Using Social Media
# TABLE OF CONTENTS

**On Our Radar: Social Media Technologies**  
By Bruce Lund  

**A Look Inside: Using Social Media**  
By Bruce Lund  

**The New Challenge of Data Inflation**  
By Jon Gosier  

**Structuring, Analyzing, and Leveraging Social Media Data**  
By Ioana Roxana Stanoi, Rajasekar Krishnamurthy, and Shivakumar Vaithyanathan  

**Is Social Media the Human Geography of the Emerging Human Sensor Network?**  
By Sean Gorman  

**On (Robot) Influence: Theory, Practice, and Strategic Implications**  
By Tim Hwang  

**Extracting Intelligence from Social Media**  
By Rishab Aiyer Ghosh  

**Basics of Social Media Analytics for Counterterrorism Specialists**  
By Rebecca Garcia  

**Tech Corner**  
A technology overview from IQT portfolio company Visible Technologies  

**In the News**
SOCIAL MEDIA TECHNOLOGIES

By Bruce Lund

Governments are increasingly finding that monitoring social media is an essential component in keeping track of erupting political movements, crises, epidemics, and disasters, not to mention general global trends.

Today social media is an information torrent fed by countless sources that can originate from almost anywhere on the planet. Information shared from any Internet-connected computer or text-enabled phone communicating with a hosting/distribution service, such as Twitter or Facebook, can potentially be viewed by millions of people.

Absolute statistics on the use and growth of social media become obsolete overnight. But currently, Facebook has about 800 million active users, and approximately 70 percent of those users are outside the U.S. In terms of social media-enabled mobile devices, wireless carriers activate 350,000 phones running Google’s Android operating system every day.

What is driving social media content and consumption? A key factor is the low barrier to entry. Consumers don’t need smart phones with monthly data plans or even e-mail accounts; they can post 140-character SMS messages via virtually any mobile phone made in the last decade. Where texting services have been shut down, companies such as Google have stepped in with voice-to-text-to-tweet services.

Advances in consumer technology have also been a factor. SMS mobile text messaging first became possible in the mid-1990s on a limited basis, and reached global critical mass by the middle of the last decade due to several factors, including increased mobile device support and marketing by mobile carriers. To this day, many social media services based on SMS, such as Twitter, limit their message size to the 140-byte length defined for SMS in the 1990s.

But mobile devices don’t tell the whole story. What started in the earlier part of the last decade as limited conversations between computer users participating in web-based chat rooms (e.g., to discuss music or the pros and cons of their favorite type of car or laundry detergent) has increased to span virtually any mainstream and special interest topic imaginable. Marketers, political analysts, and reporters took early note, and have been following — and sometimes participating in — these conversations ever since.

While computers still play a role in chat rooms, blogs, wikis, and tweets, it is people with mobile phones who are driving social media’s importance.

---

2 "Facebook F8: Redesigning and hitting 800 million users." LA Times, September 22, 2011.
3 "5 Technologies that Will Shape the Web." IEEE Spectrum, June 2011.
well beyond the initial application of interacting with friends (i.e., people who they actually know) or a few like-minded participants in a chat room. During this transition, a new sense of the word “friend” has entered the lexicon to accompany the growth.

While there are many personal reasons why individuals want to use social media, the ways that corporations and governments can use it are still evolving. In this issue of the IQT Quarterly, we’ll focus on what social media can offer government agencies.

The Role of Social Media Monitoring

Corporations are increasingly participating in social media conversations for reasons related to both marketing and customer service. This involvement can range from corporate-branded sites with their own chat rooms to Facebook pages to a company’s direct participation in real-time conversations on Twitter. Some of the most innovative applications from a technology perspective can be found in the social media monitoring space.

Who are the consumers and stakeholders in the social media monitoring and analysis marketplace? Other than the providers of the infrastructure used by social media (e.g., telecoms, Internet service providers, aggregators, hardware networking, storage providers, and handset makers), there are some key services that specialize in monitoring social media:

- **Listening Platforms/Brand Management Services.** Advertising agencies and consumer product companies use social media for brand management and sentiment analysis. Several listening platforms focus on providing major consumer brand companies (e.g., Microsoft, P&G) with information on views about products expressed in consumer forums, group or fan pages on consumer sites, and on Twitter. Example service providers include IQT portfolio companies NetBase and Visible Technologies.

- **News and Intelligence Services.** News organizations have monitored social media for several years for early indicators of emerging political unrest. Other services monitor social media for public health awareness. Researchers from Southeastern Louisiana University said they have been able to forecast future influenza rates with high accuracy by monitoring Twitter messages, and the U.S. Geological Survey is evaluating the use of tweets for localizing seismic events. Such applications may eventually be adopted by government agencies to track epidemics, disasters, and emerging political movements on a global scale.

What Social Media Companies Can Offer

Social media gives individuals the ability to reach communities across the globe with little latency (perhaps 30 seconds or less from origin to visibility). However, to ingest and store Twitter feeds alone requires a data storage and analysis infrastructure to service up to 100 million new tweets per day (approx. 1,000/sec.). Many of the real-time social media search companies also offer feeds from Facebook, Digg, foursquare, Flickr, and others.

Key reasons to monitor social media as close to real-time as possible include:

- Improving the analysis of potential indicators of societal change as well as informing predictive analysis.
- Understanding the source and direction of trending topics by measuring how ideas spread in social networks.

---

4 The information content of 140-byte messages in Chinese, Japanese, or Korean is inherently greater than those messages in Western languages, even when accounting for the multiple bytes required for Asian-language character encodings, since an individual character in such an Asian language can equate to an entire word in English. Thus, they’re denser.
Providing faster situational awareness and early warning when disasters such as earthquakes and hurricanes occur.

As with commercial news organizations, government agencies need to gain insight into trending topics and discussions in social media outlets. Specifically, they need to access, search, rank, and analyze large and exponentially-growing data sets from social media platforms as a part of open source research from a number of viewpoints and all-source analysis in multiple languages.

Future Considerations

We are likely to see changes in the business side of social media over the next few years.

- **Company Consolidation**: Consolidations in the social media monitoring marketplace have included the acquisition of BuzzMetrics (by Nielsen in 2006), Ellerdale (by Flipboard in 2010), Kosmix (by Wal-Mart in 2011), and Radian6 (by Salesforce.com in 2011). Not all consolidations are fruitful; News Corp bought MySpace seven years ago — at the time the world’s fastest-growing social network — for $580 million, then sold it in 2011 for $35 million. MySpace’s loss of users and advertisers coincided with the rise of Facebook over the same period.

- **Role of Large Search Providers**: Large companies such as Google, Microsoft (Bing), and Yahoo! offer access to major social media channels, as does Twitter itself.

- **Integration into the Core**: Instead of relying on a dedicated “app” or branded web-based portal to access social media services, users will increasingly interact with social media via invisible services within their devices’ operating systems. This core-level social media enablement of devices will give marketers unprecedented reach to potential audiences, allowing them to facilitate meetings with colleagues or exploit up-to-the-minute crowdsourced reviews of restaurants, among other things.

IQT has made several investments in the social media space, and we will continue to refine our social media investment strategy to focus on providing relevant social media data and technologies to our government customers. We will keep exploring the social media landscape, looking for ways these technologies can be used to effectively address government challenges.

Bruce Lund is a Senior Member of IQT’s Technical Staff within the Information and Communication Technologies Practice. Since 2001, he has participated in IQT investments in language processing, search, and video analytics. Prior to joining IQT, Bruce worked on Human Language Technology-related projects at the National Institute of Standards and Technology’s Information Technology Laboratory. He has a Ph.D. in computational linguistics from Georgetown University.
In this edition of the *IQT Quarterly*, we explore the use of social media from the unique perspective of In-Q-Tel and our government customers. Social media is a rapidly evolving field, and this issue focuses on how companies and researchers are innovating in the realm of social media, especially in its analysis.

In the first article, Jon Gosier from metaLayer considers big data as it relates to social media communication. He discusses the massive amounts of content created through social networks and the additional data that originates from context about content and “sharing.” He explains four different methodologies for how people deal with social data at an immense and expanding scale, in the hopes of making data from online social content actionable.

Next, Ioana Roxana Stanoi, Rajasekar Krishnamurthy, and Shivakumar Vaithyanathan from IBM Research observe that information from social media can be represented as entities such as people, organizations, and locations. Such entities enable the construction of profiles and relationships that can be better understood and analyzed. The authors go on to describe how they create this entity-centric view and how it can be applied.

In his article, Sean Gorman cites the emerging streams of mobile and social data as sources that can provide new insights into geographical analysis. Place awareness provides context, which itself is valuable in substantiating an author’s authenticity. However, location information associated with the new streams of data can be spoofed, so caution is in order. Gorman puts some of these concepts to test at GeoIQ with two case studies: using Twitter to respond to Hurricane Irene, and leveraging social data from a mobile application for crowd monitoring at the New York City Marathon.

