International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 9 (2015) pp. 23391-23397

http://www.ripublication.com

Ranking Viral Websites for Effective Digital Marketing

Easwar V Aiyer, Saikaran Balmuri, Ramprasath Karunakaran, Sriram Subramanian, Kamalaveni Vanjigounder

Dept. of Computer Science and Engineering
Amrita Vishwa Vidyapeetham(University)
Coimbatore, Tamil Nadu, India

Abstract

This project involves indexing websites that are viral. The goal of the project is to find the less popular sites that tend to get viral, thereby helping businesses to cash in on the transient popularity of the site. This will help the business to reach the same audience for a much cheaper price. The first phase of the project involves crawling social networking sites like twitter to find the links that are shared and extract the info related to the shared links. The next phase involves processing this extracted information to find the virality of each shared link followed by feeding these links to another crawler which extracts the webpage content. The meta data and the content is analysed and the links are clustered based on the category of the website.

1 **Introduction**

Many websites/blogs come up with content which interests people. But the content/service of these websites may not be good enough to attract a consistent viewer base.

Placing advertisements on websites with a consistent viewer base like Facebook, Google, Youtube, etc is pretty expensive.

But the not so popular websites which trend at certain points of time cannot cash in on their popularity as there is no means to check if a website is viral/trending in realtime.

2 **Necessity**:

There are lots of methods to estimate a website's popularity such as Google Trends, Alexa rank, SEMRush.com, compete.com etc. All these methods were invented for

digital marketers to know about their competitors. Let us look at the downside of all these methods. Consider Google Trends, they won't tell the actual traffic count and they provide data for sites having a very huge traffic. All monthly traffic creators are simply ignored. Another one is Alexa ranking which is more accurate but it tracks only users who have Alexa toolbar installed to their browsers. Unless the owner of the website reveals the exact traffic of their website, every other method is just a guess.[1]

Another approximate way is to find the number of views for YouTube or Vimeo video link embedded in the webpage. Again it gives a very rough idea on the website traffic. One common drawback of all the above mentioned methods is that they don't show the popularity for a particular instance, they show the overall popularity from the time website was hosted.[2,3]

This led us to exploit the popularity of online Social networks. 72% of internet users are active on social media. The social media is so full of tweets, shares and content. Out of all social networking sites, we will be considering two giants, Facebook and Twitter. At present, there are more than 1 Billion Facebook users. 70% of digital marketers gained customers by advertising on Facebook. Twitter has 550 million registered users which is comparatively less than Facebook but it is currently fastest growing service and it is predicted to cover 1/3 of the World population within 2019.[4,5]

Our main area of interest will be the Facebook posts and tweets with links. 50000 links are being shared per second on Facebook. Tweets that have links are 86% more likely to be re-tweeted. Adding the number of followers for the user, we can very well say that the link is trending. Taking into account of all these factors, the algorithm was formulated to calculate the most trending website at present.

3 Proposed Method:

3.1**Facebook Analysis:**

Facebook, like any Online Social Network(OSN), can be represented as a graph. Nodes of this graph represent users and edges represent connections (say, Friendship/Follower).It is possible to represent users and connections as an unweighted, undirected graph. Accurate information is obtainable only if the entire graph is analysed, but doing so would involve high computational overhead. To avoid this, it is better consider a snapshot of its structure.

Uniform crawler, as proposed by Gjoka et al. [6], Catanese et al. [7] is one of the proposed methods for Facebook crawling. Facebook assigns a 32 bit user ID number for each of its users, although it is visible to users as an alphanumeric id(for eg: https://www.facebook.com/100002891397823 is visible to users as https://www.facebook.com/Chandler.Bing.1). For uniform crawling, Rejection sampling methodology is used wherein a list of randomly generated user ID numbers are requested from Facebook. Currently the number of active Facebook users is just over 1 billion(approx. 2^{30}). Statistically, the probability of the randomly generated ID to match with that of an existing user is $2^{30}/2^{32} = 0.25$. In other words, this technique finds an existing user once every four attempts. Once an existing user is found, his/her personal friend list is extracted. One main advantage of this technique is that the

random generation of IDs ensure independence of the chosen IDs with respect to friendship.

