Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base

Marc Cheong
Faculty of IT, Monash University
Melbourne, Australia
marc.cheong@infotech.monash.edu.au

Vincent Lee
Faculty of IT, Monash University
Melbourne, Australia
vincent.lee@infotech.monash.edu.au

ABSTRACT

Twitter as a microblogging platform has vast potential to become a collective source of intelligence that can be used to obtain opinions, ideas, facts, and sentiments. This paper addresses the issue on collective intelligence retrieval with activated knowledge-base decision making. Our methodology differs from the existing literature in the sense that we are doing analysis on Twitter microblog messages as opposed to traditional blog analysis in the literature which deals with the conventional 'blogosphere'. Another key difference in our methodology is that we apply visualization techniques in conjunction with artificial intelligence-based data mining methods to classify messages dealing with the trend topic. Our methodology also analyzes demographics of the authors of such Twitter messages and attempt to map a Twitter trend into what's going on in the real world. Our findings reveal a pattern behind trends on Twitter, enabling us to see how it 'ticks' and evolves though visualization methods. Our findings also enable us to understand the underlying characteristics behind the 'trend setters', providing us a new perspective on the contributors of a trend.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous, I.5.3 [**Pattern Recognition**]: Clustering, J.4 [**Social And Behavioral Sciences**]: Sociology.

General Terms

Experimentation, Human Factors, Measurement.

Keywords

Twitter, collective intelligence, decision making, knowledge base, memetics, microblogging, trend analysis.

1. INTRODUCTION

Twitter [1] is a microblogging platform which is fast gaining popularity [2] among broad sections of society and has a global outreach spreading from developed, urban nations the likes of the United States where it has a high adoption rate [3], to developing countries in parts of Asia and South America.

The basic ethos of Twitter, in its founders' words, is "*for friends, family, and co–workers to communicate and stay connected through the exchange of quick, frequent answers to one simple question:* ***What are you doing?***" [1]. It allows users to post 'tweets' – Twitter messages, up to a maximum length of 140 characters.

However, studies [3, 4] have shown that Twitter users not only use Twitter to answer the question 'what am I doing?', but also leverage it as a means of communication and social networking; making it look like a hybridization between conventional blogging and an online social network. To illustrate: users post numerous forms of tweets, from mundane updates about their daily lives, rebroadcasting breaking news stories, to summoning help in times of crisis [2]. 'Tweeting' and 'tweeters' have become common terms referring to the act of publishing tweets and the people who use Twitter.

One interesting thing about Twitter is that it has a section on its main page entitled "Trending Topics", which displays the top 10 mentioned terms on Twitter at any given moment. This is generated based on Twitter's proprietary algorithm, but nonetheless provides an interesting *zeitgeist* into the events talked about by the Twitter community. An example of a typical Trending Topics list would be (in order of decreasing popularity):

"New Moon, #mtvmovieawards, MTV Movie Awards, Eminem, Bruno, Ben Stiller, Twilight, Susan Boyle, #Phish, Kiefer Sutherland".

In this paper, we attempt to dissect the anatomy of a trending topic to find out what makes it 'tick'. Specifically, we select at random 4 topics which appear in the top 3 category of the trending topics list and 2 control topics (which are non-trending), gather information about the posts mentioning the topic to a maximum of 1500 posts, and from there we analyze the data to investigate any patterns that occur in trending topics.

The motivation behind our work is that social networks and microblogs which are products of Web 2.0 are becoming the tool of choice for information dissemination, sharing and interpersonal communication and networking. It is a new platform being rapidly adopted by all walks of life; from politicians [5] to businessmen; young and old; for use in citizen journalism [6] to being a medium to stay close to friends and family [2]. By harnessing this information from online social network and microblogging sites such as Twitter, we can obtain an understanding about the collective "wisdom of crowds" [7], and leverage its data in policymaking, decision support, economic analysis, epidemic behavior (the "tipping points" theorem [8]) and various other applications.

SWSM'09, November 2, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-806-3/09/11...$10.00.

2. RELATED WORK

Work on conventional blogging in general include Thelwall [9] who analyzed social network content for mentions of news articles, and Glance et al. [10] who performed work on BlogPulse to monitor trends in blogs.

Fukuhara et al. [11] expanding on [10], have performed work similar to this paper, but focusing mainly on conventional blogs as opposed to our research on microblogs (Twitter). In their work, they have successfully identified "patterns of social concerns", and a correlation between "blog and real world temporal data" such as temperature and news articles. Gruhl et al. harnessed the "predictive power of online chatter" by correlating mentions of book titles in blogs to the movement of the book's selling power in real-world Amazon.com sales rankings [12]; and also "formalized … the notion of 'spike' topics generated by outside world events" as well as those in the community [13]. In the financial markets, Choudhury et al. [14] discovered a correlation between chatter in technology blogs with the actual stock market movement of technology-related companies.

Findings from [11-14] above suggest that blogs can be harnessed as a form of 'collective intelligence' to predict and forecast real world-events. Prior research on the microblogging platform Twitter has been conducted by Java et al. [3], who focused on the Twitter user spread in terms of geographic location, social networks a user is part of, and a taxonomy of "user intentions" when microblogging on Twitter. Krishnamurthy et al. [15], expanding on [3], have performed research on identifying emergent properties of "distinct classes of Twitter users and their behaviors" and also analyzing the growth of the Twitter user network on a region by region basis. The "user intentions" of Twitter users are analyzed from a humanities perspective by Mischaud [4], who states that Twitter "appears to be very much a part of the [users] who use it to send out random thoughts and details about their daily lives".