Tim Hwang’s article follows with a discussion of the influence of software agents (bots, robo-sellers, etc.) in social media. These programs, which typically are not designed to be noticed, can have unforeseen influences on online marketplaces, news, and even stock trading. As they are not constrained to only social media, Hwang believes that bots may eventually be used to disrupt the fidelity of open source intelligence.

Next we consider Rishab Aiyer Ghosh’s experience testing social media data, such as the volume of particular keywords and Twitter hashtags, against actual events during the 2011 Arab Spring using analytics provided by Topsy Labs. He offers three analyses: the overall Twitter activity by Mideast country name hashtags, a view of Tunisia-related Twitter activity mapped to a chronology of real-world events, and finally, a type of back-test in which he asks if protests organized on specific dates could have been anticipated.

Rebecca Garcia’s premise in the final article is that many of the techniques used to distill the essence of online conversations for customer service can also be used to analyze conversations among terrorist cells. She suggests that this application of social media analytics could emerge as an indispensable counterterrorism tool by enabling proactive monitoring, analysis, and engagement through extremist social networks and their associated digital properties. She also provides some practical applications of social media analytics based on her experience at SAS Federal.

The articles in this edition of the *IQT Quarterly* address some of today’s leading social media analysis topics:

- Appreciating the importance of the context of a social media message (including the message’s location of origin).
- Acquiring volumes of social media data to substantiate statistical claims and provide insights.
- Building the infrastructure to provide near real-time analytics and to scale to meet the ever-growing social media data supply.

There are many more aspects to using social media than we can address here, of course, but we hope this issue will serve as a starting point for discussion around this important technology area.
THE NEW CHALLENGE OF DATA INFLATION

By Jon Gosier

The online chatter of individuals, networks of friends or professionals, and the data being created by networked devices are growing exponentially, although our time to consume it all remains depressingly finite. There is a plethora of solutions that approach the challenge of data inflation in different ways. But what methodologies have worked at scale and how are they being used in the field?

The Infinite Age

First, let’s look at the landscape. The amount of content being created daily is incredible, and includes the growth of image data, videos, networked devices, and the contextual metadata created around these components.

The Information Age is behind us; we’re now living in an Infinite Age, where the flow of data is ever-growing and ever-changing. The life of this content may be ephemeral (social media) or mostly constant (static documents and files). In either case, the challenge is dealing with it at human scale.

In the time it takes to finish reading a given blog post, chances are high that it’s been retweeted, cited, aggregated, pushed, commented on, or ‘liked’ all over the web. An all-encompassing verb to describe this is sharing. All of this sharing confuses our ability to define the essence of that original content because value has been added at different points by people who’ve consumed it.

Does that additional activity become a part of the original content, or does it exist as something else entirely? Is the content the original blog post, or is it that post coupled with the likes, retweets, comments, and so on?

Content is now born and almost instantaneously multiplies, not just through copying, but also through the interactions individuals have with it — making it their own and subsequently augmenting it.

In the Infinite Age, one person may have created the content, but because others consumed it, it’s as much a representation of the reader as it is the author.

The Inflation Epoch

In physics, the moment just after the Big Bang, but prior to the early formation of the universe as we understand it, is called the Planck Epoch or the GUT Era (for the Grand Unified Theory that explains how all the four forces of nature were once unified). The Planck Epoch is defined as the moment of accelerated growth when the universe expanded from a singularity smaller than a proton.

We mirror this moment a trillion times each day when online content is created and augmented by the
surrounding activity from others online. From the moment we publish, the dogpile of interaction that follows is a bit like the entropy that followed the Big Bang: exponential and rapid before gradually trailing off to a semi-static state where people may still interact, but usually with less frequency, before leaving to consume new content being published elsewhere.

When people talk about big data, this is one of the problems they are discussing. The amount of content that is created daily can be measured in the billions (articles, blog posts, tweets, text messages, photos, etc.), but that’s only where the problem starts. There’s also the activity that surrounds each of those items — comment threads, photo memes, likes on Facebook, and retweets on Twitter. Outside of this social communication, there are other problems that must be solved, but for the sake of this article we’ll limit our focus to big data as it relates to making social media communication actionable.

How are people dealing with social data at this scale to make it actionable without becoming overwhelmed?

Methodology 1 – Index and Search

Search technology as a method for sorting through incredibly large data sets is the one that most people understand because they use it regularly in the form of Google. Keeping track of all the URLs and links that make up the web is a seemingly Sisyphean task, but Google goes a step beyond that by crawling and indexing large portions of public content online. This helps Google perform searches much faster than having to crawl the entire web every time a search is executed. Instead, connections are formed at the database level, allowing queries to occur faster while using less server resources. Companies like Google have massive data centers that enable these indexes and queries to take seconds.

As big data becomes a growing problem inside of organizations as much as outside, index and search technology has become one way to deal with data. A new swath of technologies allow organizations to accomplish this without the same level of infrastructure because, understandably, not every company can afford to spend like Google does on its data challenges. Companies like Vertica, Cloudera, 10Gen, and others provide database technology that can be deployed internally or across cloud servers (think Amazon Web Services), which makes dealing with inflated content easier by structuring at the database level so that retrieving information takes fewer computing resources.

This approach allows organizations to capture enormous quantities of data in a database so that it can be retrieved and made actionable later.

Methodology 2 – Contextualization and Feature Extraction

Through the development of search technologies, the phrase “feature extraction” became common terminology in information retrieval circles. Feature extraction uses algorithms to pull out the individual nuances of content. This is done at what I call an atomic level, meaning any characteristic of data that can be quantified. For instance, in an email, the TO: and FROM: addresses would be features of that email. The timestamp indicating when that email was sent is also a feature. The subject would be another. Within each of those there are more features as well, but this is the general high-level concept.

Stacked Waveform Graph used to plot data with values in a positive and negative domain (like sentiment analysis). Produced by the author using metaLayer.
For the most part, search uses such features to help improve the efficiency of the index. In contextualization technologies, the practice is to use these features to modify the initial content, adding it as metadata and thereby creating a type of inflated content (as we’ve augmented the original).

When users drag photos onto a map on Flickr, they are contextualizing them by tagging the files with location data. This location data makes the inflated content actionable; we can view it on a map, or we can segment photos by where they were taken. This new location data is an augmentation of the original metadata, and creates something that previously did not exist.

When we’re dealing in the hundreds of thousands, millions, and billions of content items, feature extraction is used to carry out the previous example at scale.

The following is a real-world use case, though I’ve been careful not to divulge any of the client’s proprietary details.

Recently at my company metaLayer, a colleague came to us with one terabyte of messages pulled from Twitter. These messages had been posted during Hurricane Irene. The client needed to generate conclusions about these messages that were not going to be possible in their original loosely structured form. He asked us to help him structure this data, find the items he was looking for (messages about Hurricane Irene or people affected by it), and extract the features that would be useful for him.

The client lacked the sufficient context to identify what was most relevant in the data set. So we used our platform to algorithmically extract features like sentiment, location, and taxonomic values from each Twitter message using natural language processing. Because it was an algorithmic process, this only took a few hours, allowing the client to get a baseline that made the rest of his research possible. Now the data could be visualized or segmented in ways that weren’t possible with the initial content. The inflated content — metadata generated algorithmically — included elements that made his research possible. We now had an individual profile of every message that gave us a clue about its tone, location of origin, and how to categorize the message. This allowed the client’s team to look at the data with a new level of confidence.

In the context of our social data challenge, these extracted features might be used on their own by an application, or they might become part of an index or database like the one mentioned previously.

**Methodology 3 – Visualization**

Visualization is another way to deal with excessive data problems. Visualization may sound like an abstract term, but the visual domain is actually one of the most basic things humans can use for relating complex concepts to one another. We’ve used symbols for thousands of years to do this. Data visualizations and infographics simply use symbols to cross barriers of understanding about the complexities of research so that, regardless of expertise, everyone involved has a better understanding of a problem.

Content producers like the New York Times have found that visualizations are a great way to increase audience engagement with news content. The explosion of interest in infographics and data visualizations online echoes this.

To visualize excessive data sets, leveraging some of the previous methods makes discovering hidden patterns and trends a visual, and likely more intuitive, process.

**Methodology 4 – Crowd Filtering and Curation**

The rise of crowdsourcing methodologies presents a new framework for dealing with certain types of information overload. Websites like Digg and Reddit are examples of using crowdsourcing to vet and prioritize data by the cumulative interests of a given community. On these websites, anyone can contribute information, but only the material deemed to be
Big data can refer to managing an excess of data, to an overwhelming feed of data, or to the rapid proliferation of inflated content due to the meta-values added through sharing.

interesting by the highest number of people will rise to the top. By leveraging the community of users and consumers itself as a resource, the admin passes the responsibility of finding the most relevant content to the users.

This, of course, won’t work in quite the same way for an organization’s internal project or your data mining project, but it does work if you want to limit the information collected to those in the crowd who have some measure of authority or authenticity.

