are anxious to improve the success of sponsored search listings (D'Avanzo, Kuflik, & Elia, 2011). Various authors have studied prediction of sponsored search campaign success from text features created from keywords or advertisement text. For example, Jansen and Schuster (2011) investigate whether keywords associated with different phases of the buying funnel (Consumer Awareness, Research, Decision, and Purchase) have different success rates. A number of authors have attempted to determine whether semantic features of keywords impact sponsored search success (Rutz & Bucklin, 2007; Rutz, Trusov, & Bucklin, 2011; Shaparenko, Çetin, & Iyer, 2009).

2.2. Advertisement content modeling

Textual content, including that found in advertisements, can be analyzed within a data or text mining context based on the numeric representations of stylometric, sentiment, and semantic features of the text (Abrahams, Coupey, Zhong, Barkhi, & Manasantivongs, 2013; Aggarwal & Zhai, 2012; Haans et al., 2013; Nielsen, Shapiro, & Mason, 2010).

Stylometrics describes the readability and stylistic variations of a portion of text using numerous metrics such as characters per word, syllables per word, words per sentence, number of word repetitions, and Flesch Reading Ease (Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2014). Tweedie, Singh, and Holmes (1996) and Ghose and Ipeirotis (2011) describe the value of stylometric modeling of text using several machine learning techniques.

Sentiment content refers to the emotional or affective communication embedded in text. Feldman (2013) provides an overview of business applications of sentiment analysis in areas such as consumer reviews, financial market blogs, political campaigns, and social media advertising. When studying the characteristics of an advertisement, the representation and interpretation of its sentiment content is important because advertisements not only deliver objective descriptions about branded products or services, but also can induce measureable emotional reactions from current or potential customers. Heath and Nairn (2005) proposes that advertisements that approach readers' feelings rather than knowledge can be processed with low attention and can result in increased buying behavior.

Semantics refers to the meaning of the text, such as the categories of items referred to in the text. Stone, Dunphy, and Smith (1966) describe a semantic tagging method that extracts word senses from text and classifies the words into concept categories. Abrahams et al. (2013) illustrates how this type of sematic tagging of advertisement content can be useful for audience targeting.

2.3. Data mining

Numerous data mining and machine learning techniques have been used to create models that predict important sponsored search performance metrics such as clickthrough rate, conversion rate, and bounce rate. Logistic regression, Support Vector Machines (SVM) and Bayesian models, as well as other techniques, have been applied to these predictive problems. Search engines are primarily concerned with modeling the click-through rate for both new and ongoing ads, because revenue depends on their ability to rank PPC ads relevant to searchers with as high a click-through rate as possible. In contrast, advertisers are more likely to focus on the conversion rates associated with their PPC campaigns.

Richardson, Richardson, and Ragno (2007) argues that modeling click-through rates for ads with known run times is a straight-forward process, while modeling the click-through rate for a new ad presents unique challenges. Because of the rapid growth in the inventory of new PPC ads, along with the fast turnover of these ads, the authors indicate that it has become increasingly difficult to estimate plausible click-through rates based on historical data. The authors present a logistic regression model fit using a feature set created from the actual ad text and numeric attributes (e.g., landing page, keywords, title, body, clicks and views). The authors discuss how their base logistic regression model was more accurate than an ensemble of boosted regression trees.

Wang et al. (2012) add to the research stream of ensemble learning within a sponsored search context by introducing an ensemble for click-through prediction. Wang et al., were team members who participated in the 2012 KDD Cup (http://www.kddcup2012.org/) and built an ensemble consisting of four base classifiers: maximum likelihood estimation, online Bayesian probit regression, Support Vector Machines, and latent factor modeling. Feature creation is a reoccurring research theme within the context of sponsored search because the unit of analysis, the PPC advertisement, typically provides relatively few independent variables. The authors' main contribution is their novel approach for combining