Meanwhile, a similarity between Twitter chatter and blog postings have been suggested in [3, 4], leading us to conduct research on Twitter (as a microblogging platform) to be able to provide an insight into how the Twitter community's chatter accurately reflects the properties of real-world events represented in the form of trending topics. Gruhl's research [13] has also covered characterization of the individuals contributing to a surge of a 'spike' topic in the blogosphere; however his work is limited to profiling them based statistically on the number of posts that they have written.

It is to be noted that our work differs from existing studies in a few aspects. Firstly, we emphasize the use of microblog posts (Twitter's 'tweets') rather than conventional, full-sized blogs. Secondly, the methodology behind our analysis is with the application of visualization techniques and AI-based data mining approaches (such as Kohonen's self-organizing maps or SOMs [16]). Thirdly, we not only explore the properties and features of a trending topic, but also the properties of the users (or 'trend setters') that contribute 'tweets' regarding a trending topic, which makes them a part of the trend.

3. METHODOLOGY & PRELIMINARIES

3.1 Topics surveyed

The Twitter public timeline is observed for trending topics on the week beginning 11th May 2009. We use a Perl script that harnesses the capabilities of the Twitter Search API provided by Twitter [1] to search for four trending topics (chosen at random) on the 14th to 15th of May 2009, and two non-trending terms as a control (Table 1 and 2). The end of the work week is chosen as the time to harvest data as it has been identified [2] to be immediately following the days in which Twitter activity levels are at a peak.

Table 1. List of trending topics selected

Keyword	Description
Grey's Anatomy	Trending topic on the 15th of May 2009. This refers to a television drama series which just had its season finale aired on primetime television.
H1N1	Trending topic that first peaked on 1st-2nd of May, and peaked again on the 15th. H1N1 refers to the Swine Flu pandemic.
Nizar	Trending topic on the 11th of May 2009. Nizar is the name of a politician involved in a constitutional crisis in a Malaysian state [17].
TwitHit	Trending topic on the 15th of May 2009. This keyword appears as a result of Twitter users who supply their login credentials to a dubious spamming site [18].

Table 2. List of non-trending topics (control) selected.

Keyword	Description
Coffee	Non-trending but common topic, included here for comparison.
Revolverheld	Non-trending and relatively obscure topic (a German alternative rock band), included here for comparison.

3.2 Data for trend pattern detection

Among the constraints of the Twitter Search API is that it returns up to a hard upper bound of 1500 tweets, and a soft limit on the date range (approximately backdated to about 20 days) due to technical limitations. Therefore the maximum search results are obtained by topics with more than the 1,500-tweet limit; else the search is only constrained to the earliest date available for capture. For all the topics above, 7215 total tweets have been collected, with only the control term 'Revolverheld' having less than 100 tweets due to the fact of its obscurity.

The raw results available from the Twitter Search are dumped into a comma-separated value file for ease of parsing. In each result, 12 attributes – among them the 140-character message, the username, and the timestamp – are obtained. In addition to the 12 attributes, 3 Boolean attributes are synthesized from the 140-character message text via the use of a "tagger" Perl script:

Retweeted: represents the presence of the keyword "RT" in a message, which is used by a user to 'forward' another tweet by another user, akin to, a "Fwd:" tag in the subject body of an email. This behavior is included in our dataset as we try to see the significance of how a message 'retweet' contributes to the dissemination of a trend, in the same way the habit of email forwarding [19] works.

Replied: the presence of an 'at' (@) symbol in the tweet is used to signify a reply to a user, in the format "*@username this is a reply message*". Research has been done [20] which indicated approximately 96% of all usages of the 'at' symbol in Twitter messages exclusively deal with the notion of communication with another user or group of users. This is important as we try to analyze a trend in the perspective of Twitter users composing tweets to another user as contributing to the overall development of a trending topic.

Trended: this detects the presence of the word 'trend' in a message in the context of the user 'piggybacking' on the trending topic but not genuinely discussing the topic in a proper context. Examples of such messages would be "*xyz is now a trending topic!*" The set of messages in which this behavior occurs is directly contributing to the increase in rank of a trending topic, creating a 'self-fulfilling prophecy'.

3.3 Demographics of Twitter users contributing to a trend

Besides the data about tweets mentioning the topics above, data about the users' postings are also collated.

In each topic above, usernames who posted data on each of the thousand-odd messages are identified and filtered out to remove duplicates. From each of the unique user lists, a representative sample of approximately 13% of the population is randomly selected, and a simple HTML file containing the URIs of the user profiles are generated. Both steps above are automated with a "sampler" Perl script.

The following pieces of data are then obtained: device used for Twittering, gender, primary usage pattern, and country. All of the data (except the device used for Twittering, which is available by cross-referencing the data from 3.1) are obtained by visiting their Twitter profile page of the users. Note that the data is not verified against any third-party source but taken 'as is' written by the users themselves.

Client and device used is available from the 'tweets' data obtained (as discussed in Section 3.1 above). The device (computer, mobile phone, or culled from external data sources) can be ascertained from the codename of the Twitter client application used (similar research has been applied with respect to this aspect [3]).

