INTELLIGENT RETRIEVAL
Open Journal

Review

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

Marc Cheong*;Vincent Lee
Faculty of IT, Monash University Melbourne, Australia

***Corresponding author**
Marc Cheong
Faculty of IT, Monash University Melbourne, Australia

Article information
Received: November 10th, 2022; **Revised:** December 31st, 2022; **Accepted:** January 17th, 2023; **Published:** February 22nd, 2023

Cite this article
Cheong M, Lee V. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. 2023; 1(1).
doi: https://doi.org/10.70705/ppp.ir.2023.v01.i01.pp7-13

ABSTRACT

The microblogging site Twitter has a lot of room to grow into a centralized database of information useful for gathering perspectives, facts, ideas, and feelings. Activated knowledge-base decision-making with communal intelligence retrieval is the topic of this research. Because we are analyzing tweets from microblogs rather than the more typical "blogosphere," our technique is distinct from that of the current research. To further differentiate our process, we combine visualization tools with data mining methods based on artificial intelligence to categorize communications related to the trend subject. In addition to trying to correlate a Twitter trend with real-world events, our technique examines the demographics of the people who post such tweets. Through the use of visualization techniques, we are able to deduce a pattern from Twitter trends and see its 'ticking' and evolution. We get a fresh view on the people who start trends from our research, which helps us identify the 'trend setters' and their underlying traits.

Keywords
Twitter; Collective intelligence; Decision making; Knowledge base; Memetics; Microblogging; Trend analysis.

INTRODUCTION

The microblogging service Twitter is rapidly becoming popular among many people all over the world. Its reach extends from developed, urban nations like the US, where it is widely used, to developing nations in South and Southeast Asia.

Hong Kong, China, SWSM'09, November 2, 2009. This work is protected by copyright 2009 ACM. The ISBN is 978-1-60558-806-3/09/11.ten dollars.

According to its creators, "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?" sums up Twitter's mission. [1]: Yes. Users are able to compose and publish messages, or "tweets," on Twitter, with a character limit of 140.

Nevertheless, research [3, 4] has shown that Twitter users do more than just answer the question 'what am I doing?' They use it as a communication and social networking tool, giving the impression that it combines traditional blogging with an online social network. As an example, tweets may take several forms, including individuals' ordinary life updates, rebroadcasts of breaking news, and pleas for assistance during emergencies [2]. The words "tweeting" and "tweeters" have entered general use to describe both the practice of tweeting and the individuals who participate in it.

An intriguing feature of Twitter is its "Trending Topics" section, which shows the ten most popular phrases used on the platform at any one time. An intriguing glimpse into the events discussed by the Twitter community is provided by this, which is created using Twitter's proprietary algorithm. The following is a sample of a typical list of Trending Topics, arranged in decreasing order of popularity:
"New Moon, #mtvmovieawards, MTV Movie Awards, Eminem, Bruno, Ben Stiller, Twilight, Susan Boyle, #Phish, Kiefer Sutherland."
We try to figure out what makes a popular subject "tick" by breaking it down into its component parts in this study. We take a random sample of four topics that are currently trending (plus two non-trending control topics), collect data on all posts mentioning those topics up to a limit of 1500, and then analyze the data to look

for patterns in the data.

We are driven to do this work because microblogs and social networks—all born out of Web 2.0—are quickly replacing traditional means of communication and networking as the go-to means of disseminating and sharing information. People of various ages and backgrounds are jumping on the bandwagon of this new platform, which has many potential uses, including but not limited to: citizen journalism [6], staying in touch with loved ones [2], and political engagement [5]. Twitter and other online social networks and microblogging platforms provide a wealth of data that can be used for a variety of purposes, including policymaking, decision support, economic analysis, understanding epidemic behavior (the "tipping points" theorem [8]), and more.

2. RELATED WORK

Activities related to traditional blogging often include Glance et al. [10], who worked on BlogPulse to track blog trends, and Thelwall [9] who examined social network material for references to news items. Our study on microblogs (Twitter) is distinct from that of Fukuhara et al. [11], who built on [10] and conducted comparable work, but with a primary emphasis on traditional blogs. Their research has found "patterns of social concerns" and linked "blog and real world temporal data" like weather reports and news stories. By linking blog mentions of books to changes in the books' real-world sales power, Gruhl et al. were able to tap into the "predictive power of online chatter." rankings for Amazon.com [12]; it also "formalized... the notion of 'spike' topics generated by outside world events" [13] and community-generated ones. Choudhury et al. [14] found that in the financial markets, there is a connection between the discussion on technology blogs and the real stock market movement of firms associated to technology.

