Peckalytics: Analyzing Experts and Interests on Twitter

Alex Cheng  
Computer Science  
University of Toronto  
hyc@cs.toronto.edu

Nilesh Bansal  
Computer Science  
University of Toronto  
nilesh@cs.toronto.edu

Nick Koudas  
Computer Science  
University of Toronto  
koudas@cs.toronto.edu

ABSTRACT

We provide a description of Peckalytics, its technology and functionality. Peckalytics processes the entire Twitter data stream in real time and provides a flexible search interface to identify experts in any topic area as well as users with interests in any topic. It provides flexible analytics around sets of experts, their followers as well as sets of users with specific interests. The system is implemented to scale for large data sizes. At the time of this writing it operates on an archive of 30 billion tweets, with 220,000 new tweets crawled every minute. In addition to raw tweets, the social graph of users, and profile information, Peckalytics makes novel use of Twitter lists to assess the expertise of different users. Our aim is to facilitate targeting and optimization of advertising campaigns on the Twitter platform.

Categories and Subject Descriptors
H.0 [Information Systems]: General; H.3.3 [Information Systems]: Information Search and Retrieval

1. INTRODUCTION

Social media have transformed the way we interact online as individuals and consumers. At the same time it is transforming the way businesses aim to interact with their customers and fans online. Before social media became mainstream, marketers and advertisers resorted to the collection of behavioral online information on individuals to target their messages. Individuals were primarily targeted based on the topical focus of the sites they visited. For example sports sites like ESPN would display sports related advertising or advertising related to the perceived interests of sports fans. The general interests of sports fans would be derived based on third party market research (e.g., males aged 25-35 with interest in sports are also interested in certain types of movies or specific male grooming products).

In the early stages of the social web, bloggers on particular topics with wide following were identified in order to endorse or sponsor specific products. At the same time, bloggers started serving advertisements on their blog real estate. Social media is transforming the way marketers and advertisers spend their budgets. Novel ways to market online are gaining traction both from an academic [2] as well as practical point of view [1]. In particular influencer based targeting in social media has emerged as a very popular way to market in social platforms (such as Twitter and Facebook). Individuals are identified as online experts in particular topics; they are either incentivised to participate in sponsored advertising spreading the messages to their followers or the platforms automatically insert sponsored messages in their activity streams (as in the case of Twitter/Facebook advertising). The goal is to increase brand awareness, by increasing the number of impressions (e.g., how many followers see a particular tweet) and click-throughs to particular campaign (how many click on the link embedded in the tweet) with the ultimate goal to track conversions (how many end up purchasing a product).

The demonstration describes the Peckalytics system running on top of the real-time Twitter data stream, processing more than 330 million pieces of content daily and aiding influencer marketing and advertising on Twitter. Are data in our system are collected via the public Twitter api. We utilize carefully crafted algorithms along with solid state drives (SSDs) to ensure that the system is able to provide results within seconds despite the large dataset size consisting (at the time of this writing) of 30 billion tweets and increasing in size daily.

With more than 250M users, Twitter has emerged as a prominent marketing and advertising vehicle in addition to being a prominent social communications platform. Advertising and marketing on Twitter involves two crucial steps. First being able to identify who are the “experts” on any topic on the platform and second being able to identify sets of users with active “interest” on a particular topic. In the context of Twitter an expert in a particular topic is an account (user) that primarily produces and shares content related to the topic and has a wide following that actively engages with the produced content (sharing, re-tweeting etc). A user has interest in a particular topic if the user follows a number of experts in the topic and engages with the content they produce.

Peckalytics offers a number of important functions that aid marketing and advertising campaigns on Twitter. First it can identify expert accounts on any topic (queries on Peckalytics are topics). Second, it offers analytical functions on the set of expert accounts on a specific topic, such as what other topics they are experts at, what conversations
they participate in and what types of content they share online. For example, using Peckalytics, one can learn that experts in ‘social marketing’ on Twitter are also experts in ‘seo’ (search engine optimization), ‘social media’ and ‘pr’ (public relations) among other things. They participate in discussions with hashtags (#smplus, #blogwell, etc) and share content from sites such as practicalcommerce.com and tweetedtimes.com.

The Twitter advertising platform works primarily in one of the following three ways:

- First, the advertiser provides a set of Twitter user handles, and Twitter targets advertisements to the followers of these accounts. The ability of Peckalytics to identify sets of experts at any topic readily aids advertisers to identify the most relevant accounts to provide while instigating a Twitter advertising campaign.