Gender is identified by the writing style of the user (e.g. "username misses his/her friends") in the profile information and also by the profile image that is publicly available on Twitter. Besides the male and female sexes, a third category, the neuter gender is included for Twitter users that are created by a workgroup or an organization.

Primary usage pattern is based on a survey on the first page of that user's Twitter updates; and if inconclusive, a visit to the user's homepage (publicly available as a link on their profile). They fit into the following categories:

1. **Personal**: majority of the postings are personal communication, and social networking; examples would be messaging friends, sharing information.
2. **Group**: would be a not-for-profit user group with common interest, such as fan clubs or for researchers to network.
3. **Aggregator**: predominantly publishing or collating information as part of their job (for example news agencies, Twitter accounts linked to RSS feeds, politicians' message to his constituents) with little or no personalized content or messaging performed.
4. **Satire**: a page created for humorous, satirical, or parodying purposes.
5. **Marketing**: a page created to push a product; however, the majority in this category comprise of spam, unsolicited postings, and possibly harmful sites.

Country is based on the user profile's 'location' field (which can take a form of a city – 'Adelaide' for instance, or a GPS coordinate). Sometimes, the 'location' is located in the user's homepage URL as provided by the user.

4. OBSERVATIONS

4.1 Anatomy of a trend

We conduct our testing on the full set of 7215 tweets containing trend- and non-trend-related keywords as in Section 3.2. From that we could identify several patterns of topics for both trending or 'spiking' [13] topics and non-trending topics. Instead of using the date and time directly as the manipulated variable on the x-axis, we instead have another approach – using the unique message identifier (UID) generated by Twitter for each message. No prior research as far as we have seen deals with using the UID as a measure of time. The benefits of using the UID instead of time include its relative ease of use, and the frequency of UID generation over time is more or less stable. Should the frequency of UID generation increase or decrease (for example due to increased popularity of Twitter, or conversely, declining usage) in the future, the UID frequency can be a reliable indicator for future trend/spike analysis. The UID frequency (determined by ratio of the date range and UID difference between the first and last messages for each of the 6 case studies and obtaining the average) is approximately 111 UIDs per second, which is a good reflector of the current rate of message flow on Twitter.

The emergence of "Retweet" messages is obvious in almost all of the trending topics noted, contributing to the overall chatter of a trend not unlike how email forwards [19] and blog linking [21] behave in contributing to the memetic spread of a topic. "Replies" are predominantly found on topics with high user interaction (with the exception of TwitHit, as will be discussed later). The interesting part is the prevalence of the keyword "Trend" only on topics which already have been included in the Trending Topics list: messages such as these are usually discussing about the trend or 'piggybacking' on the term to generate more views, typically tactics employed by aggressive marketing campaigns and spammers.

We classify each of our case studies into 3 categories:

1. **Long-term topics**: such topics are sparsely discussed about due to obscurity, and should they have any spike, the spike will decay relatively quickly. Topics like this

can be accessed by the Twitter Search API up to approximately 20 days, but rarely exceed the maximum retrieval results of 1500 tweets.

2. **Medium-term topics**: such topics are either generic terms which are commonly talked about but do not warrant a high number of tweets; or sustained 'trailing patterns' [11] due to a pre-existing spike that occurred beyond the API-imposed 1500 tweet boundary, but the discussion on the topic is trailing. Such topics can range for a period from half a day to approximately a few days.
3. **Short-term topics**: such topics are high volume in nature and can be a very commonly talked about term which does not exhibit spiking/'bursting' behavior; or topics captured using the Twitter API in the middle of a spike. Topics like these, during the moment of data capture, will only backdate up to a few hours when accessed. Topics as these can be categorized as those in the 'graduated increase pattern' or in the middle of a 'periodic pattern' [11].

4.1.1 Long-term topics

Several topics have a relatively low occurrence in the public timeline. The control term 'Revolverheld' (a German alternative-rock band) was used to study the trend of an obscure, non-trending topic, while the trending topic 'Nizar' (Malaysian politician) was chosen to study the trend of a quick spike. Quick spikes are better known as *the Slashdot effect* (named after the Slashdot website where featured articles gain a spike in popularity), also termed by Fukuhara et al. as a *sensitive pattern* [11, 22].

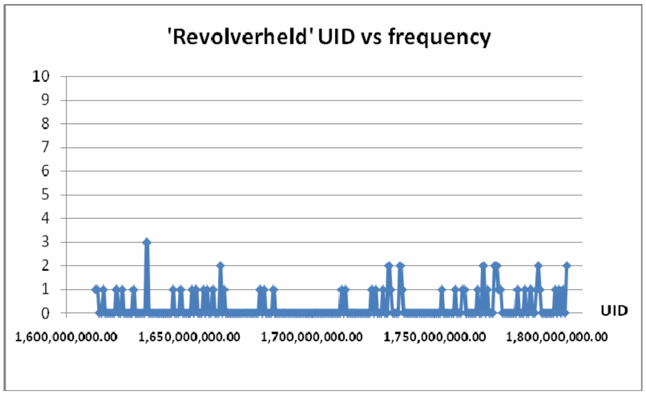

Figure 1. UID vs frequency plot for 'Revolverheld'

Figure 1 refers to the pattern captured by the obscure topic 'Revolverheld', captured over a period of approximately 20 days. The obscurity of the topic 'Revolverheld' is seen by the fact that the frequency of mentions of this topic remain at a minimum level – 3 or lower per 770 thousand UIDs = approximately a period of 2 hours.