Blogs may be used as a "collective intelligence" tool to foretell and anticipate real-world occurrences, according to the results from [11–14] up there. Previous studies on Twitter have been carried out by Java et al. [3]. These studies attempted to categorize "user intentions" while microblogging on Twitter, as well as the geographical distribution of Twitter users and the social networks to which they belong. Further study on the identification of emergent traits of "distinct classes of Twitter users and their behaviors" and an analysis of the evolution of the Twitter user network on a region by region basis has been conducted by Krishnamurthy et al. [15], building on this work by [3]. 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," according to Mischaud [4], who analyzes Twitter users' "user intentions" from a humanities standpoint.

While previous studies have found similarities between blog posts and Twitter chatter (as a microblogging platform), we set out to investigate this phenomenon in order to shed light on how the Twitter community's chatter reflects the characteristics of real-world events portrayed as trending topics. Although Gruhl's study is confined to profiling people based on statistical analysis of the amount of posts they have authored, his research [13] has also dealt with characterizing the individuals contributing to a surge of a "spike" issue in the blogosphere.

It should be mentioned that there are a few ways in which our study varies from previous research. First, we stress the importance of microblog entries, or "tweets" on Twitter, over traditional, full-sized blogs. Furthermore, our study is based on visualization tools and data mining methodologies that are based on artificial intelligence (e.g., SOMs, developed by Kohonen [16]). In the third place, we look at the characteristics of both trending topics and the people (or "trend setters") who are a part of them by contributing "tweets" about them.

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.

Due to technological restrictions, the Twitter Search API has some limits, such as returning up to 1500 tweets as a hard upper bound and a soft limit on the date range (roughly backdated to around 20 days). The search is limited to the earliest possible date unless the subject has more than 1,500 tweets, in which case it returns the maximum number of results. A total of 7215 tweets covering the aforementioned issues have been collected; however, the control phrase 'Revolverheld', being obscure, has less than 100 tweets.

To make analyzing the raw results from Twitter Search easier, they are dumpted into a comma-separated value file. Twelve characteristics are retrieved from each result, including the date, the username, and the 140-character message. Using a "tagger" Perl script, we extract three Boolean qualities from the 140-character message content in addition to the twelve:

The inclusion of the phrase "RT" in a message allows one user to "forward" another user's tweet, similar to a "Fwd:" tag in an email. electronic mail. In an effort to understand the impact of a "retweet" on the spread of a trend, similar to the practice of email forwarding [19], we have included this behavior in our dataset.

For example, "@username this is a reply message" would be a formatted way of saying "I'm responding to you" in a tweet. Approximately 96% of all occurrences of the 'at' sign in Twitter messages are devoted to discussing contact with another person or group of users, according to research [20]. This is critical for our trend analysis because it allows us to see how tweets from one person to another shape the evolution of a hot issue. When the term "trend" appears in a message, it means that the person is trying to capitalize on the current fad without really addressing it. The message "xyz is now a trending topic!" is an example of this kind of communication. This "self-fulfilling prophecy" is happening because the sequence of communications in which it happens is helping a hot issue rise in popularity.

3.3 Demographics of Twitter users contributing to a trend

We also compile data about the users' posts in addition to the data about tweets addressing the themes mentioned above.

For each of the aforementioned subjects, we have identified the users that contributed to the thousand or so messages and eliminated any duplicates. We build a basic HTML file with the URIs of the user profiles after randomly selecting a sample from each of the individual user lists that represents around 13% of the population. A Perl script called a "sampler" automates both of the stages mentioned above.

Next, we collect information about the nation, gender, main use behavior, and Twittering device. All of the data is retrieved from the individuals' Twitter profiles (with the exception of the device used for tweeting, which can be found by comparing the data from 3.1). Please be aware that the data is presented "as is," without any kind of verification by third parties.

The data collected from 'tweets' (described in Section 3.1 above) provides information on the client and device utilized. It is possible to determine the device (computer, mobile phone, or data collected from other sources) by looking at the codename of the Twitter client program (this is not the first time this has been studied [3]).

The gender of a Twitter user may be inferred from their writing style (e.g., "username misses his/her friends") and the publicly visible profile picture. For Twitter accounts made by a company or group, there is a third gender option in addition to the traditional male and female ones.

A user's primary use pattern is determined using a survey that appears on the first page of their Twitter updates. If the poll does not provide conclusive results, the user's homepage (which may be found publicly on their profile) is visited. They may be grouped into these types:

1. Individual: most posts are about people chatting with one another or exchanging information with one another; this includes both social networking and personal messaging.

2. Group: This might be a non-profit user group for academics or a fan club for those with similar interests.

Thirdly, an aggregator is someone whose primary function is to disseminate or compile information; this may be a news agency, a Twitter account that is connected to an RSS feed, or a politician who is sent a message to his people.

4. A page that is made for the aim of comedy, satire, or parody.

Fifthly, marketing: a website designed to promote a product; unfortunately, most marketing websites are spam, unsolicited posts, and potentially dangerous.

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 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 everydayvocabulary; 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]. Howev-

er, 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 adisproportionately 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'.