Another commonly used traversal algorithm is the Breadth-first-search. This starts from a seed profile (the profile that is logged into).This is followed by extracting the friends list of the seed profile and enqueuing them in a FIFO queue. These queued profiles are visited in order to retrieve further sub-levels of friendships.

One limitation that is common to both these technique is the security measure adopted by Facebook. When friends list is requested from Facebook, it responds with a maximum of 400 friends.This can be overcome by scraping information directly from the page by using a bot (to simulate human behaviour). Although, the legality of the above mention technique is debatable.

3.2 Twitter Analysis:

First, all tweets with links have to be streamed. Also we have to know the number of followers for each user who tweets and the time at which the link was shared. We also try to find the gender of the user using the username which gives an insight on the viewing audience. Finding the gender has many drawbacks because a Tweet does not give gender so we will be guessing the gender which is again very erratic due to pseudo usernames. But not everyone uses pseudo names. When the number of shares is really large, the marketers gets to know whether his/her target is male or female. Once all these details have been extracted from Tweets, we try to rank the websites using a set of formulas mentioned below.

$$TotalCount = TotalCount + Number\ of\ Followers \quad [1]$$

$$Weight = TotalCount \div (CurrentTime - LastSeen) \quad [2]$$

$$TotalCount = TotalCount - (Weight * 60) \quad [3]$$

$$NormalisedWeight = Weight \div N \quad [4]$$

$$N = \sqrt{(\sum (Weights)^2)} \quad [5]$$

3.2.1 Explanation:

1. TotalCount is the measure which gives us an idea of how many people the link has reached. Even though TotalCount doesn't give us the exact picture of how many people have clicked the link to visit the content of the website, a link with more TotalCount has a higher probability of site visits.
2. The time at which the tweet/post is shared is noted for each link. The current time minus the latest time of the link shared is the difference.The difference is a measure to show how popular the link is at that particular point of time.The popularity of a website is inversely proportional to the difference. Thus we arrive at the formula 2.
3. The link with more weight is a more popular web page. But formula 2 does not take into account links that have reached a lot of people at a time (say

TotalCount is really high). And for some months the link is not shared. Then one tweet/post shares this link. Thus the Difference will be very small, giving us an improper picture of the popularity of the site. To overcome this, we introduce formula 3.This will periodically reduce the Total Count. And ensure that the time factor is also included while keeping track of the popularity.

4. But in formula 3 we face a problem where the upper limit and the lower limit of the weight is not defined. To resolve this we introduce the concept of Normalised weight in formulas 4 & 5. Now the lower limit is 0 and the upper limit is 1.This can be used to delete links which fall below a certain normalised weight.

3.2.2 Time Complexity:

To find top k popularsites at a given point of time. We need to find the top k links with the highest weight. Instead of sorting the whole database which has a worst case complexity of O(n*log n) to O(n^2). We can index the site by searching for the link with maximum weight by going through the whole database k times. If the size of the database is n. The worst case complexity is O(k*n).

3.2.3 Categorisation of Webpages:

Once all links have been indexed, we have to find the category to which the webpage belongs. This is to make the marketers' job easier. Ranking the top viral websites in each unique category helps the marketer to decide on which webpage is more suitable to advertise his/her product. For this purpose, we use uClassify, an open source API. The classification is done with Naïve Bayesian classification as the core. The probabilities of a webpage to belong to a class is obtained as the result. The primary category of the website is determined based on the highest probability.

A custom made classifier with 10 different categories was trained based on the content of the webpage. The training sets contain samples of popular websites under each category.

