SerpentTI: Flexible Analytics of Users, Boards and Domains for Pinterest

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

Mary Malit
Computer Science
University of Toronto

Chuanxi Zhang
Computer Science
University of Toronto

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

ABSTRACT

Pinterest is a pinboard style photo sharing web service that allows its users to manage, share and express their interests via a collection of theme based photos. A few design choices of Pinterest makes it highly desirable to social media practitioners and marketers as a new, high quality data source for deep analysis, or as a complimentary data stream to existing social data such as Twitter and Facebook. The analysis capabilities at the current Pinterest site are minimal however as the focus is currently on user experience. We provide a description of SerpentTI, a system that currently crawls, indexes and aggregates more than 31 million users, 96 million boards and 3.1 billion pins from Pinterest to enable flexible and deep analytics.

Categories and Subject Descriptors

H.0 [**Information Systems**]: General; H.3.3 [**Information Systems**]: Information Search and Retrieval

Keywords

Pinterest, social media, analytics

1. INTRODUCTION

Social media web services are mainstream with billions of people participating, generating massive quantities of both labelled and unlabelled data. Popular services such has Twitter have over 200 million users generating more than 350 million tweets per day. Novel analytical services based on content from popular services such as Twitter and Facebook have surfaced. However, analytical services for new sites such as Pinterest are lacking. This may be due to a multitude of reasons, such as lack of accesible APIs and limited data availability.

SIGMOD'14, June 22–27, 2014, Snowbird, UT, USA.

ACM 978-1-4503-2376-5/14/06 ...$15.00.
http://dx.doi.org/10.1145/2588555.2594518.

Pinterest[2] is a social media web service that allows its users to express their interests by sharing photos of their liking via a pinboard style interface to group related photos by themes. Pinterest is notable for having acquired many active users in a short period of time [3]. As of July of 2013, Pinterest has 70 million users, with 30% of the users having actively engaged with the site (either pinned, repinned or liked and image). More specifically, Pinterest encourages its users to share pins. A pin is usually composed of an image, a descriptive text about the image, with the URL that the pin was retrieved from, when applicable. Pinterest allows domain owners to embed a *Pin It* button on their website in order to facilitate the process of pinning from other domains, similar to Twitter and Facebook website integrations for sharing content. Pinterest requires that the pins be grouped into *boards*. Users routinely create boards containing pins of similar theme, and assign the boards textual descriptions as they see fit. (e.g. A user can pin an image of a fruit punch bowl from target.com, share it on Pinterest in a board that he/she creates, and assign the board title *For Fridayś party*). Users also have the option to assign each board one of 33 categories (e.g. Food & Drink, Weddings, Gardening, Technology ...) predefined by Pinterest. Once the set of boards and pins are in place, Pinterest encourages its users to stay up to date with each other by following other users. Users of Pinterest can also interact with each other by liking others' pins, and commenting on others' pins. If they so find a particular pin really interesting, they may wish to *repin* that pin onto a board of their own, in effect, sharing that pin with their own followers.

Several design choices of Pinterest make it attractive for high quality data mining and analytics. First, Pinterest positions the users' primary action around pin/repin onto boards. This is a direct indication of users' interest. (hence the name, *pin-interest*). Both the pins and the boards often have user assigned descriptive titles, and the user usually assigns one of the predefined categories to the boards. From our collection of 96 million boards, 51% of the boards have a user assigned category. As a consequence, Pinterest enables its users to build a massive *interest graph*. Second, board titles often contains highly actionable data about a user as often one indicates strongly, the kind of goods and services one desires and annotates them with a descriptive board title. (e.g. the board *"For My Wedding"* may contain pins about wedding gowns, flowers, cakes or the type of