Our main emphasis is on the current US drama series "Grey's Anatomy" and its short-term popularity. The red cluster accurately portrays the drama series' target audience: American women who use Twitter mostly for personal tweeting. A change in demographics is indicated by the fact that most of the Twitter users in this cluster actively contribute to Twitter not only through the web but also through other social media clients and mobile clients, which isn't immediately apparent.

Transitioning from the desktop online environment to a more mobile, social-based one, away from the typical use of online 2.0 applications like Twitter [23]. People in the blue cluster are mostly female and use Twitter for personal contact. They are located all over the world, but mostly on the European and Asian continents, and they mostly utilize the web interface to contribute to the microblog conversation. The other half of the US drama series discussion (the yellow cluster) is from aggregator users, who either seem like they're compiling entertainment/television industry news feeds or are marketing spammers "piggybacking" on a popular keyword. There is some demographic information that isn't immediately obvious in the last cluster, which is green. This group is made up of Canadians who utilize a combination of different ways to post to Twitter and who are mostly female. The media, advertising, and television production industries greatly benefit from data like the ones shown above, which is why this study is being conducted.

According to Section 4.2, which stated that the Twitter message data was collected during breakfast time in the GMT+0 time zone, the majority of users for the 'coffee' keyword are male and female from the United Kingdom and the United States. These users tweet about coffee in a 'personal' context, describing part of their daily routine. Twitter accounts associated with news aggregation and coffee-related marketing efforts make up the rest of the sample set (in red); a few of these accounts have been banned or suspended for rules violations.

5. CONCLUSION AND FUTURE WORK

In this research, we have introduced a novel method for Trend pattern analysis on the microblogging site Twitter. Using decision-making strategy based on demographics of the set of Twitter users contributing towards the debate of a certain trend, our technique used information retrieval on the collective intelligence described in the Twitter message pool and user base. This has shown promise in a variety of relevant domains, including marketing, corporate intelligence, epidemic research, and others.

To improve decision-making and trend analysis, future work in this field might use the Twitter API to access additional features and qualities that are easily accessible, such as the real-time message stream, instead of offline data. Extra empirical research is now under progress to generalize a hypothesis based on trending patterns.

REFERENCES

"Twitter" was created in 2009 by Twitter Inc.

Second, in 2009, O'Reilly Media, Inc. published The Twitter Book by T. O'Reilly and S. Milstein.

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

E. Mischaud's Master's Thesis, "Twitter: Expressions of the Whole Self," 2004. London School of Economics and Political Science (Media@LSE), 2007.

5. M. Harris, "Barack to the future," in Engineering & Technology, vol. 3, IEEE, 2008, p. 25.

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

J. Surowiecki's 2005 book, The Wisdom of Crowds, was published by Abacus in London.

[8] Malcolm Gladwell, The Tipping Point: The Power of Little Changes. Second edition, 2002, New York: Back Bay Books.

The article "No place for news in social network web sites?" was published in the Online Information Review in 2008 and can be found in volume 32, pages 726-744.

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

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

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

Referenced in [13] "Information Diffusion Through Blogspace" by D. Gruhl, D. Liben-Nowell, R. Guha, and A. Tomkins, published in 2004 at WWW 2004, New York, pages 491–501.

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

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

[16] "Self-organization and associative memory" by T. Kohonen appeared in Applied Optics, volume 24, pages 145–147, on January 15, 1985.

The Star Malaysia, Kuala Lumpur: Star Publications (M) Bhd, 2009, pp. 17–18 (written by M. Mageswari and L. Goh).

"TwitterHIT: Turning Twitter into a Junk Traffic Exchange" by P. Cashmore was published in Mashable in 2009, edited by A. Ostrow. Visit http://mashable.com/2009/05/16/twitterhit/ for more information.

Referenced in "Forward thinking" by M. A. Smith, J. Ubois, and B. M. Gross in 2005 at the Conference on Email and Anti-Spam (CEAS 2005), proceedings.

20. "Beyond Microblogging: Conversation and Collaboration via Twitter," presented by C. Honeycutt and S. C. Herring at the 2009 42nd Hawaii International Conference on System Sciences, Chapters 1–10.

As stated in [21] S. Arbesman's 2004 paper "The Memespread Project: An Initial Analysis of the Contagious Nature of Information in Social Networks," which can be found on pages 1-9.

The Slashdot Effect: An Analysis of Three Internet Publications, 1999, by S. Adler [22]. You may find it at this URL: http://ssadler.phy.bnl.gov/adler/SDE/SlashDotEffect....

The article "Social Network Sites: Definition, History, and Scholarship" was published in 2007 in the Journal of Computer-Mediated Communication and was written by d. m. boyd and N. B. Ellison.

[24] In News, Gender, and Power, edited by C. Carter, G. Branston, and S. Allan, published by Routledge in