3.2.4 Implementation:

In order to extract the tweets from Twitter, we use Twitter4j API, an unofficial open source Java library for the Twitter API. It is used to integrate java application with the Twitter service. Once stream is established, we are selecting only the English Tweets with links. Most of the links extracted are shortened links, hence we have to expand them before feeding them into the database by calling expandurl function in Twitter4j API.This is done using the filter module shown in figure 1.

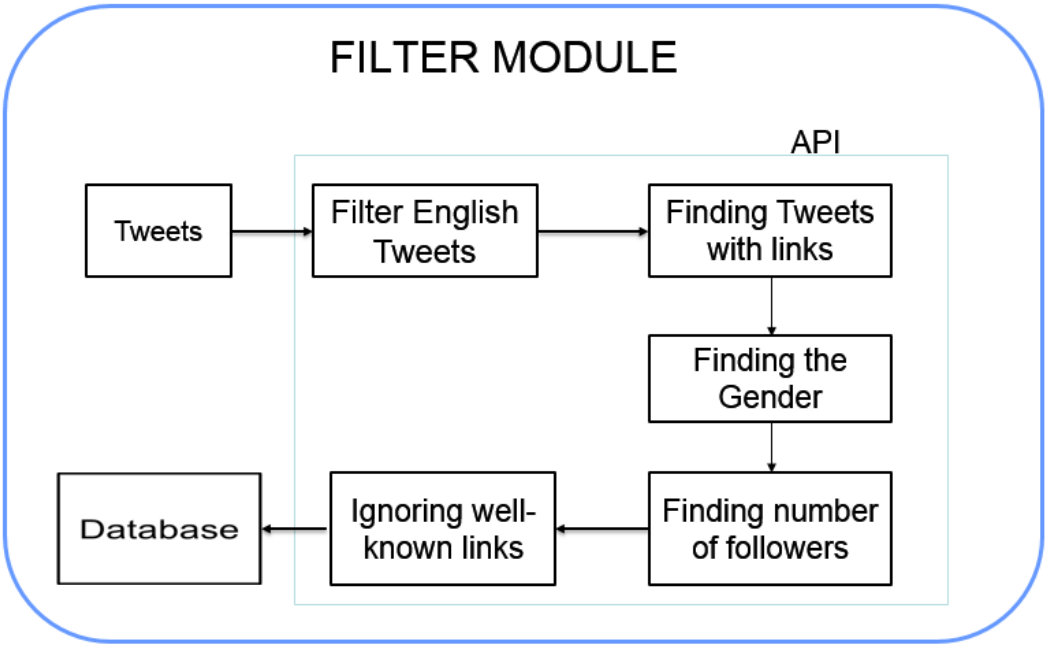

Figure 1: Filter Module

The extracted links are fed into the database along with the TotalCount (formula 1) and gender. We find the gender using NamSor Omonastic API which is a name recognition API which is accurate up to 87% for English names.Tweets have to be streamed for an entire day and periodically (for every 1 hour) weight for a link should also be calculated and TotalCount has to be updated based on the formulas proposed. The link with the highest normalised weight will be the most trending website. The extracted URLs are then fed to the classifier function.

4 Results:

Since Twitter has restricted the amount of tweets (around 3000) to be streamed per day, we were able to run the algorithm for 30 minutes. As a results, we got the most trending linksbased on the normalised weight as shown in figure 2 and figure 3. Figure 2 shows the database table which has columns for links, category, total count, normalised weight, male-count, female-count and last seen time.