'Nizar' the trending topic, on the other hand, spiked at 11th May at approximately 07:00 GMT directly corresponding to the real-world event of the Malaysian courts passing judgment on Mr. Nizar's case [17]. The spike registered 329 topics during a window of approximately 2 hours. Over the observation period, 'Nizar' peaked to #3 on the Twitter trending topics list before gradually fading off, corroborating the sensitive decay pattern studied in [11].

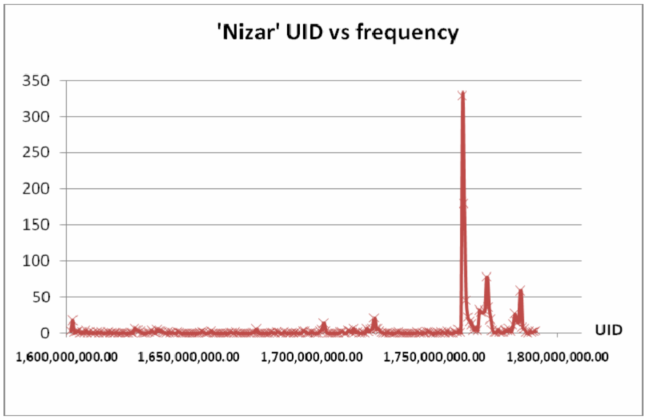

Figure 2. UID vs frequency plot for 'Nizar'

4.1.2 Medium-term

Medium-term topics have a significantly shorter range of UIDs (and consequently, time) compared to long-term topics. Medium-term topics when fetched from the Twitter API normally reach as far back as the hard retrieval limit of 1500 tweets.

The data from Figure 3 is the data for Twitter chatter about the 'H1N1' swine flu pandemic, a Trending topic. This data is captured in the middle of the Trend. This trend is classified as a 'trailing pattern' [11], as activity regarding this topic has sustained its inclusion in the Trending Topics list ever since the outbreak of H1N1 began in early May 2009.

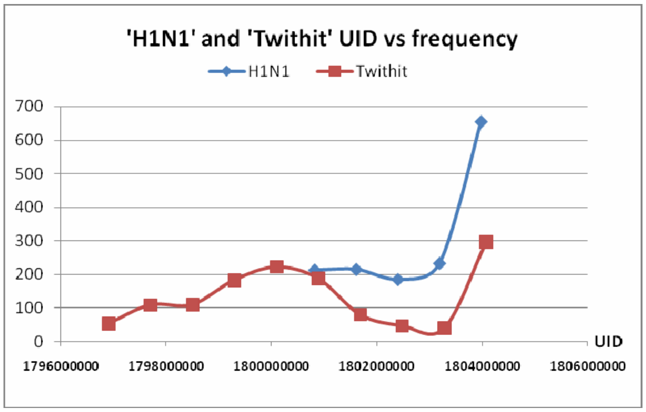

Figure 3. UID vs. frequency plot for 'H1N1' and 'TwitHit'

The UID for such a medium-term topic has a range of 3.1 million UIDs (compared with the short-term topics with range of 200 million UIDs). Based on the same histogram interval as above – per 770 thousand UIDs, approximately 2 hours – we see that medium-term trending topics generate hundreds of mentions, which corroborates with our definition of a spike in the case of the 'Nizar' keyword in the previous section.

Another occurrence of medium-term topics with spiking tendencies can be found in the case of the keyword 'TwitHit' (Figure 3) which stemmed from spamming activity from a dubious website [18]. This trend can be roughly categorized as per Fukuhara et al.'s definition [11] of a 'sleeper hit' – a topic which rose in popularity from relative obscurity to a trending topic. The 'sleeper hit' pattern is discovered in many such Twitter trends, for example mentions of Twitter scamming epidemics, or an unexpected catastrophe of great significance to the news. The range of UIDs for this term is approximately 8 million UIDs (nearly triple the one from H1N1); however it still fits our definition of a medium-term trend (the same histogram plot interval of approximately 770 thousand UIDs or 2 hours applies).

For this case study, we capture the data from the first occurrence of this spamming outbreak. From the scatter plot, we can easily see that the 'spike' or trending characteristic of the keyword begins at the 4th data point (approximately 8 hours from the start of the outbreak). It is to be noted that for cases such as these, the occurrences of "reply" tweets do not occur until near the end of the trend, as the first part of the message growth trend is generated by automated spamming methods; "replies" at the end are about users discussing about their experiences being hit by the scam.

4.1.3 Short-term

Finally, certain trending topics belong to the short-term category. This means that the retrieval limit of 1500 messages from the Twitter API stretches barely 4 to 5 hours back. This is the result of a disproportionately large amount of messages generated by Twitter users relating to the highest trending keywords on Twitter at any given moment.

Trends such as these include the high chatter succeeding the season finale of the US drama series "Grey's Anatomy" (Figure 4). The range of UIDs for short-term trending topics such as this is relatively the smallest (950578 UIDs, a period of approximately 2.4 hours). The histogram interval used in graphing this topic is approximately 190 thousand UIDs, or roughly 30 minutes; the data capture for this trend is obtained after it jumped to first position.