The news curation service Storyful.com is a great example of using crowd-sourced information to report on breaking events around the globe. Its system works not because the masses are telling the stories (that would lead to unmanageable chaos), but because the staff behind Storyful has pre-vetted contributors. This is known as bounded crowdsourcing, which simply means to extend some measure of authority or preference to a subset of a larger group. In this case, the larger group is anyone using social media around the world, whereas the bounded group is only those around the world that Storyful’s staff has deemed to be consistent in their reliability, authority, and authenticity. This is commonly referred to as curating the curators.

Curation has risen as a term over the past decade to refer to the collection, intermixing, and re-syndication of inflated content. It is used to refer to the construction of narratives without actually producing any original content, instead taking relevant bits of content created by others and using them as the building blocks for something new. This presentation of public data as edited by others represents a new work, though this “new work” may be made up of nothing original at all. By carefully selecting curators whom you know will be selective in what they curate, the aggregate of information produced should be of a higher quality.

**Conclusion**

Big data, like most tech catchphrases, means different things to different people. It can refer to managing an excess of data, to an overwhelming feed of data, or to the rapid proliferation of inflated content due to the meta-values added through sharing. Pulling actionable information from streams of social communication represents a unique challenge in that it embodies all aspects of the phrase and the accompanying challenges. Ultimately, if content is growing exponentially, the methods to manage it have to be capable of equal speed and scale.

*Jon Gosier* is the co-founder of metaLayer Inc., which offers products for visualization, analytics, and the structure of indiscriminate data sets. metaLayer gives companies access to complex visualization and algorithmic tools, making intelligence solutions intuitive and more affordable. Jon is a former staff member at Ushahidi, an organization that provides open source data products for global disaster response and has worked on signal to noise problems with international journalists and defense organizations as Director of its SwiftRiver team.
STRUCTURING, ANALYZING, AND LEVERAGING SOCIAL MEDIA DATA

By Ioana Roxana Stanoi, Rajasekar Krishnamurthy, and Shivakumar Vaithyanathan

Social media is an interactive vehicle for communication, accessed on a daily basis by hundreds of millions of people. Unlike conventional media that is a one-way street for information exchange, social media enables people to write content as well as provide feedback and recommend content to other users. Sources of social media vary in purpose and data formats.

On blogs, small groups of users comment on specific topics, while forum users discuss or comment on a broader range of topics. Hundreds of millions of users form communities on social networking sites like Facebook and Twitter. Furthermore, professional networking sites like LinkedIn attract over 100 million users who exchange information about professional experience, interests, and events. By analyzing and integrating insights from communication across these sources, one can explore different scenarios and answer questions as varied as:

- Which people or organizations have had their accounts or systems hacked in the past week?
- Are there trends in these hacking attacks? Who is affected by this account and what has been the impact of the attack?
- Given a set of “persons of interest,” can we identify the people closely related to them? Do personal and professional relationships and common affiliations surface new connections between two or more of these members?
- Who are the people planning to attend a public demonstration in a given city? What advanced indicators, if any, are available on the relative magnitude of these demonstrations and the mood of the demonstrators? What is the ratio of
people who agree with the protesters to people who oppose the movement, and what are their most common arguments?

- Who are the individuals that intend to buy a particular product? Are there any consumer segments that are more likely to respond favorably to a particular marketing campaign?

- Are there any incidents of food poisoning or other health-related issues linked to a particular product or location? Can one identify outbreak patterns based on the demographics and recent activities of the affected individuals [e.g., travel, attendance at a public event or concert]?

Note that the answers to the previous questions require deep understanding of the individuals, their relationships and opinions, along with micro-segment or group-level aggregation analysis.

Benefits of Leveraging Social Media

Unique to social media is the wealth of content that includes timely, detailed, and continuous information about hundreds of millions of entities such as people, events, and organizations. Opinions are persisted as soon as they are expressed; events are documented through words, images, and video, often as soon as they happen. Tracing and connecting these dots of information throughout their development becomes a question of extracting, structuring, and integrating relevant data from text. For instance, a person’s profile accumulated over time can expose detailed personal experiences, interests and hobbies, travel patterns, intents, and, very importantly, relationships and affiliations with other people and organizations.

At IBM Research, we currently identify over 35 types of relationships between people, ranging from acquaintance and friend to various family and business relationships.

Facts about people can be understood within the context represented not just by their identifiable information such as name, location, and date of birth, but also in the context of different groups they are associated with. One obvious way of grouping people is using explicit connections such as family ties, relationships, or affiliations with institutions. However, a population can also be segmented by individual attributes or a combination of attributes such as location, age, items purchased, participation in events, hobbies, occupation, and life events.

Representing the information from social media as entities [people, organization, location, etc.] enables the building, over time, of comprehensive profiles and relationships, which can then be used by multiple applications in healthcare, national security, marketing, and competitive intelligence. The following describes how to create an entity-centric view from social media data, and two applications that leverage this view.

Building Individual Profiles and Micro-Segmentation

The figure on the following page describes the various categories of information that can be obtained about individuals from social media data. As mentioned, building and maintaining these profiles is the foundation of a platform that can answer the various questions mentioned earlier in the article. The profile information includes personal attributes such as
identifying information and interests, life events such as getting married or changing a job, relationship information (both personal and business), behavioral information such as purchase intentions, browsing and interaction history, and timely insights such as current travel plans and event participation information.

The sequence of steps in our process required to construct individual personal profiles from social media is described below.

1. Extraction. Each social media post is analyzed and relevant profile values are surfaced through high-quality information extractors. For instance, the message "Congrats @jdoe and @mbeth on your second kid" reveals a relationship between two individuals and also the fact that they recently had their second child. Similarly, the message "My bank account got hacked" identifies a recent victim of fraudulent activity. The accuracy of information extraction is critical for subsequent individual and group-level analysis and any conclusions drawn thereof. Furthermore, mentions of related entities such as products, brands, and organizations need be cleansed, normalized, and mapped to standard categories (e.g., "I like bb bold" refers to BlackBerry Bold).

2. Source-Level Profile Integration. The output of the extraction is either integrated into an existing profile or used to start a new profile. This step requires understanding how the new data relates to what is already known about that particular profile, resolving any conflicts that arise during the process, and establishing trust on profile values at a very granular level. For instance, the message "Moved to LA this week love the new job" implies that the current location of the individual has changed to Los Angeles.

3. Linking Profiles Across Sources. Individuals typically have a social media presence on multiple sites. Resolving these multiple identities into a single real-world entity is an important operation. The resolution process does need to account for the fact that identity information may be noisy and sparse. At the same time, it needs to exploit the fact that people leave links to their other social profiles encouraging...
friends and followers to read their views on multiple platforms. For instance, a personal blog may contain a reference to a Twitter profile or a Facebook page (e.g., “follow me on twitter @jdoe”).

4. Summarization. Data stored in profiles should be unambiguous and the footprint of the stored data should be minimized without losing any useful information. Redundant, obsolete, and unnecessary data should be removed and values collapsed to the appropriate level of granularity. For instance, knowing the exact location of an individual at various points in time (through analysis of geolocation data in social media messages) may be critical for certain applications, while in others, it may be sufficient to maintain a summarized temporal location history.

Constructed profiles can be used in multiple ways. They can be aggregated across different dimensions such as affiliations, interests, and location. In some cases, more complex statistical analysis can be used to detect trends and even predict outcomes. There may also be, for specific applications, the need to enhance individual profiles with application-specific information. Below we describe two applications in the context of some of the questions posed at the beginning of this article.

**Hacking Attacks.** On September 9, 2011, the Twitter account of NBC News was hacked and breaking news was posted on its Twitter account about an airplane that crashed in New York at the site of Ground Zero. When journalistic accounts are hacked, fake news can quickly be disseminated through social networks. In the case of the NBC News account, there was an almost instant increase in social media buzz related to hacking incidents. Early detection of such an increase could enable proactive damage control. Furthermore, information directly related to hacking incidents of journalistic accounts, when combined with demographic data (e.g., geographical location of journalists and opinions of news organizations), sheds light on which accounts are more likely to be future targets.

Performing the above analysis requires identifying when hacking incidents are reported in social media messages, the nature of the accounts hacked (e.g., personal versus business, financial versus email), and even identifying the industry category of the business accounts (e.g., journalist, blogger). Social influence is an indicator of reachability and therefore the potential damage that can result from a hacking attack. Social influence can be calculated based on various criteria including the number of followers and number of times social media messages from that account have been reposted. More complex statistical analysis can expose general hacking trends, calculate the risk that a site will be targeted, and predict the effects of such an incident.

**Disease Outbreaks.** The Centers for Disease Control and Prevention (CDC) track occurrences of different diseases and conditions, mostly based on reports from medical offices and hospitals. For instance, influenza information is summarized and made available by the CDC on a weekly basis, while data on other health conditions such as food poisoning is available less frequently, sometimes only on a yearly basis. By contrast, social media can be used to accumulate and analyze mentions of diseases, medical conditions, and
symptoms as they happen. Detailed information in social media profiles of the affected individuals allows for understanding causes and trends, and predicting risks and outcomes.

