October 03, 2016

Prediction of Traffic for Internet Domains

Joya Zuber
Hugues Vincent
Isaac Andres

Follow this and additional works at: http://www.tdcommons.org/dpubs_series

Recommended Citation
Zuber, Joya; Vincent, Hugues; and Andres, Isaac, 'Prediction of Traffic for Internet Domains', Technical Disclosure Commons, (October 03, 2016)
http://www.tdcommons.org/dpubs_series/283

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.
Prediction of Traffic for Internet Domains

Authors: Joya Zuber, Hugues Vincent, Isaac Andres

ABSTRACT

This disclosure describes techniques to generate predictions regarding genuine traffic that a domain is likely to receive without access to historical data about the domain. Domain traffic data for known domains is utilized to learn characteristics of domains that are likely predictors of traffic that the domains receive. Characteristics of a query domain are used to predict traffic that the query domain is likely to receive. Traffic predictions can be used to identify domains that buy referral traffic and to evaluate domains.

KEYWORDS

- Domain traffic prediction
- online advertising
- internet domains
- parked domains

BACKGROUND

Online advertising networks serve advertisements on large numbers of Internet domains e.g., on web pages associated with the domains. In some instances, advertisements may be served on “parked” domains. A parked domain is a domain that has been registered but is not in active use. Advertisers charge a price for the advertisements that are served at the different domains. For example, an advertiser may “pay per click” for an advertisement that a user has clicked on. Distinguishing between genuine and non-genuine traffic and serving
advertisements on domains that have genuine incoming traffic is valuable to advertisers and advertising networks.

Genuine traffic has characteristics that distinguish it from other traffic. For example, genuine traffic to a domain may originate from a user typing the domain name in a browser application. In another example, genuine traffic may originate from users clicking on a link on a web page that refers to the domain. Internet domains may also get traffic that is not genuine. For example, traffic originates from an expired link. In another example, traffic is purchased from third parties. Some information sources provide traffic source data for websites that is based on actual usage statistics that are gathered from various data sources. Data from such information sources may be sparse and unreliable.

It is difficult to determine the source of traffic on a parked domain. For example, referral data is often hidden or not traceable and third parties may own the Domain Name System (DNS) record. Many parked domains do not receive enough traffic to reliably estimate traffic sources e.g., by using historical data and statistical methods. Techniques to quickly and automatically determine the source of traffic on a domain are valuable to both advertisers and advertising networks. Further, estimating the amount of organic traffic a specific domain is likely to receive is valuable to domain owners and advertisers that place ads on those domains.
Fig. 1

Fig. 1 shows an example analytics system (110) that is used to identify genuine traffic and to generate predictions of traffic that a particular domain is likely to receive. Analytics system accesses domain traffic data (120). Domain traffic data (120) includes traffic information for existing domains that are known to receive a relatively large amount of natural (or genuine) traffic. Domain traffic data includes data on traffic among the existing domains. Further, domain traffic data includes Uniform Resource Locators (URLs) that are associated with the existing domains. For example, a domain may be www.sandwichshop.com. URLs such as delivery.sandwichshop.com, reserve.sandwichshop.com, order.sandwichshop.com, contact.sandwichshop.com, and menu.sandwichshop.com may be associated with the www.sandwichshop.com domain. Domain traffic data includes data on incoming traffic for the domain, and specific data on traffic for the individual URLs associated with the domain. Domain traffic data includes sources of the incoming traffic for the domain. For example, www.sandwichshop.com receives incoming traffic from a search engine, a map website, a website that lists restaurants etc.
Training set and learning

Analytics system (110) utilizes all or a portion of domain traffic data (120) as a training set (125). Learning module (140) of analytics system (110) is configured to perform a variety of analyses on the training set. For example, learning module (140) determines various characteristics of domain names in the training set e.g., by parsing the domain names. In some example scenarios, such characteristics include a length of domain names in the training set or whether the domain names use hyphens or other characters. In some example scenarios, such characteristics include whether the domain names include known words and a language corresponding to the known words. In some example scenarios, such characteristics include the Top Level Domain (TLD) in the domain names (e.g., .com, .net, .biz, .gov etc.). In some example scenarios, such characteristics include similarity to frequently visited domains.

Learning module (140) utilizes traffic data for domains in the training set to determine characteristics that are associated with domains that receive genuine traffic, and characteristics that are associated with domains that receive relatively high traffic. For example, domains that have names that include words in English language may receive relatively high traffic compared with domains that do not include English words. In another example, domains with relatively short domain names may see higher traffic than those with longer domain names. Learning module (150) applies heuristics and statistical techniques to determine and weight the determined characteristics as predictors of traffic. Learning module (150) generates a composite score for each domain in the training set based on the determined characteristics.
Traffic prediction

Analytics system (110) receives queries that include a query domain (130). In various examples, the query domain includes an existing domain, a parked domain, or a domain that has not been registered. Analytics system (110) utilizes prediction module (150) to generate predictions (160) regarding genuine traffic that the query domain receives. Prediction (160) is based in part on a likelihood that the query domain will receive genuine traffic. In some examples, prediction (160) includes a predicted amount of traffic.

Prediction module (150) utilizes characteristics determined by learning module (140) as likely predictors of genuine traffic. In operation, prediction module (150) parses the query domain to determine characteristics of the query domain. In some examples, prediction module (150) generates a composite score for the query domain based on the determined characteristics. In some examples, the composite score is calculated by applying weights to the different characteristics. For example, different TLDs may have different weights. Prediction module (150) compares the composite score for the query domain with composite scores for domain names that are known to be associated with relatively high genuine traffic.

Based on the foregoing comparison, prediction module (150) generates the prediction prediction (160). In an example scenario, prediction module (150) estimates a high probability of genuine traffic to a query domain that resembles a commonly visited domain. Continuing with the above example, the query domain includes a misspelling of the domain e.g., “www.sanwichshop.com.” In another example, the query domain includes a composite score that matches domains the score of that see relatively high genuine traffic, and therefore the prediction module (150) generates a reliable prediction that the query domain is likely to experience genuine traffic.
The techniques described here can rapidly and automatically generate traffic predictions for very large numbers e.g., millions of query domains without access to historical data for those domains.

Examples of use

Prediction (160) is provided to domain purchasers/advertising network owners (170). In some examples, the prediction is used to value new domains that have not been registered. For example, such domains include domains on a new TLD (Top Level Domain) such as .food or .store.

Predictions can be used by purchasers of domains to evaluate whether a domain is likely to receive organic traffic, without the need to access actual historical visit data for that domain. Domain purchasers utilize predictions generated by analytics system (110) to purchase new domain portfolios, with confidence that the domains in the portfolios are likely to receive genuine, sustainable traffic.

The described techniques (e.g., as implemented in analytics system 110) predict traffic for query domains without requiring actual historical traffic data for the domain, which may often not be available or may be inaccurate. Advertising network owners can use the predictions to classify domains that are new to the advertising network prior to observing their traffic patterns. Predictions can also be used to identify domains that receive significantly more or less traffic than predicted by the system. For example, such identification allows advertising networks owners to find domains that engage in activity that increases traffic through non-organic means or domains that buy referral traffic. Such identification can help advertising network owners manage their networks and find possible areas of abuse.
Predictions can also be used to determine market value of a domain, because the amount of genuine, natural traffic that a domain receives is an important component of the market value. With the explosion new top level domains, predictions may also be used to find domains under a new TLD that has high potential for organic traffic.