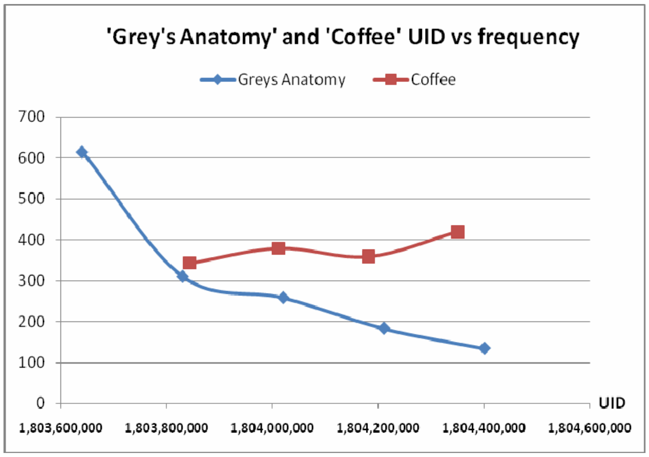

Figure 4. UID vs frequency plot for 'Grey's Anatomy' and 'Coffee'.

Note that the voluminous number of tweets make it impossible to go beyond the 1500 message boundary, as such, our range of UIDs (and correspondingly, time) is rather limited. It is important to note that the keyword 'coffee', although it exhibits behavior of a 'long-term trending topic', is not defined as a Trending Topic. This is because 'coffee' is a relatively common term in everyday vocabulary; such terms are removed from consideration in spike detection studies for being too common and not being a proper noun [13].

4.2 'Trendsetter' demographics

From the 7215 tweets published in our dataset, we calculate the number of unique users contributing to the public message timeline, removing duplicates. We then randomly sample 485 unique users and obtain demographic information from their profile as detailed in the Methodology section.

4.2.1 Message uniqueness

In the process of determining unique users in our dataset, we have identified a trend in the case of the 'TwitHit' search term, which is a result of a dubious spamming site. Of the 1328 messages obtained using the search API, only 622 unique users contributed to the message pool. This indicates that spam activity on Twitter caused by TwitHit and its equivalents comprise of similar repeated messages generated by an affected user.

Another trend, 'Nizar' (being the topic of 1328 tweets) only has 439 unique users contributing to the topic being trended is just 439. This is due to a number of participants corresponding back and forth ("@" messages) and retweeting sources of news ("RT" messages) as it happens.

4.2.2 Distribution of topics by Twitter client/device used

A majority of all users sampled contribute to the Twitter timeline using the main web interface at http://www.twitter.com/ [1]. However, the amount of users from mobile devices is significant, reflecting a shift to adopting social networking and microblogging sites from the computer to a mobile environment [23].

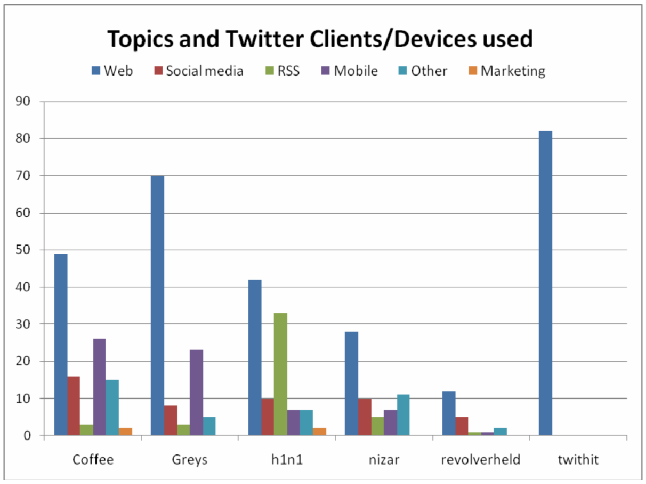

Figure 5. Distribution of topics by Twitter clients and device classes.

Another trend noticed from the data is that besides mobile devices and the Twitter main interface; users have also adopted Social Media applications – that work by synchronizing messages on various social media platforms beside Twitter, for example Facebook and Windows Live – suggesting that a section of Twitter users also participate in other Web 2.0 social networking platforms.

The usage of RSS and other Twitter content-generating/online marketing tools used as content, rather than users generating original content is evident in the case of the term 'H1N1', which indicates that part of the hype regarding the H1N1 flu virus is actually repeating same stories on Twitter, to inflate its risk and impact. For the case of trend 'TwitHit', the main Twitter web interface is used 100% in all samples, which indicates a possible exploitation of the Twitter web interface in propagating junk postings and spam.

4.2.3 Distribution of topics by country/geographic region

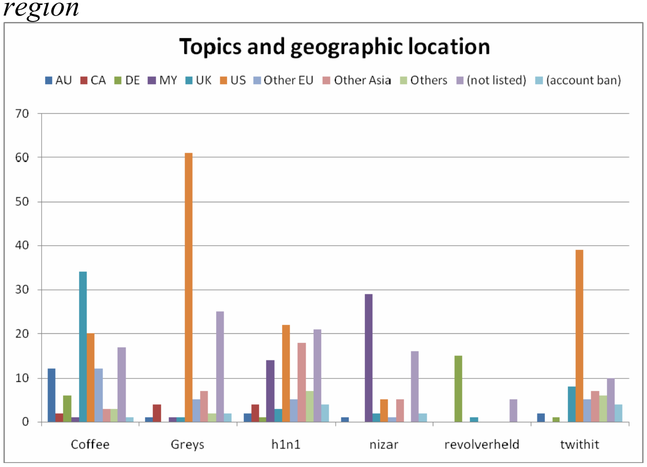

Figure 6. Distribution of topics by geographic location.