To perform this analysis, mentions of occurrences of diseases can be extracted from social media posts. These mentions can cover a variety of health conditions such as flu, food poisoning, diabetes, asthma, and allergies. When combined with profile data such as location, travel patterns, places frequented, and food habits, a deeper understanding of the progress and causes of diseases and conditions becomes possible. For example, detailed information on food poisoning incidents not only enables one to zoom into the geographical location where there may be an outbreak, but also to study travel patterns and habits of affected people. In addition, simple aggregations can expose clusters of outbreaks, while more complex predictive models can be used to further calculate causes and risks, enabling proactive measures to be taken.

**Conclusion**

Data generated by social media contains valuable information that can be used to build comprehensive profiles of people, events, and organizations. These profiles, together with additional application-specific information, can help surface previously unavailable insights into public health, national security, marketing, and competitive intelligence. In this article, we discussed how applications can leverage the volume and timeliness of information in social media. As social media continues to expand and its content becomes richer, it will undoubtedly become an indispensable tool for many applications.

---

**Ioana Roxana Stanoi** is a researcher at IBM Research — Almaden. Her current work focuses on scalable entity extraction and integration, applied to Social Media Analytics. Her experience also includes event-driven processing, mobile e-commerce, query processing, index optimization, and transaction support. Ioana received her Ph.D. in Computer Science and a B.A. in Physics from the University of California Santa Barbara.

**Rajasekar Krishnamurthy** manages the Intelligent Information Integration group at IBM Research — Almaden. His research interests are primarily in large-scale unstructured data management, with emphasis on text analytics and entity integration technologies. He holds a B.Tech in Computer Science and Engineering from the Indian Institute of Technology Madras, and a Ph.D. in Computer Science from the University of Wisconsin–Madison.

**Shivakumar Vaithyanathan** is the Department Manager of the Intelligent Information Systems at IBM Research — Almaden. His department is involved in building systems for Scalable Text Analytics and Entity Integration, Enterprise Search, and large-scale machine learning and mining. Shivakumar is an Associate Editor of the Journal of Statistical Analysis and Data Mining.
Is Social Media the Human Geography of the Emerging Human Sensor Network?

By Sean Gorman

Of the 318 million mobile handsets shipped in 2011, 79.9 percent will be GPS-enabled. By 2014, 18 percent of laptops and 42 percent of portable video games will be GPS-enabled (iSuppli 2010). In Kenya, the hottest selling new phone is Huawei’s $80 GPS-enabled Android phone, which has already sold 350,000+ units (MIT 2011). In the United States, 72.2 percent of users accessed social media sites and blogs through their mobile devices — up 37 percent from 2010 (ComScore 2011). Over 40 percent of content on Twitter came from mobile devices in 2011 (GigaOm 2011).

While these stats may seem disconnected, they are indicative of a shift happening in how the world interacts with digital information. We are rapidly converging to a point, in both the developed and developing worlds, where our daily lives will be streamed through mobile devices that attach a location and time to our collective actions. The exhaust of data being generated by mobile devices is potentially the richest collection of human geography/terrain data ever generated. While the potential is immense, there are significant challenges facing the use of social media and mobile content for understanding human geography.

This shift in social and mobile interaction provides us with the potential to address some of the Intelligence Community’s top priorities. One of the top priorities expressed by Letitia Long, Director of the National Geospatial-Intelligence Agency, is:

“Create new value by broadening and deepening our analytic expertise. By providing deeper, contextual analysis of places informed not only by the earth’s physical features and imagery intelligence, but also by ‘human geography.’”

Harnessing the emerging streams of mobile/social data as new insight to human geography is exciting, but not without criticism and shortcomings. We will look at the criticisms, challenges, and potential for leveraging social media for a better understanding of human geography through use cases and current research. Specifically, we will consider how social media was used to aid in response to Hurricane Irene, and also to coordinate with FEMA for crowd tracking during the New York City Marathon.

Criticism of Social Media as a Data Source

While the potential for leveraging social media is great, it is helpful to start by addressing the criticisms of using it as a data source. It is easy and partially accurate to view social media as simply a collection of pre-teens and bots1 discussing Justin Bieber. While this evaluation is somewhat hyperbolic, it exposes four fundamental criticisms of social media as a viable source of information:

1. Social media content lacks substance.
2. Social media content lacks authority.
3. Social media content is biased.
4. Social media content can be spoofed.

---

1 “Bots” are software applications that run automated tasks over the Internet – in the case of social media, often impersonating humans to create more content on a topic, event, or brand.
There is truth in all these criticisms. Much of the content created in social media applications consists of vapid topics by users who have little experience or authority to speak on the topic. The users of social media, and specifically those who create content, are not a representative sample of the population. Bots that generate automated content are a common part of many social media applications, and users can be spoofed and accounts hacked. That said, these four criticisms are only part of the story, and methodical approaches and use of the appropriate tools can keep shortcomings in balance when it comes to performing useful analysis. They can’t, however, be ignored if you want to leverage social media for accurate answers.

**Addressing Substance and Authority**

One approach to addressing these shortcomings is leveraging the location component of social media content. As location-enabled mobile devices become more pervasive, location information enables new tool sets, which provide value in multiple areas. By knowing the location of a user who is contributing data to a social media source, we have a much better idea of his or her ability to create content of substance and authority. A user posting on the progress of the Arab Spring who is located in Tahrir Square in Cairo will have a level of authority inferred purely by being at the location of the event. This also enables the user to provide substance through text, photos, or video.

While we can use location to filter social media data to identify more substantive and authoritative data, there is still a challenge in the data being biased. This poses two problems. Relying on only a few pieces of social media information could create a false positive — the amount of data is not statistically significant. The data could also only reflect one segment of the population — for instance, young affluent males. Even if you do have a large amount of data, it could fail to reflect the totality of society in a location.

**Quantifying Bias and Combating Spoofing**

To address these problems we can again leverage location. For a geographic region, we can find the number of data points observed and calculate the margin of error based on the total population there. We see this with news media polls all the time. For example, based on a sample of 1,000 people we have a +/- 5 percent margin of error. If we wanted to be 95 percent confident with a margin of error of +/- 4 percent in a geography with a population of 100,000, we’d need to observe at least 451 social media data points.

Further, the social media data sample could be correlated with demographics of the region to see what strata of society are not being represented. For instance, assume the 18-25 year-old population represents 80 percent of the geographic variation in the data, but the 60-67 year-old population only explains 10 percent. Now, we know that the data sample is not providing a meaningful reflection of the 60-67 year-old population, and we need to use other techniques to discover data on that population. In many ways, this bias can be helpful because it gives further insight into the detailed human geography of emerging data sources. It also gives us a method to interconnect the data with more traditional data sources. Location is the common thread that ties together both new and old data sources, so we have appropriate context for informed analysis.

Location from mobile applications also gives us another tool to combat spoofing. On Twitter, 24 percent of tweets in 2009 were created by bots, (Sysomos 2009). While many social media networks actively work to exclude bots from generating content, it is still a pervasive problem. This also causes problems for analysis since it makes it relatively simple to spoof information into an application. While creating a bot, which automates content to a social network is simple, creating a bot that mimics a GPS-enabled mobile phone is much more difficult. The GPS coordinates attached to social media created by a mobile device become a helpful tool for identifying human-generated content.

If a human is spoofing information from a mobile device to get around the challenges of a bot creating believable data, there is still a problem of scale. In order to spoof data you must achieve a scale that will outweigh the network of real people generating information on events. For example, a group of individuals trying to spoof false information about a protest would need enough humans to reach a scale that drowned out the real data about the protest coming from mobile devices. This makes it very hard to prevent a sample of sufficient size not to converge on truth.
Putting Concepts to the Test

While discussing the potential of social media to provide meaningful human geography is a useful exercise, it is more valuable to put social media to the test with a real-world event. To give the opportunity a proper scope we’ll cover two scenarios: 1) using Twitter to respond to Hurricane Irene and 2) leveraging social data from an NYC Marathon mobile application for crowd monitoring.

Hurricane Irene

As Hurricane Irene barreled towards the Eastern seaboard in August 2011, the disaster response community was interested in leveraging social media to identify where citizens needed help. To accomplish the task, they tapped into the live stream of data coming from Twitter, and filtered it to capture just those tweets that talked about the hurricane and flooding. Mapping those tweets with geographic information, they were able to analyze the data to discover areas where the public said flooding was occurring. The visualization in Figure 1 shows those tweets overlaid with the storm surge risk data, then aggregated to identify which locations were at greatest risk based on the fusion of hazard and social media data.