domain_name	Category	total_count	weight	normalised_weight	male_count	female_count	last_seen
http://gofundme.com/savethekitchen	Society	2552942	141830	0.0082445	330	746	2014-11-18 22:01:32
http://www.thebrisky.com/hot-celebrity-rides/	Adult	230027	6970	0.00332	400	20	2015-02-04 08:46:08
http://af12.mindpics.com/top-destinations-to-save-up-for-2015/	Entertainment	104700	5816	0.00299	250	345	2014-11-18 22:01:40
http://rcobanus.greatpix.me/mind-twisting-quotes-that-will-change-you-forever	Entertainment	13522	4507	0.001198	115	90	2015-02-04 08:16:47
http://www.amazon.com/IMPOSTER-The-Protectors-Series-Book-ebook/dp/B004JHZ1NQ	Business	90975	4135	0.00115	210	67	2014-11-18 21:57:17
http://affiliates.chacha.com/custom/dlmi.php?aid=CD27318&bid=36275&uabid=36274&ibid=36273&subid1=DopeBoiBrandon	Recreation	56975	2713	0.00098	91	163	2014-11-18 21:58:05
http://www.bbc.co.uk/news/world-middle-east-31124166#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa	News	63237	1621	0.0009	345	29	2015-02-04 08:40:06
http://www.theguardian.com/sport/2014/nov/18/horse-racing-tips-wednesday-19-november	News	25947	1621	0.0009	124	244	2014-11-18 23:03:22
http://www.bbc.co.uk/news/world-us-canada-31124170#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa	News	63237	1580	0.00089	101	239	2015-02-04 08:39:51
http://www.bbc.co.uk/news/world-asia-31124837#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa	News	63237	1580	0.00089	234	20	2015-02-04 08:39:57
http://missdd.me/do-you-have-trypophobia-1	Health	51286	1192	0.00074	112	132	2015-02-04 08:36:25
https://docs.google.com/a/democracypr.com/document/d/1mQ8aBj4u7GdVzTOzD4DkPQDBOzyRF	News	7731	966	0.00073	222	23	2014-11-18 23:43:58
http://www.entweak.com/gaming/guild-wars-2%E2%80%B2s-tangled-paths-update-goes-live-today/?utm_source=twitterfeed&utm_medium=twitter	Games	9331	933	0.000733	91	10	2014-11-18 23:09:05
http://www.mb103.com/lnk.asp?o=6712&c=918271&a=159213	Entertainment	14286	892	0.00071	20	4	2014-11-18 23:03:31
http://aroundthefoghorn.com/2015/02/01/san-francisco-giants-rely-hunter-pence-balance/?utm_source=dlvr.it&utm_medium=twitter	Sports	4205	841	0.00071	23	34	2015-02-04 12:14:13
http://simonshinerocks.blogspot.co.uk/	News	4644	774	0.00068	657	1	2015-02-04 12:13:32
http://www.amazon.com/Preppers-Road-March-Ron-Foster/dp/1466225394	Business	21237	707	0.00068	245	876	2015-02-04 08:49:15
http://iexp.in/index.php?id=Kcj124944	Business	10817	676	0.00063	45	345	2014-11-18 23:03:05
http://amazingpic121.dailyfunnypics.me/r/2HkJs	Entertainment	19511	573	0.00061	33	11	2015-02-04 08:45:42
http://www.nanjingnk.com/redirect/?url=http://FLIGHTX9235.APPSPOT.COM	Business	2619	523	0.00059	6	1	2015-02-04 12:14:53