Many trending topics on Twitter are localized to a specific region, while others are issues of global concern and scope. In our analysis, there is a specific link between Twitter user's geographic information and the content of the tweets surveyed:

1. Coffee is mentioned primarily by UK and US residents, which coincides with breakfast mealtime near the GMT time zone.
2. Grey's Anatomy and Revolverheld are part of the US and German entertainment and media culture, hence it is naturally followed more closely by Twitter users of their respective countries.
3. Nizar is a Malaysian politician; Twitter messages mentioning him are mainly from Malaysian Twitter users.
4. H1N1, the swine flu epidemic, is a global issue, which describes its distribution of geographic location.
5. It is interesting to note that TwitHit has more of an appearance among US Twitter users, suggesting that TwitHit might have originated there.

4.2.4 Distribution of topics by gender

Gender information can also reflect the composition of the users that contribute to a topic. In the example of Grey's Anatomy, the target demographic of that particular TV show is among females, and is reflected on the large proportion of female Twitter users.

News stories such as political and environmental news are 'tweeted' by predominantly male Twitter users; discussion on the real world of gender dynamics in news production can be found in [24].

A separate category, 'neuter' which refers to a Twitter account owned by an organization or group such as a media agency or government department, is featured in a proportion of 'H1N1' messages. Our interpretation to this, based on reviewing the H1N1 sample set, is due to the role of organizations using Twitter as a channel for broadcasting updates and latest news regarding the swine flu pandemic.

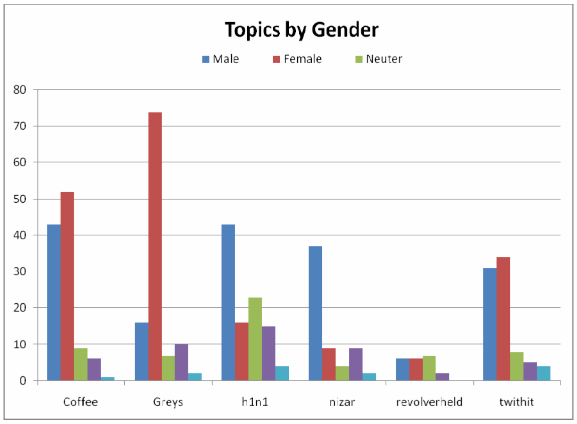

Figure 7. Distribution of topics by gender.

4.2.5 Distribution of topics by Tweeting habits

Finally, we look at the tweeting habits of the users contributing to aforementioned topics. The majority of Twitter users participating in chatter are users who talk about their personal life and use Twitter as a form of communication and social networking – this corroborates the findings on "main [Twitter] user intentions" [3].

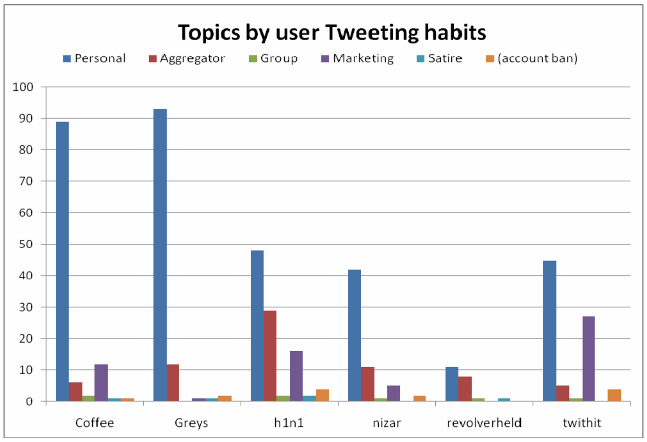

Figure 8. Distribution of topics by their users' Tweeting habits.

In the case of 'H1N1' however, much information is generated by 'Aggregator' users (i.e. users categorized as frequently rebroadcasting information, this relates to the discussion of using RSS feeds and marketing tools in the earlier section about Twitter clients). 'Marketing' users which comprise of users on Twitter which are aggressively promoting a product or website tend to contain junk posts or spam - this is evident in the "TwitHit" search term whose sole motivation is to 'phish' for user login credentials [18]. It is also interesting to note that many 'marketing' users piggyback on the trending topic 'H1N1', which in our investigation turns out to link to a website populated with a disproportionately large number of advertisements in an attempt to gain profits from page views.

4.3 Clustering of demographics

The usage pattern of Twitter in each of the 6 case studies above have been fed into an implementation of Kohonen's [16] Self-Organizing Map (SOM) algorithm, Viscovery SOMine. All the sample data for each of the cases (as categorized in section 4) are used as the training data for SOMine, while applying the program's default parameters for generating SOMs for clustering. SOM-based clustering can give us an idea of categorizing and clustering the users contributing to a trend based on their demographic data. This could potentially be useful in decision-support (e.g. policy making, socio-economic planning), where the clustered data can give us representative characteristics of the users contributing to a particular Twitter topic.