NYC Marathon

To promote the 2011 New York City Marathon, organizers had a mobile application built that allowed spectators to interact with the event through a variety of social features. Both the event organizers and government response agencies like FEMA were interested in being able to monitor crowd interaction with the event. The application was used on 31,686 unique devices and they collectively generated 3.6 million events with both a location and time stamp. This provided a large sample of data on the location of crowds as well as characteristics about their actions. It provided a real-time sensor network and dynamically updating situational awareness of the event’s crowd activity across the city. Both the organizers and government support agencies were interested in understanding when crowds in a geographic location grew quickly. To fill this need, real-time analytics were constructed to aggregate the crowd activity by block groups across New York City, and update that analysis every ten minutes. This
analysis was then linked to a visualization that updated the crowd counts by block group as a map. The result can be seen in Figure 2.

This type of approach to dynamic data opens up the potential to also have calculations run to detect when the crowd in any one block increases rapidly. For example, when a crowd in a block group increases by more than 100 percent or exceeds 2,000 people, an SMS alert could be sent to police within a 500 meter radius of the location.

The use of social media for Hurricane Irene response and the NYC Marathon demonstrates where social/mobile data opens up entirely new possibilities for the analysis of human geography. Since the data is dynamic, analysis can also be dynamic, enabling change detection to proactively identify hot spots and areas of interest. Not only do these emergent data sources give us new insight into human geography, they also open the door to new analytics that could provide immense value over traditional approaches. There are still challenges with leveraging new data streams from social/mobile content, but with awareness and solid methodological approaches to address issues, there is immense potential for the entire analytic community.

Sean Gorman is the founder of IQT portfolio company GeoIQ, a collaborative web platform for geographic data analysis. Previously he was in academia as a research professor at George Mason University and before that a graduate student at the University of Florida. In between he worked for startups in the DC area building online communities, providing Geo-IP location, and mapping telecom infrastructure. His research has been featured in Wired, Der Spiegel, ABC News, Fox News, The Washington Post, Business 2.0, and CNN.

REFERENCES

Sysomos – http://www.sysomos.com/insidetwitter/mostactiveusers/
ON (ROBOT) INFLUENCE: THEORY, PRACTICE, AND STRATEGIC IMPLICATIONS

By Tim Hwang

Three Stories

In March 2011, The Huffington Post reported on a heretofore unnoticed correlation: whenever the actress Anne Hathaway appears in the news, the stock price of Warren Buffett’s Berkshire Hathaway always seems to take a jump upwards.1 While not ultimately conclusive, further digging seemed to suggest that this otherwise inexplicable pattern might have been the result of high-frequency algorithmic trading programs playing the stock market according to detections of the word “Hathaway” in the news.2

A month later, in April, a professor at UC Berkeley reported on an odd finding: the presence of two third-party sellers on Amazon.com selling The Making of a Fly — a classic of developmental biology — for the astronomical prices of $23,698,655 and $18,651,718, respectively.3 Closer observation revealed that these two sellers were not real sellers at all. Rather, they appeared to be computer programs implemented by wholesalers to post and price books based on the behavior of other sellers in the marketplace. These systems usually price books competitively and reasonably. However, the collision of two of these programs in this case produced an endless escalation as one robo-seller would raise the price on its copy of the book, and the other would respond by raising its own price, and so on. Without any reasonable cap on the pricing of the product, the two bot sellers continued this escalation until the prices reached the stratospheric levels at which they were eventually observed.

In October, a columnist for the New York Times discovered a book for sale online curiously named and focused on the topic of “Saltine Cracker.”4 Searching for other titles by the author of this volume, Lambert M. Surhone, showed that he was the writer or editor of

---

more than 100,000 titles on Amazon. While calls to the publishing company were answered with assurances that the books were produced by human writers and editors, closer investigation revealed numerous obvious and “robot-like” errors. For instance, one book on English rock band The Police was decorated by a cover design featuring actual police officers.

What can we draw from these stories? These vignettes are rightly classified as types of revealing errors — failures that show us the usually invisible ways in which technology affects our lives. It is important to recognize that none of the previously discussed systems were ever designed to be noticed. Indeed, algorithmic systems, like those that make trades on the stock market, often go entirely undetected by the human users that interact with them. We are only able to glimpse the hidden influence of these systems when glitches in their design make their presence obvious.

One important lesson is that digitization enables botification. That is to say, the movement of a certain arena of human activity onto a digital platform allows software to act in lieu of actual people for that activity. Amazon, for instance, moved book sales and retail online. In turn, book wholesalers were able to automate the pricing and posting of books.

It is important to note that these automated identities acting as people — “bots” — can have a massive effect on the system as a whole. Indeed, just how significant these bots can be was on full display on May 6, 2010. For a heart-stopping few minutes, the Dow Jones Industrial Average dropped a staggering 1000 points, and then, absurdly, proceeded to regain all those losses within the subsequent few minutes. Later investigation by the SEC suggested that subtle, accidental interactions between numerous automated high-frequency stock trading programs were largely responsible for the failure. The sudden “Flash Crash” was itself a kind of revealing error, and it underscores the power of these bots to shape the human systems that they operate within.

**The Rise of the Socialbots**

Digitization enables botification. The examples discussed previously certainly seem to show that this holds true for online marketplaces. But the Internet obviously contains much more than that. Most notably, the massive migration of community activity onto platforms like Facebook and Twitter accomplishes a social “digitization” that platforms like Amazon and eBay accomplished for commerce. This obviously begs the question: could bots act in lieu of humans on these platforms in the way that they do in online marketplaces? And, if so, could these bots similarly influence the behavior of human actors in scaled and imperceptible ways?

Our team at the Pacific Social Architecting Corporation set out with this question in mind. The initial project was to hold an event that invited teams to design their own “socialbots” to compete on their ability to provoke certain behaviors in a social network. This competition took place in early 2011 on Twitter. Three teams were given three weeks in which to review a network of 500 human users and program a bot to interact with these targets. Once the bots were launched, the competition itself lasted for two weeks in which we monitored the success (or failure) of the bots in interacting with these unsuspecting individuals.

Each team received a single point for every mutual friend relationship created with a user in the target network. Teams also received three points for every message sent to the bot from a human user in the more than 100,000 titles on Amazon. While calls to the publishing company were answered with assurances that the books were produced by human writers and editors, closer investigation revealed numerous obvious and “robot-like” errors. For instance, one book on English rock band The Police was decorated by a cover design featuring actual police officers.

What can we draw from these stories? These vignettes are rightly classified as types of revealing errors — failures that show us the usually invisible ways in which technology affects our lives. It is important to recognize that none of the previously discussed systems were ever designed to be noticed. Indeed, algorithmic systems, like those that make trades on the stock market, often go entirely undetected by the human users that interact with them. We are only able to glimpse the hidden influence of these systems when glitches in their design make their presence obvious.

One important lesson is that digitization enables botification. That is to say, the movement of a certain arena of human activity onto a digital platform allows software to act in lieu of actual people for that activity. Amazon, for instance, moved book sales and retail online. In turn, book wholesalers were able to automate the pricing and posting of books.

It is important to note that these automated identities acting as people — “bots” — can have a massive effect on the system as a whole. Indeed, just how significant these bots can be was on full display on May 6, 2010. For a heart-stopping few minutes, the Dow Jones Industrial Average dropped a staggering 1000 points, and then, absurdly, proceeded to regain all those losses within the subsequent few minutes. Later investigation by the SEC suggested that subtle, accidental interactions between numerous automated high-frequency stock trading programs were largely responsible for the failure. The sudden “Flash Crash” was itself a kind of revealing error, and it underscores the power of these bots to shape the human systems that they operate within.

**The Rise of the Socialbots**

Digitization enables botification. The examples discussed previously certainly seem to show that this holds true for online marketplaces. But the Internet obviously contains much more than that. Most notably, the massive migration of community activity onto platforms like Facebook and Twitter accomplishes a social “digitization” that platforms like Amazon and eBay accomplished for commerce. This obviously begs the question: could bots act in lieu of humans on these platforms in the way that they do in online marketplaces? And, if so, could these bots similarly influence the behavior of human actors in scaled and imperceptible ways?

Our team at the Pacific Social Architecting Corporation set out with this question in mind. The initial project was to hold an event that invited teams to design their own “socialbots” to compete on their ability to provoke certain behaviors in a social network. This competition took place in early 2011 on Twitter. Three teams were given three weeks in which to review a network of 500 human users and program a bot to interact with these targets. Once the bots were launched, the competition itself lasted for two weeks in which we monitored the success (or failure) of the bots in interacting with these unsuspecting individuals.

Each team received a single point for every mutual friend relationship created with a user in the target network. Teams also received three points for every message sent to the bot from a human user in the

---

5 On this particular concept I’m indebted to a project by Benjamin Mako Hill, which tracks various revealing errors in technology and elsewhere. See Revealing Errors, http://revealingerrors.com/ (last visited Nov. 6, 2011).