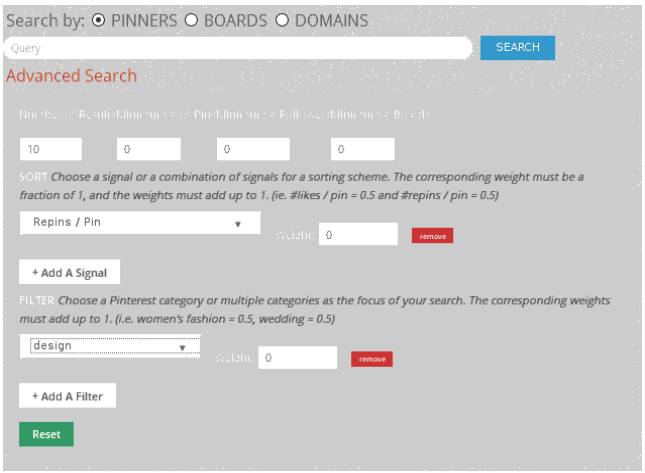

Figure 1: The SerpentTI Search Interface

```
{/*pin*/
   'domain':'...',
   'like_count':'...',
   'images':'...',
   'id':'...',
   'price_currency':'...',
   'pinner':'...',
   'comments':'...',
   'access':'...',
   'comment_count':'...',
   'board':'...',
   'method':'...',
   'attribution':'...',
   'description':'...',
   'price_value':'...',
   'is_playable':'...',
   'link':'...',
   'is_repin':'...',
   'is_uploaded':'...',
   'description_html':'...',
   'repin_count':'...',
   'created_at':'...',
   'dominant_color':'...',
   'embed':'...',
   'rich_summary':'...',
   'is_video':'...'
}

{/*user*/
   'last_name':'...',
   'domain_verified':'...',
   'following_count':'...',
   'image_medium_url':'...',
   'like_count':'...',
   'full_name':'...',
   'image_small_url':'...',
   'id':'...',
   'first_name':'...',
   'secret_board_count':'...',
   'location':'...',
   'indexed':'...',
   'is_partner':'...',
   'website_url':'...',
   'board_count':'...',
   'username':'...',
   'repins_from':'...',
   'twitter_url':'...',
   'facebook_url':'...',
   'follower_count':'...',
   'pin_count':'...',
   'about':'...',
   'has_board':'...',
   'image_large_url':'...',
}

{/*board*/
   'description':'...',
   'pin_thumbnail_urls':'...',
   'collaborator_count':'...',
   'owner':'...',
   'pin_count':'...',
   'id':'...',
   'category':'...',
   'name':'...',
   'privacy':'...',
   'url':'...',
   'follower_count':'...',
}
```

Figure 2: Pinterest JSON schemata

chapel that a user desires). In terms of user management, Pinterest requires its users to login with their email, Twitter or Facebook account. This makes Pinterest a central hub that implicitly links existing social media profiles. Using our repository of Pinterest user database, 80% of the users have logged in through their Facebook account, while 20% have logged in using their Twitter account. Furthermore, due to the lack of a publishing API, Pinterest is less prone to the effect of massive spamming content. Despite these good qualities, Pinterest offers only a rudimentary, content based search functionality over pins and boards on its website with a static, fixed result ranking function.

We present SerpentTI [1] [1], a system that crawls, aggregates and indexes Pinterest with the goal to enable flexible and rich analytics of users, boards and domains. As can be seen in Figure 1, SerpentTI presents a familiar search box, with the option to toggle user, board or domain analytics with advanced filtering and scoring customizations. SerpentTI enhances its users' ability to extract value from the Pinterest interest graph in several practical ways:

Granular indexed fields Users perform full textual boolean search over the titles/descriptions of boards and pins, domain URLs or the contents of the pins. One particularly interesting aspect of this is that queries can be formed to capture users' desire for products or services, restricted to certain criteria. (e.g. "places I want to go" for users in California that have pinned information about yachts).

Flexible ranking functions If required, the result ranking can also be altered by selecting a combination of various statistics and filtering attributes. For example, a user may wish to re-rank the list of Pinterest users from Toronto that are into real estate, ranked by the average followers per pin, as a measure of influence. SerpentTI renormalized all supported ranking attributes to the range of 0 to 100 and it supplies an interface to specify arbitrary weighted combinations of attributes. This allows the scoring function to be very flexible in practice (e.g. a user might want to weight comments / pin twice more important than the fraction of the boards that are in mens fashion).

Deep user/board insight a "More Info" button is available with each user and board result on SerpentTI to deliver further insights including hourly posting activity, related keywords, followers of various authority and activity statistics.

Similar user/board suggestions For search results returning boards and users, SerpentTI provides an efficient and high quality user and board suggestion based on an analysis of the pin content signature of users and boards.

Domain analytics Domain analysis is possible with SerpentTI by aggregating and indexing the domain of origin of pins. This tool is useful for domain owners in order to measure engagement of their domain on Pinterest. Domain analytics provides a dashboard to quickly visualize popular pins, top pinners with authority, gender breakdown, with a word cloud summarizing the biography of repinners and the contents of their pins (to aid better understanding of those engaging with the domain). The domain analytics tab also makes it possible to visualize the volume of repins from the domain as a function of time.