The attributes used in the SOM are as discussed in Section 3.3 - the Twitter client (aggregated by device platform or type), country (aggregated according to geographic area), gender, and Twitter user type. A set of 29 maps are generated for all the keywords studied in this paper, the final combination of clusters are shown in the subsequent sections and discussed on. Banned or suspended accounts are included in the following cases as a separate entity, denoted with an 'X' as they represent a significant class in studying the demographics of a trends knowledge base. Unknown or anonymous information is denoted with a '*'.

4.3.1 Long-term topics

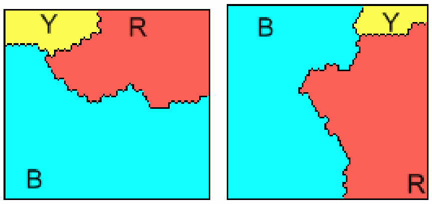

Figure 9. (a and b) Kohonen's SOM clusters generated for the long-term topics 'Revolverheld' and 'Nizar'.

The SOM for control (non-trending) topic 'Revolverheld' reveals that the majority of the users contributing to the chatter (blue cluster) are females in Germany who mainly contribute personal chatter on Twitter using the web interface. The red cluster represents German males/organizations which aggregate news regarding the 'Revolverheld' band using social media clients; and the yellow cluster depicts anonymous users (with no geographic location nor gender information accessible) contributing to the discussion anonymously.

For the long-term trending topic Nizar, the majority of the conversation is generated by Malaysian users (relevant, since the topic is a Malaysian news story) of both genders who mainly use Twitter for personal microblogging The red cluster consists of predominantly males from other countries, using Twitter as a form of 'citizen journalism' to aggregate and publish news. It is interesting to note that there are a proportion of users (the smallest cluster) which are organizations which either aggregate data or perform aggressive marketing while 'piggybacking' on a Trending search term; almost all feeds from this category of users are culled from RSS feeds.

4.3.2 Medium-term topics

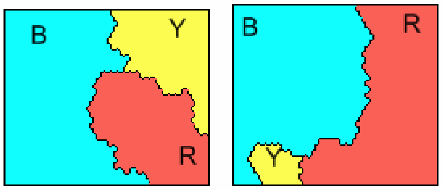

Figure 100. (a and b) Kohonen's SOM clusters generated for the medium-term topics 'H1N1' and 'TwitHit'.

For the medium-term 'H1N1' trending topic (which is a global affair), several interesting trends can be observed from the generated Kohonen SOM. The blue cluster comprising the majority of the user sample comprise of male Twitter web users who microblog about personal matters, situated in Malaysia, the United States, and other countries in Asia genuinely discussing about the flu pandemic. The yellow cluster consists predominantly of news aggregators (by organization-based Twitter accounts) sourcing data from RSS feeds that contribute to the heavy hype about the flu pandemic on Twitter; what brings attention to this cluster is that a subsection of this consists of users whose accounts have been banned by Twitter for account violation. The red cluster is closely related to the previous cluster, where the majority of users singled out in this cluster are 'marketing'-based 'anonymous' Twitter users including banned accounts whose sole modus operandi is piggybacking on the 'H1N1' topic for spamming and deceitful advertising purposes.

Studying the SOM for 'TwitHit' reveals the demographics behind users who fall prey to internet scamming/spam-based sites. The red cluster represents the majority of users falling prey to the scam – majority of users in this cluster are American regular Twitter users of both sexes who use Twitter for typical personal microblogging. The blue cluster represents dubious accounts which have the US and the UK as country of origin, using Twitter for mostly aggressive marketing and spamming activities – which possibly indicates the root cause of the problem. A small cluster of Australian-based users of Twitter who use Twitter as a form of social networking and also personal microblogging are the next affected set of users outside of countries on the Western side of the globe.

4.3.3 Short-term topics

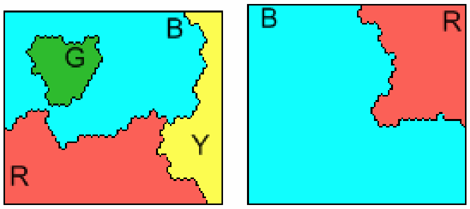

Figure 11. (a and b) Kohonen SOM generated for the short-term topics 'Grey's Anatomy' and 'coffee'.

We focus our attention on the short-term trending topic 'Grey's Anatomy' (for the US drama series). The cluster in red successfully reflects the demographics of the drama series – female Twitter users based in the United States, using Twitter mainly for personal microblogging. However, this cluster reveals some of the demographics that aren't obvious from cursory inspection – the majority of the Twitter users in this cluster actively contribute to Twitter not only through the web, but rather mostly through other social media clients and also using mobile clients, indicating a shift from traditional usage of Web 2.0 services such as Twitter from the desktop web environment to a more mobile, social-based environment [23]. The blue cluster represents users (also predominantly female and use Twitter for personal communication), however their geographic location spans the continents of Asia and Europe and their main contribution to the microblog chatter comes mainly from the web interface. The remainder of chatter on the drama series from the US – marked as a yellow cluster - comes from aggregator users, that either exhibit characteristics of collating news feeds from the entertainment/television industry, or part of marketing spammers 'piggybacking' on a trending term. The final cluster (in green) also reveals demographic data that isn't readily apparent; this group consists of (predominantly) female Twitter users using Twitter as a personal communications medium, entirely from Canada who use a hybrid of methods of posting to Twitter. Data such as the ones illustrated above are highly valuable to people in the media, advertising, and television production industry, which illustrates the motivation behind this research.