9 In the parlance of Twitter, the team would receive a point if their bot “followed” a user, and that user in turn followed back.
target network. Finally, teams were constrained in the types of bots that could be designed by a longer set of rules intended to minimize the possibility of harm on the target group and ensure that there was a fair test of the ability of these automated identities to create changes in the social network.  

To give a sense of how these bots actually worked in practice, the winning socialbot model operated largely through a database of generic questions and responses. These included general questions about various topics, as well as general purpose responses like, “That’s very interesting.” It was designed to message users in the target network randomly with these questions, and respond when it received a reply from one of the users. The bot was also designed to aggressively follow users in the target network in order to maximize the total number of points that the team received. In two weeks, this bot received 198 messages from the target users and built 107 mutual connections with users.

Another model of note was deployed by a Boston-based team. Specifically, their bot was designed to identify two groups of users focusing on similar topics but unconnected by any relationships on Twitter. Then, it would follow users in the target group while copying the content of other unrelated users on Twitter as its own. While extremely rudimentary, this model in practice becomes a powerful method for building relationships with human users. In one trial, a bot copying the tweets of a community of competitive runners released a message stating that it had hurt “her” knee while training recently. The human community that it was embedded in sent messages expressing concern and inquiring whether or not it would hurt the bot’s chances for an upcoming marathon.

In the end, the teams were able to foster connections with over 50 percent of users in the target network. Most dramatically, the pattern of connections generated by the socialbots during this time period resulted in a significantly altered social topology with the bots becoming deeply enmeshed in the structure of the target network. Even within a short time period, we observed the bots being able to attract and develop relationships with human users that increasingly drew them away from the other human communities of which they were originally a part.

---


Social Hacking and the Future of Influence

The bots deployed in the previously described competition are more than just automated agents for chatting casually with unsuspecting users. The outcome of the socialbots competition suggests that relatively simple automated systems are able to provoke some notably large effects within a target online social group in a relatively short period of time. The ability to reliably do so (and at scale) would then seem to create an opportunity to consciously direct and shape the large-scale topology and behavior of social groups online.

Our team at the Pacific Social Architecting Corporation is currently experimenting with the deployment of swarms of these automated identities to affect the large-scale structure of social networks and norms on the level of tens of thousands, if not hundreds of thousands, of users. Early tests show that the ability for the bots to influence social networks scales accordingly. The team is also designing its current swarms with the ability to automatically assess the behavior of human networks and optimize strategies based on that data.

The ability to pull off large-scale “social architecting” of this sort is significant. Emerging research suggests that the structure of relationships in a community has a big impact on the spread of information in social groups, political activities, and personal practices. Insofar as series of bots are able to collectively and actively shape these behaviors, the technology seen in the socialbots competition may emerge as a remarkably powerful tool in social influence.

Implications

As this technology continues to develop and mature, there are a number of strategic implications posed in the national security and intelligence context.

First, and most obviously, is that swarms of bots may be used to disrupt the fidelity of open source intelligence. The bots deployed during the socialbots competition were realistic enough to fool average human users of the platform, and further development will generate ever more believable identities. To that end, an adversary may potentially be able to create sufficiently believable communities of false identities online to manipulate the inferences drawn from data on social platforms like Twitter, Facebook, and networks abroad. A grassroots movement, for instance, may appear larger than it actually is through the deployment of sufficient numbers of bots. Sufficiently sophisticated adversaries may also be able to taint open source data in ways that make it difficult to successfully assess the importance (and reality) of

---

“influential” players in shaping the discussion online. This ability to shape assessments of activity through the Internet may become particularly pronounced as more traditional news outlets with broader reach come to increasingly rely on these platforms to provide live “on the ground” news reports.

Second, automation permits the democratization of information operations. The resources to develop and deploy this technology are relatively small. Beyond a small monetary prize given to the winning team in socialbots, the entire competition was deployed for less than $500. For that amount, we were able to successfully and significantly shape the information environment of 500 human users online. Even a modest increase in resources might allow these types of efforts to scale considerably to broader sets of users.

Moreover, these bots do not need to be overly complex to be effective. Rather, even relatively simple systems seem to be successful in triggering user responses and overall changes in the structure of a social network. This suggests that small, dedicated third-parties may soon be able to launch ever more effective and subtle campaigns to shape opinion or spread disinformation in ways previously only available to actors with significantly broader access to resources. The easy design and deployment of these “weapons of mass persuasion” might make them an attractive option for adversaries in attacking the credibility and public effectiveness of their targets.

To hedge against these threats, tools should be developed that not only draw on open source intelligence on social networks, but also provide systems to evaluate the reality of the observed activity and actors. Like the dynamic space of computer security, it is likely that the social influence “exploits” discovered by people deploying these types of tools will emerge as quickly as they can be patched up.

Regardless, one thing seems clear: the next few years will see the emergence of new tactical and strategic doctrines in the active deployment and defensive methods of automated systems for social influence. Like the evolution of strategy that has been required in the development of drones in actual combat, or in the mass leverage of botnets in conventional cybersecurity, thought can and should be given to how similar automated systems might influence the characteristics of the battlefield of influence and narrative.

EXTRACTING INTELLIGENCE FROM SOCIAL MEDIA

By Rishab Aiyer Ghosh

Homs

The image is dark and blurred, but you hear voices raised, slow, rhythmic drumming, and the chants of hundreds. The camera turns and you can see men, women, even children standing and slowly clapping on the street, holding white balloons and Syrian flags. Families crowd the balconies above. Then, people gather around, link arms, and dance in a circle. The crowd goes wild, cheering.

The people of Homs are making a stand. They are out in force asserting their rights and their resistance to government troops whose artillery has heavily shelled the city over the weekend. More than any reports, more than any dry statistics, these sounds and images convey the spirit of the people of Homs and provide invaluable intelligence on how the protests in Syria are progressing.

The videos¹ of these events were posted to YouTube on October 10, 2011, after the weekend of shelling. Within a short while, they were highlighted as Trending Top 100 Videos on Topsy.com.

Arab Spring

When the protests began last winter in Tunisia and Egypt, and started generating headlines around the world, I had a personal interest in what was going on. I had visited Tunisia and Egypt several times, and had friends who were camped out in Cairo’s Tahrir Square. People on the ground were using social media to post pictures, videos, and text descriptions of things as they happened, but it was not easy to find this content. Searching on Google Images for Egypt showed (and still shows) only pictures of pyramids, camels, and maps, as Google is poor at ranking fresh content. Twitter’s own search engine mostly shows tweets as they happen, without any history and with very little ranking to highlight what’s really important.

Topsy.com, however, provides real-time search based on social media and is able to provide both fresh results and ranking with historical context. Instead of pyramids and camels, Topsy’s consumer website showed images captured on camera phones in Tahrir Square, ranked by the impact the images had on Twitter. Topsy’s “Expert Search” shows, for any search

¹ http://www.youtube.com/watch?v=nvU6yXkGmng and http://www.youtube.com/watch?v=yJE0HnYBOc
term, the influential people whose posts significantly impact ranking for search results for that term. It was gratifying, and a validation of Topsy’s algorithms, to see people I knew personally ranked as the top experts for Egypt, Cairo, and Tahrir.

From Publication to Conversation

This experience was also a validation of something larger — a shift that has taken place over the past few years in how information is collated and disseminated around the world. Information used to be spread through publication. People with some level of professionalism would find out what was going on, collate it, filter it, edit it, and disseminate it through a publication. This took the form of newspapers, magazines, wire services, and other media such as radio and TV. The rise of the Internet didn’t completely change this model; websites merely became new forms of publishing.

Indeed, when Google introduced authority-based ranking to search a decade ago, it did so with a “publication” model of the Internet, treating web domains as publications with identities, and links between them as citations through which authority flowed. But in the past few years, the dramatic rise of social media has led to a breakdown between authors and websites, people and web domains, information and publication. The best source of signal today is often from social media, because information is no longer being disseminated only through publication, but also through conversation.

Conversation was always the primary form of information dissemination. But conversations, even public conversations, were never really accessible, and had to go through a process of collating, filtering, and editing mostly by professional journalists until the information was published more widely. The only way to access information in conversations directly was to have some form of human intelligence: people on the ground, reporters, or insider contacts.

With social media, public conversations are now recorded, widely accessible, and searchable. Open sources are no longer only published sources, but now include the millions of people who are publicly conversing online. We don’t need professional third parties intermediating our experiences. We can all be in Tahrir Square, virtually.

The scope of open source intelligence has, therefore, vastly and rapidly expanded. It now provides access to information shared through those public conversations that previously could only be accessed through human intelligence embedded into the contexts where public conversations take place.

Human Intelligence Obsolete?

By greatly expanding the scope of open source intelligence, public conversations could theoretically replace a large part of human intelligence. They certainly provide the human element, as the examples of on-the-ground camera phone pictures show. But to what extent is this actual intelligence?