- Second, the advertiser bids on a list of topics on Twitter. Twitter using their own proprietary algorithms identifies which users are interested in the topic and subsequently targets them with tweets “promoted” by the advertisers (inserting them in their tweet stream). Peckalytics can assist in this case as well, by analyzing related topics for a topic of interest, advertisers can identify possibly cheaper topics to bid on. For example, if the price for ‘social marketing’ is too high, Peckalytics could suggest ‘seo’ as a related topic, which may have a relatively lower bid price. The net effect however in the campaign will be the same, as largely the same audience will be targeted, since the followers of experts in ‘social marketing’ and the followers of experts in ‘seo’ are highly related as per Peckalytics.

- Third, the advertisers bid on search keywords (to target searches posed to the Twitter search feature). Information on twitter is temporal by nature and events evolve with time, thus the keywords used in searches evolve over time. When a keyword is used during a search query on Twitter for which an advertisement exists, the platform will display promoted tweets (as advertising) along with the search results. Peckalytics readily assists the keyword bidding process, as for each query identifies the keywords used currently in tweets related to the query, keyword associations prevalent in tweets as well as discussions (hashtags) of interest aiding the advertiser to bid on relevant to the query keywords.

2. TECHNOLOGY

There are several components that need to operate together to make Peckalytics a functioning system. All our servers are from Dell, have 4 quad-core processors, 96GB main memory, and 12TB raw disk storage. At the time of writing, the total data size was over 30 billion tweets spanning three months and growing daily.

The system processes and adds more than 330 million new tweets per day, collected via the public api. The stream is stored in across seven servers. Each server materializes the entire tweet, both content and meta data, utilizing a compressed row format into a local MySQL database. We refer to the collection of the seven MySQL instances as the data store.

A separate process, scans the tweets as they are materialized across the data store, producing a stream of tweets towards the indexer process. The indexer process runs on a node by itself and it is a multi threaded process that materializes a table for each day. Each row in the table is a unique twitter account identifier and a list of all tweet identifiers the account produced that day. This is a fairly heavy read/write scenario as in a day we materialize more than 25 million rows that are read and written on demand as new tweets arrive, for a total of more than 330 million update operations daily. A compressed row format is utilized here as well. Special care is taken to avoid deadlocks, and therefore we run with relaxed transactional semantics to increase throughput across multiple threads reading and writing the table. The tables for the last two weeks are materialized in solid state drives (SSDs) for increased performance. We refer to the collection of tables keeping the association between twitter account identifier and tweet identifiers produced for each day as the index store. The index store can easily retrieve for any day, the identifiers of all tweets produced that day for any set of twitter accounts. The collection of all tweet identifiers can then be provided to the data store to retrieve the actual tweets.

Information about which twitter accounts follow others is constantly crawled from Twitter, materializing the social graph. This graph is stored in a separate MySQL instance; which given a twitter user as a query, returns all twitter users following the queried user. Along with each tweet, a set of meta data are appended by twitter associated with the twitter user. These meta data contain, aggregate number of followers, self disclosed personal information, location information etc. All these account profile related meta data are stored in the profile store and can be queried on demand.

Next is the process for uncovering the expertise of twitter accounts and their interests. Twitter supports the concept of a “user list”. These are human created lists that any Twitter user can create. They contain a descriptive name (supplied by the creator) and a set of twitter accounts associated, according to the creator, with this list. For example a list on ‘machine learning’ will contain all accounts that according to the creator are experts or very related to the topic of machine learning. We source these lists by constantly crawling them. Lists are dynamically modified and new lists are created all the time. These lists are processed utilizing a Apache Lucene index; for each twitter account we produce the set of all lists the account is associated with; we refer to this as the expertise vector. That information is then directed to Lucene to populate the index of topics associated with the account. The index supports full Lucene query syntax, including phrase queries and boolean logic. At the same time, the social graph, provides related information about ‘user interests’. Since for each twitter account I we can also determine all twitter account I follows, the union of their expertise vectors produces the interest vector for I. Commonly by following a twitter account (assuming a non spam scenario) we are interested in the information produced by that account, based on the expertise that account has. Expertise vectors for each identifier are supplied to lucene for indexing.

On top of all this information several algorithms are implemented to extract useful information and conduct analysis.

3. REFERENCES