As for the 'coffee' keyword, the majority of users (to corroborate our observation from Section 4.2) are from the UK and the US from both sexes, who tweet about coffee in a 'personal' context, describing part of their daily routine (recall from Section 4.2 which stated that the Twitter message data was collected during breakfast time in the GMT+0 time zone). The remainder of the sample set (in red) comprises of Twitter user accounts involved in coffee-related marketing campaigns and news aggregation; some of the accounts in this cluster have been suspended or banned based on policy violation.

5. CONCLUSION AND FUTURE WORK

We have presented a new approach of analyzing Trend patterns for the Twitter microblogging platform in this paper. Our approach exploited the information retrieval on the collective intelligence characterized in the Twitter message pool and user base, and applied decision-making stratagem using the demographics of the set of Twitter users contributing towards the discussion of a particular trend. As demonstrated, this is potentially useful in the areas of business intelligence, marketing, epidemic research and other related fields.

Future work on this area include using the Twitter real-time message feed as opposed to using offline data, and also leverage other properties and attributes which are readily available by using the Twitter API for better decision-making and trend analysis. To generalize a trending-pattern theory, more empirical studies are underway.

6. REFERENCES

[1] Twitter Incorporated, "Twitter." 2009.

[2] T. O'Reilly and S. Milstein, The Twitter Book: O'Reilly Media, Inc., 2009.

[3] A. Java, X. Song, T. Finin, and B. Tsen, "Why We Twitter: An Analysis of a Microblogging Community," in 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis: Springer-Verlag, 2009, pp. 118-138.

[4] E. Mischaud, "Twitter: Expressions of the Whole Self," Master's Thesis. Department of Media and Communications (Media@LSE)., University of London, 2007.

[5] M. Harris, "barack to the future," in Engineering & Technology. vol. 3: IEEE, 2008, p. 25.

[6] H. N. Shams, "Twitter finally arrives in Malaysia," in The Malaysian Insider Kuala Lumpur: The Malaysian Insider, 2009.

[7] J. Surowiecki, The Wisdom of Crowds. London: Abacus, 2005.

[8] M. Gladwell, The Tipping Point: How Little Things Can Make a Big Difference. New York: Back Bay Books, 2002.

[9] M. Thelwall, "No place for news in social network web sites?," Online Information Review, vol. 32, pp. 726-744, 2008.

[10] N. S. Glance, M. Hurst, and T. Tomokiyo, "BlogPulse: Automated trend discovery for weblogs " in WWW 2004 New York, NY, 2004, pp. 1-8.

[11] T. Fukuhara, T. Murayama, and T. Nishida, "Analyzing concerns of people using Weblog articles and real world temporal data," in WWW 2005 Chiba, Japan, 2005, pp. 1-12.

[12] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The Predictive Power of Online Chatter," in 11th ACM SIGKDD Chicago, IL: ACM, 2005, pp. 78 - 87.

[13] D. Gruhl, D. Liben-Nowell, R. Guha, and A. Tomkins, "Information Diffusion Through Blogspace," in WWW 2004, New York, 2004, pp. 491 - 501.

[14] M. D. Choudhury, H. Sundaram, A. John, and D. D. Seligmann, "Can blog communication dynamics be correlated with stock market activity?," in ACM Hypertext 2008, Pittsburgh, PA, 2008, pp. 55-60.

[15] B. Krishnamurthy, P. Gill, and M. Arlitt, "A Few Chirps About Twitter," in WOSN'08, 2008, pp. 19-24.

[16] T. Kohonen, "Self-organization and associative memory," Applied Optics, vol. 24, pp. 145-147, January 15 1985.

[17] M. Mageswari and L. Goh, "Nizar is Perak MB: Details of the court ruling," in The Star Malaysia Kuala Lumpur: Star Publications (M) Bhd, 2009.

[18] P. Cashmore, "TwitterHIT: Turning Twitter into a Junk Traffic Exchange," in Mashable. 2009, A. Ostrow, Ed., 2009. Available at http://mashable.com/2009/05/16/twitterhit/.

[19] M. A. Smith, J. Ubois, and B. M. Gross, "Forward thinking," in Proceedings of the Conference on Email and Anti-Spam (CEAS 2005), 2005.

[20] C. Honeycutt and S. C. Herring, "Beyond Microblogging: Conversation and Collaboration via Twitter," in The 42nd Hawaii International Conference on System Sciences, 2009, pp. 1-10.

[21] S. Arbesman, "The Memespread Project: An Initial Analysis of the Contagious Nature of Information in Social Networks," 2004, pp. 1-9.

[22] S. Adler, "The Slashdot Effect: An Analysis of Three Internet Publications," 1999. Available at http://ssadler.phy.bnl.gov/adler/SDE/SlashDotEffect.htm.

[23] d. m. boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," Journal of Computer-Mediated Communication, vol. 13, pp. 210-230, 2007.

[24] C. Carter, G. Branston, and S. Allan, News, gender, and power: Routledge, 1998.