Open source intelligence drawn from publications relied on the editorial process to provide the intelligence. Without that editorial process, the valuable signal in the huge mass of public conversations is easily drowned in noise. If you were actually in Tahrir Square, but forced to hear everything everyone was saying there, you wouldn’t understand anything. When virtually in Tahrir Square, and also in every other location in the world where public conversations are taking place, most of what is going on is noise — whether it’s spam, celebrity gossip, or simply not relevant to your interests.

Extracting signal from all this noise is clearly essential in order to extract valuable intelligence from online public conversation. This signal extraction has to work for the peculiar nature of public conversations: they take place in certain locations; they are comprised of things individual people say about other people or things; and they consist of things said in particular sequence, in response to things other people say. A conversation cannot be understood without context. Signal from tweets, in other words, cannot be extracted without building a model that represents who is saying what about what to whom in response to what else.

Topsy’s technology stack was built from the ground up to do this analysis, filter out the noise, and rank the signal high. This is essential for search results, but also provides powerful analytical tools.
Anticipating Events Using Relative Trends

Historical activity levels of what people are discussing within the social web can provide excellent proxies for predicting what events might soon occur by providing an accurate estimate of the traction they are receiving in public conversation.

It is now possible to access deep historical counts of tweets by query terms so the data can be used to back-test a variety of hypotheses around how social media mentions impact real-life events. Everyone is aware that Twitter was used to communicate on-the-ground activities during the Arab Spring uprisings, but how correlated was public conversation online to events that took place? And to what extent would it be possible to determine, prior to an event taking place, some likelihood of it happening?

To test the relationship between the social volume of keywords related to real-world events during the Mideast uprisings, our team at Topsy employed share of voice (SOV) analyses to measure the relative change in activity for a given group of related keywords mentioned on Twitter over time. SOV analysis plots the proportion of mentions — typically, pre-filtered to limit noise — for a term relative to a basket of related terms, over time. Metrics used in SOV analysis can be scoped for a specific language, social data source, or geographic area. This is a useful technique for measuring the relative importance of something being mentioned on the social web over time within a given category of related keywords or phrases and other parameters.

Analysis

The following analyses cover three different types of activities during the January to March 2011 timeframe. The first analysis examines overall activity by Mideast country hashtags, followed by an analysis of Tunisian activity mapped to a chronology of real-world events. The final analysis is to see if protests organized on specific dates could have been anticipated.

The SOV analysis for countries shows interesting trends by country hashtag over time as depicted in Figure 1.

This analysis shows that #iran had historically dominated Mideast social communication, but dropped in relative importance when the Tunisian uprising started. Tweets related to Tunisia began appearing on 12/17/10, which is when Mohamed Bouazizi, the fruit vendor in the central town of Sidi Bouzid, lit himself on fire in front of a local municipal office. This act sparked real-world protests...
as well as a dramatic increase in online public conversation. Tweets related to Egypt actually spiked on 1/1/11, well before any major Egyptian protests began getting coverage in the mainstream media. But tweets about Tunisia overtook Egyptian activity. Could the Egyptian spike in late December have been an early warning of things to come in Egypt?

Tweets relating to Bahrain began to increase on 2/14/11, and persisted through the beginning of March, corresponding with real-world protests that occurred there. Tweets relating to Saudi Arabia initially flared up on 2/20/11, but were overtaken by Bahrain and then expanded in relevant importance during the first week of March.

A second analysis looks specifically at the relationship of mentions of Tunisia on Twitter to real-world protest events, as compiled by Al Jazeera.²

As can be seen in Figure 2, there is a correlation between events occurring and the chatter on Twitter. It’s not clear whether the social communication caused the events, but it’s certainly clear that the social communication amplified the information dissemination around each real-world event, and could have been used as an indicator of events unfolding on the ground.

Understanding the relative volume of social communication leading into an event can provide a potentially excellent signal for gauging the relative size of a pending event. This can be done by interpreting real-time data from social streams relative to historical baselines, quantifiable in terms of trending metrics such as velocity and momentum.

We tried to answer the question: Is it possible to anticipate which planned protests will occur and which will fail?

When people begin rallying around common causes on Twitter, they typically come up with a hashtag to describe the area of interest. This allows for postings to be grouped around a common “key” so that everyone’s communication can be easily referenced. In the Mideast, the syntax “(#date)” evolved as a way to organize people around specific protest dates, providing a convenient way for us to examine the activity volume and share of voice for groups of hashtag dates. Figure 3 shows SOV analysis around some of the hashtags used to represent rallies across different countries. We ran this analysis against a number of dates and filtered out those for which there was obviously little or no traction.

It is interesting to note that some hashtag dates became significant while others peaked low and faded. It is clear that the SOV analysis by protest dates can be used to identify efforts to organize protest dates, providing an excellent leading indicator of the interest for each potential protest. Assuming “hashtag date” syntax continues to be used to organize events, this seems like an excellent indicator to gauge the popularity of an event leading into that date. Similarly, if there are other keywords (hashtags or any other

² http://english.aljazeera.net/indepth/spotlight/tunisia/2011/01/20111414223827361.html
term] used to describe an upcoming event, monitoring the activity for these terms relative to other related terms would provide an excellent way to understand the relative popularity of that event.

**Conclusion**

Millions of people are using social media to hold public conversations, sharing facts and opinions about the world around them every day. These conversations can be aggregated and measured in real-time, replacing a large proportion of human intelligence, as long as the right technologies are applied to filter signal from noise, generating open source intelligence from raw, noisy public information. As shown in the example analysis, such intelligence can be used to learn about, understand, and anticipate real-world events.

The people of Homs were expressing their resistance to the regime, but by using technology, they were able to express this to the whole world. Their voices deserve to be differentiated from all the background noise, and their message heard.

---

*Rishab Aiyer Ghosh* is Co-Founder and VP Research of Topsy Labs, Inc. In this role, Ghosh has developed a number of key patented or patent-pending technologies, including intelligent reputation attribution platform and estimating social network influence [U.S. Pats. 7831536 & 7991725], real-time ranking entities based on social network influence, trend detection in social networks, and the identification of relative expertise across social networks (several patents pending). He started “First Monday,” the most widely read peer-reviewed journal of the Internet, in 1995, with Ed Valauskas, Esther Dyson, and Vint Cerf. In 2000, at the University of Maastricht in the Netherlands, Rishab started the Collaborative Creativity Group, the leading research group on the economics of free/open source software, Wikipedia and other forms of collaborative innovation. He has extensively researched and published how reputation works and motivates in online communities for more than 12 years, collaborating with Stanford, Oxford, Cambridge, and Tsinghua Universities, with grants from the U.S. National Science Foundation and European Commission. Rishab is also a board member of the Open Source Initiative.
BASICS OF SOCIAL MEDIA ANALYTICS
FOR COUNTERTERRORISM SPECIALISTS

By Rebecca Garcia

It is well-known that extremist organizations have embraced the Internet for its vast and immediate worldwide reach for recruitment. In 2006, the University of Haifa’s Gabriel Weimann published groundbreaking research in his book *Terror on the Internet: The New Arena, the New Challenges*, and discovered approximately 4,300 terrorist-related websites affiliated with 40 active terrorist groups as defined by the U.S. Department of State. This parallels the increasing role websites play in funneling information to all of us. As consumers spend more time using social media, it should not be surprising that extremist organizations are as well. They see social media as a primary tool for bringing together communities of interest, regardless of geographic location, and for discussing ideas in a seemingly anonymous manner.

In recent years, SAS has been working with commercial interests in the area of social media analytics, helping make sense of the vast number of conversations on the Internet to better determine how organizations are perceived. This can entail culling textual and numeric data from websites in various formats (semi-structured and unstructured), including blogs, message boards, chat rooms, and other social media and social networking sites to summarize the conversations about a given topic. It involves analyzing nouns to determine relevance to a given topic and understanding the use of adjectives to determine positive, negative, and neutral sentiment across many different languages. Ultimately, a variety of insights can be delivered, whether it is the trending of negative sentiment associated with a brand, detailed analysis of key phrases that resonate among certain groups of people, or determining which individuals have the most influence in a given social network.

The premise of this article is that many of the techniques used to distill the essence of online conversations for businesses serving their customers can also be applied to analyzing conversations among terrorist cells. The area of social media analytics could feasibly emerge as an indispensable counterterrorism tool.
tool by enabling proactive monitoring, analysis, and engagement through extremist social networks and associated digital properties.

The Challenge: Information Overload

It has become obvious that the Internet is not only the communications infrastructure of choice across the full spectrum of society; it also serves as an effective operational infrastructure for various organizations, including terrorist cells. It can be accessed from anywhere at low or no cost, knows no physical boundaries, provides the means to transfer funds, offers media-rich training to facilitate distance learning, operates in real-time, and is always on. And much as a system administrator has tools to understand and diagnose traffic on an internal network, businesses increasingly expect, and are realizing, similar capabilities in dealing with social media chatter. But analyzing this information is not without its challenges, including:

- Too many websites to monitor/too much information flowing through in real-time.
- Messages are unstructured text, written in various languages, and subject to language nuances, such as slang, colloquialisms, abbreviations, misspellings, etc.
- Difficulty of determining a user’s identity.
- Difficulty of differentiating legitimate actors from the casual visitors.
- Difficulty of seeing any trends or emerging topics due to volume, veracity, and velocity of data.
- Significant subject matter expertise required to understand the messages.
- Because of all of the above, there is no easy way to automate the data mining process.