Figure 2: Database table

The top 25 Viral websites(according to February,2015) are as follows:
1.http://gofundme.com/savethekitchen
2.http://www.thebrisky.com/hot-celebrity-rides/
3.http://af12.mindpics.com/top-destinations-to-save-up-for-2015/
4.http://rcobanus.greatpix.me/mind-twisting-quotes-that-will-change-you-forever
5.http://www.amazon.com/IMPOSTER-The-Protectors-Series-Book-ebook/dp/B004JHZ1NQ
6.http://affiliates.chacha.com/custom/dlmi.php?aid=CD27318&bid=36275&uabid=36274&ibid=36273&subid1=DopeBoiBrandon
7.http://www.theguardian.com/sport/2014/nov/18/horse-racing-tips-wednesday-19-november
8.http://www.bbc.co.uk/news/world-middle-east-31124166#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa
9.http://www.bbc.co.uk/news/world-asia-31124837#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa
10.http://www.bbc.co.uk/news/world-us-canada-31124170#sa-ns_mchannel=rss&ns_source=PublicRSS20-sa
11.http://missdd.me/do-you-have-trypophobia-1
12.https://docs.google.com/a/democracypr.com/document/d/1mQ8aBj4u7GdVzTOzD4DkPQDBOzyRFyGEbak4lLVlL6k/edit
13.http://www.entweak.com/gaming/guild-wars-2%E2%80%B2s-tangled-paths-update-goes-live-today/?utm_source=twitterfeed&utm_medium=twitter
14.http://www.mb103.com/lnk.asp?o=6712&c=918271&a=159213
15.http://aroundthefoghorn.com/2015/02/01/san-francisco-giants-rely-hunter-pence-balance/?utm_source=dlvr.it&utm_medium=twitter
16.http://simonshinerocks.blogspot.co.uk/
17.http://www.amazon.com/Preppers-Road-March-Ron-Foster/dp/1466225394
18.http://iexp.in/index.php?id=Kcj124944
19.http://amazingpic121.dailyfunnypics.me/r/2HkJs
20.http://www.nanjingnk.com/redirect/?url=http://FLIGHTX9235.APPSPOT.COM
21.http://tmzlivenow.com/look-at-these-images-twice-photo-album
22.http://crazyfunstuff.com/deflated-dreams-a-gallery-of-celebrity-booty-implants-looking-absolutely-horrible-1
23.http://www.kj103fm.com/main.html#ldr_widget
24.http://teespring.com/invites/2xqduc
25.http://feeds.feedburner.com/~r/infowarsnewsfeed/~3/aL_toJy8nVo/?utm_source=feedburner&utm_medium=twitter&utm_campaign=earththreats

Figure 3: Ranked Websites

5 Conclusion:

Given the current trend, a phenomenon like social network crawling shows great potential in the development of Digital Marketing. In our work, Facebook and twitter, the two most widely used online social networking websites, were chosen. As a result of the restrictions imposed by Facebook and a vast portion of the profiles being marked as private, Facebook proved to be an unviable option. On the other hand,

twitter provided a far friendlier environment to base our work on. The most trending websites were extracted with the help of our algorithm. The algorithm takes into consideration the transient nature of popularity of websites and bases its output on the same. With Facebook and twitter granting more access to their content, this work can be further improved upon by installing a dedicated server to run the algorithm throughout the day.

References:

[1] Tart, N.(n.d.).How much traffic a website gets.*incomediary.com.* Retrieved 9 April 2015, from www.incomediary.com/how-much-traffic-website-gets

[2] Bullas, J.(17 Jan. 2014). Social media facts and statistics you should know in 2014.*JeffBullas.com.*Retrieved 9 April 2015 fromwww.jeffbullas.com/2014/01/17/20-social-media-factsand-statistics-you-should-know-in-2014

[3] Nagarkar, V.(15 May 2014). Why twitter stock will not go the Facebook way.*AmigoBulls.com.* Retrieved 9 April 2015, from www.amigobulls.com/articles/why-twitter-stock-will-not-go-the-facebook-way

[4] Cooper, B. B.(19 Aug. 2013). Twitter statistics to help you reach the followers.*BufferApp.com.* Retrieved 9 April 2015, from www.blog.bufferapp.com/10-new-twitter-stats-twitter-statistics-to-help-you-reach-your-followers

[5] International Association of Chiefs of Police. Fun Facts.*Iacpsocialmedia.org.*www.iacpsocialmedia.org/Resources/FunFacts.aspx (Accessed 9 April 2015).

[6] Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou,A. (2010) Walking in facebook: A case study of unbiased sampling of osns. INFOCOM, 2010 Proceedings IEEE, pp. 1-9. IEEE.

[7] Catanese, S. A., De Meo, P., Ferrara, E., Fiumara,G., and Provetti, A. (2011) Crawling facebook for social network analysis purposes. Proceedings of the international conference on web intelligence, mining and semantics 52. ACM.