2. TECHNOLOGY

Our hardware is from Dell with 72GB of RAM, 12TB of mechanical disks and 16 physical processors. Since there is no publically available content API for Pinterest, we have wrote several specialized crawlers to consume content from Pinterest. Pinterest continuously updates each of the 33 category pages (refered to as **C** in what follows) with the newest activities. Our crawler continuously extracts the set of pins, user and boards from **C**. Another set of crawlers then performs bread first search outwards from this seed set to further collect pins, boards, followers of boards and users. We also materialize different versions of the user, and board profile in order to display the follower growth history of boards and users on SerpentTI. In order to expediate the crawling process, we have deployed more than 200 processes across a cluster of 16 machines to handle each of the different crawling tasks. The crawlers materialize the

[1]"SerpentTI" is an anagram of "Pinterest"

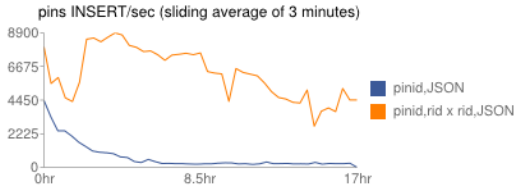

Figure 3: Update throughput for Pin storage: update throughput VS hours into benchmark

pins, users and boards data as JSON and commits them into a MySQL database for easy retrieval. A subsequent set of consumers then derives secondary statistics (e.g. number of pin per day for a specific domain) that is also used by SerpentTI. As materializing all these data from various JSON sources can be painful to manage, we have developed a small but general extract, transform and load (ETL) framework in Java that is capable of loading streamed JSON and perform database related business logic in a batched setting, which expediates database write throughput. The set of JSON consumers on ETL is also distributed across the same cluster of 16 machines to allow parallel data loading. One interesting statistic that we are able to materialize using this crawler architecture, is the growth over time for the number of likes, comments and repin counts for every pin that shows up in **C**. Since Pinterest pushes pins with new comments, repins to the top of **C**, our crawler will periodically observe these activities, and will be able to infer the state of the pin from the JSON metadata. Apache Lucene is used to support searching for users and boards. Due of the amount of data involved, our indexing process is also pipelined into three stages to minimize database stalls. In terms of the database write throughput, we observed the update throughput for pins gets worse very quickly as the database gets bigger. This is due to page misses on the pins table. As the pins get updated to the databases in random pin ID order, the page miss rate increases as the database gets physically larger. We can make much better use of the page cache by partitioning these types of updates into two tables. Effectively, the first table is an append only table where the payload for pins is appended to the table with an auto increment row ID. A second table then establishes the link between the auto increment row ID with the pin ID. As we can see in figure 3, the steady state performance of this second strategy is at least 20 times better than the original, straight forward strategy for storing pins. We make one further optimization to compact multiple pin payloads into one row to achieve better compression. Another secondary optimization SerpentTI uses is to place a part of the data on solid state drives to minimize the read/write latency. At this time, our crawlers have crawled more than 3.1 billion pins, and can update profiles for 30 million users in about 12 days and the board profiles for the 96 million boards in roughly 45 days. After each iteration of crawling, the set of newly discovered users and boards will be shuffled with the existing set of users and boards, before proceeding with next round of distributed crawling on the cluster.

3. DEMO EXPERIENCE

SerpentTI is online at *santiago3.cs.toronto.edu/pinterest/*. A search box that supports the full boolean semantics is shown to the user. On the top of the search box is a list of toggle buttons for the user to indicate the types of analytics desired. SerpentTI currently supports analysis for *boards*,*users*, and *domains*. As usual, a few predefined fields are indexed and can be queried using the regular boolean syntax. For example, a user can find all the boards owned by a Toronto user with regards to jewelry by writing: *jewelry AND location:toronto*. The list of fields that indexed include: *location*, *about*, *boardNameDesc*, *pinDesc*, *domains*, *hasFacebookUrl*, *hasTwitterUrl*, and *hasWebsiteUrl*. With these fields indexed and ready to go, users can write queries such as *domains:homedepot.com AND ("for the" OR "for my")* to search for the users that are interested in products from homedepot.com, effectively performing a search by *need*. As the search results are returned, users will be presented with a *More Info* button per result item for further drill down. For the case of a board result, *More Info* shows the user a few thumbnails of the pins from the board, followed by basic board statistics including the number of pins, followers and collaborators. After that, a series of normalized signal values follow. These include followers per pin, average boards per pin, unique domains per pin, repins per pin, likes per pin and comments per pin. A list of followers for the board is also returned and partitioned by their own follower count into three buckets, corresponding to high, medium and low follower counts. Lastly, a set of similar boards (utilizing our own algorithms) is presented with a snapshot of the pin descriptions at the bottom. Similar drill down capability is also presented for Users allowing deeper analytics. SerpentTI also offers a unique domain analytics tool. Users select the *domains* toggle button to analyze pins shared from a particular domain on Pinterest. By entering *target.com* as the query, the user will now obtain analytics for pins shared from target.com. On the top of the domain result page, a list of highly repinned pins is presented, followed by the daily volume graph, showing the number of pins from the domain as a function of time. After that, a gender breakdown, authority break down of the pinners followed by the familiar summary word cloud of the pinner's contents and their profile biography. As can be seen in Figure 1, SerpentTI provides the possibility of a customized ranking function per search. For the case of a *user* search, one can control the result ranking as a combination of normalized authority measures such as repins per pin, likes per pin, comments per pin or average followers per pin. The user can also further combine these scoring metrics with board category weights. Similar advanced scoring manipulation is also present for *board* search. With these constructs available, one can express fairly sophisticated result semantics such as "people showing pinning activity in the technology category who have pinned jewelry images, ranked by repins per pin". SerpentTI is online and fully functional. All functionalities outlined are available today. During the demo users will have the opportunity to interact with the system in real time. Users will be able to issue queries and perform all analysis outlined above.

4. REFERENCES

[1] `http://santiago3.cs.toronto.edu/pinterest/`.
[2] `http://www.pinterest.com/`.
[3] J. Slegg. Pinterest tops 70 million users; 30% pinned, repinned, or liked in June. `http://bit.ly/13rl0h0`, July 2013.

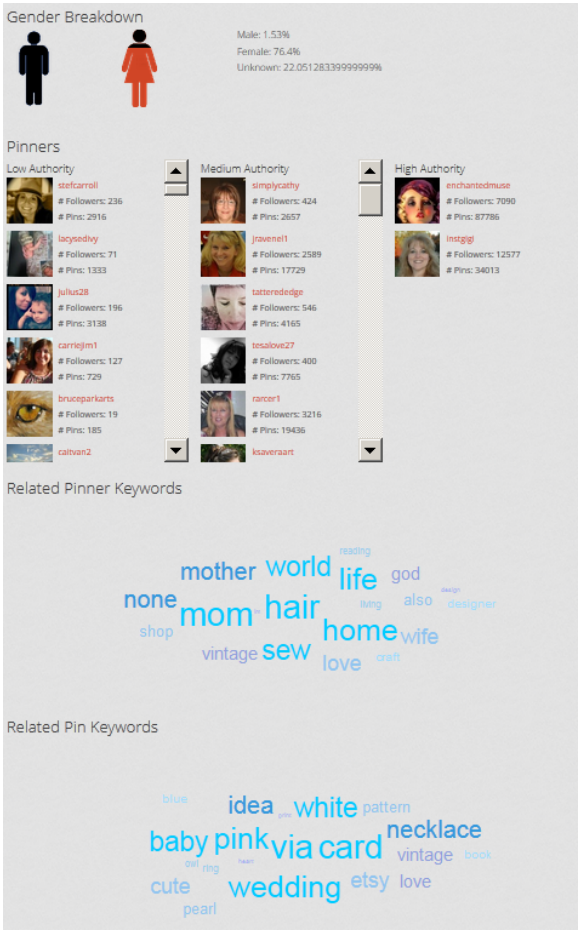

Figure 4: Screenshot of Domain Analytics on SerpentTI

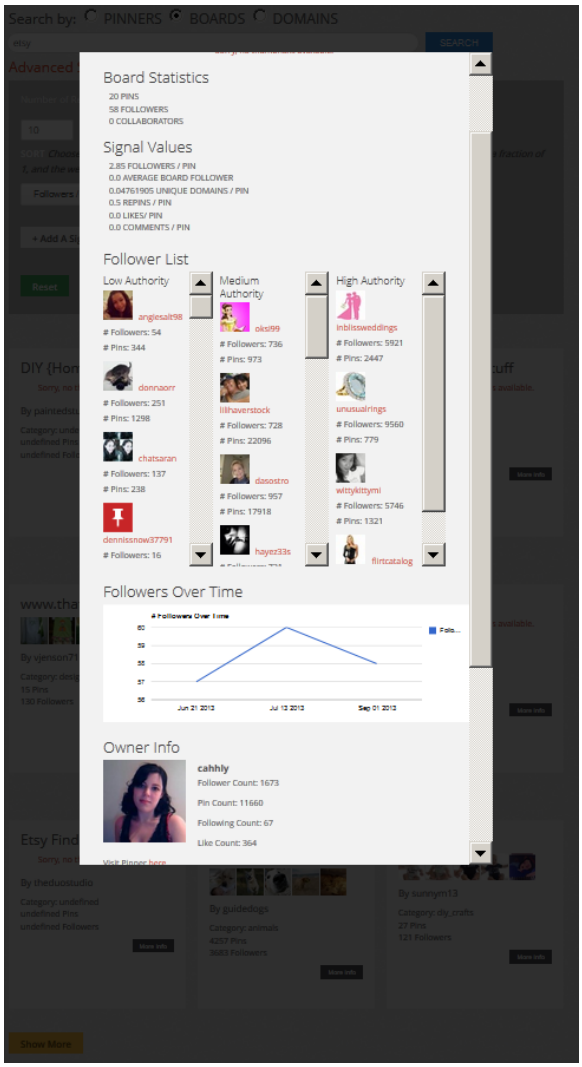

Figure 5: Screenshot of Board Analytics,More Info on SerpentTI