The Promise of Social Media Analytics

The area of social media analytics has emerged to help companies wrestle with many of the previously listed challenges. It allows companies to listen to and monitor details of conversations about their products and services on social media and online channels, ranging from Twitter to customer transcripts residing on internal systems. By continuously accumulating and archiving these conversations, companies can understand trends and predict future conversation volumes to determine how they will affect business results.

The biggest immediate benefit companies see from social media analytics is eliminating the need to analyze social media data in a piecemeal and ad hoc fashion. Instead of logging into hundreds of discrete sites, the information is all represented in one place. And instead of a chaotic, jumbled mess of sentences, businesses can impose order, categorizing sentences on the basis of how their nouns map to specific business attributes. For example, a sentence containing a complaint about phone call wait times becomes categorized as a negative instance specific to customer service. Millions of sentences add up to millions of insights on challenges and opportunities like pricing, corporate reputation, customer loyalty, and quality.

Lessons Learned in Social Media Analytics

As businesses have embraced the brave new world of social media analytics, there are lessons that can be applied to counterterrorism analytics.

1. Treat online conversations as information assets.

Many organizations in the private sector track and archive customer activity through sales data, yet they often have no capture or archival strategy for analyzing and retaining what those customers say about the business publicly. Similarly, intelligence organizations often have detailed information on the criminal history of an individual, yet lack the ability to connect that with public conversations.

As businesses have learned, what is being said online can be a leading indicator of what is about to happen to that business. Anyone who has read a negative hotel review on Expedia or a negative product review on Amazon can attest to the power of crowds to influence
opinion. But the ability to link online perception to offline activities is completely undermined if those same conversations are not properly categorized as relevant to an organization, or not archived long enough for a good analyst to discover that linkage.

For these reasons, the foundation of any social media analytics effort should be the collection, integration, and storage of online conversations for several years of historical analysis, combined with information from internal systems to allow for deeper, more holistic insights. It is crucial to have this depth and breadth of conversation history to understand the difference between a fad and a trend. This is relevant to the counterterrorism community as well. Correctly categorizing a written reference to a known terrorist, belonging to a known terrorist cell, pre-9/11, in Afghanistan, is only valuable if it is actually available for analysis and is easily matched with other intelligence data.

2. **Understand the sentiment of conversations linked to specific topics and issues.**

Understanding the web traffic generated by specific social media activities — or assessing the tone of references to a specific topic — does not go far enough in helping develop insights that drive action. To begin understanding the issues, sources, and trends specific to the counterterrorism mission, you must be able to select which topics, media sources, and key conversations to analyze, and then choose how to classify them.

In the past, the level of detail collected from online conversations was insufficient to guide important decisions. For commercial interests, it has long been difficult to accurately determine topics that were important and perceptions that were associated with a particular topic. Unable to link conversations to topics, organizations could not quantify the impact of online content on their overall performance. This is another area where social media analytics provide a critical link. In terms of counterterrorism, the sentiment analysis capability can be used to measure online reaction to changes or initiatives to specific topics, such as “U.S. foreign policy,” “military operations,” “foreign presence,” and “foreign aid.” This effectively structures the unstructured, making public sentiment analysis possible.

3. **Eliminate analyst bottlenecks.**

Lack of data is not the pressing issue for most intelligence professionals. It is lack of insights. Even the best data analysis is worthless if it does not reach decision makers on time with clarity. Therefore, it is critical that the very best thinking go into data visualization techniques that are designed for decision makers, not just analysts, with turnkey capabilities in sentiment trending, author analysis, threat tracking, and phrase clustering. In this way, identification of escalating issues, the most influential actors, and emerging trends and topics are delivered to those in the best position to capitalize on that information.

Even the best data analysis is worthless if it does not reach decision makers on time with clarity.
Practical Applications for Counterterrorism

Through the approaches described previously, there are a number of application areas that social media analytics offer counterterrorism strategies. They include:

The What. Businesses expect a real-time read on social sentiment relative to their brands, and counterterrorism specialists should expect the same of the "brand" of any terrorist. Understanding the key phrases, sentiment, and volume of conversations encompassing specific terrorists can be an indicator of activities.

The Virtual Where. In a world with billions of websites, simply understanding which online properties have a critical mass of disaffected, relevant people is not a trivial undertaking. Being able to consume chatter across all types of websites and determine their relevance to your organization’s mission is among the most important benefits to any listening platform.

The Physical Where. Much of the global population’s use of social media sites is driven from mobile devices, which increasingly contain coordinates of a user’s location. This allows organizations to begin to understand geographic characteristics of extremist chatter, and more quickly identify domestic sources of unrest. Furthermore, the Internet protocol (IP) addresses of devices connected to networks can increasingly serve as accurate sources of location-based insights.

The Who. There is a treasure trove of information surrounding the connections between people. This can yield an organizational hierarchy of individuals that can be used to determine the more important users across the forums.

The Why. Connecting social insights to real events on the ground is the ultimate goal. Correlating online chatter to historical events can yield clues as to why certain chatter is noise, while other chatter can be an indication of more serious events to come.

Conclusion

In the post-9/11 world, government organizations with complex challenges would often see enormous programs created with budgets of $500 million to $1 billion awarded to large systems integrators. More recently, the administration mandated that these enterprise programs be scaled down to much smaller projects, in an effort to reduce costs and, more importantly, to facilitate more rapidly successful solutions to critical challenges.

Counterterrorism organizations have to straddle technology investment that is proven and current, while staying within budget and ensuring adoption. But technology is only a means to an end; it is ultimately people who understand the specific issues inherent to counterintelligence who will ensure success. Therefore, the human intuition of experts in intelligence must inform social media analytics technology on what to look for, so that computer intelligence is better able to inform the humans about the content, correlation, and trajectory of social conversations. It is only through this approach that technology of any kind can meet the 21st century needs of government intelligence, and perhaps is the most important prerequisite in evaluating a social media analytics effort.

Rebecca Garcia is a Director at SAS Federal, where she focuses on the Intelligence Community (IC). In this role she is responsible for ensuring SAS delivers analytic solutions to the IC to help solve mission and IT critical challenges. Rebecca has supported this community since 1997 and has worked for The Boeing Company, Northrop Grumman Corporation, and KPMG Federal. She has a Bachelor of Arts degree in Economics from the University of North Carolina at Greensboro.
To supplement the IQT Quarterly’s focus on technology trends, Tech Corner provides a practitioner’s point of view of a current challenge in the field and insight into an effective response.

DATA AT THE SPEED OF SOCIAL

A technology overview from IQT portfolio company Visible Technologies

Imagine you try out a new search startup that wants to compete with Google. The first thing that they ask you to do is set up an account, where you provide the ten queries that you plan to run on their system. After several hours or days, you receive word that you are cleared to start running queries, as long as you only run one of the 10 queries provisioned.

Most social media monitoring and analytics vendors avoid the “big data” problem and instead limit users to focused topics over short time periods. Each topic is then a small data problem where traditional databases and indexing techniques work reasonably well. In some cases, these vendors use technologies like Hadoop to store large-scale social data. However, access to this data is only available through high-latency, batch-oriented actions like account setup and topic provisioning.

While this strategy can work for monitoring and measuring known topics, the agility required to experiment, learn, and adjust is missing. Even simply reacting to emerging topics and feedback through social engagement can be difficult.

Big data initiatives abound and most use Hadoop (or its variants) to store and analyze data due to its elastic linear scaling and fault-tolerant architectural features. With the increasingly widespread adoption of these technologies and their availability through cloud vendors such as Amazon Web Services, scaling out to petabytes is not the difficult and expensive problem it once was. So why aren’t we seeing vendors that can support ad hoc analysis of the entire social web in near real-time?

Increasingly, the problem has become speed, or “Fast Data.” Batch-oriented map/reduce architectures work very well when measuring latency in hours; they do not work well for interactive, near real-time applications.

An Illustration of Fast Data

Social media is an ever-growing source of insight into human perspectives on nearly every imaginable topic, from Bieber to baseball to Burberry. A look at the British royal wedding in the spring of 2011 shows the scale of a social media event.

If we limit ourselves to retrospective analysis of tweets containing the hashtag #royalwedding during the course of the event, it’s a simple problem with only about 1.5 million posts to analyze. But can big data technologies keep up with Fast Data and provide insights in near real-time?

Testing 2 billion social media documents stored in an 85-server cluster running Hadoop and HBase, we captured the following average latency measures:

- Random access to retrieve a single document by its primary key: 0.015 seconds
- Random access to retrieve documents by a non-key attribute (such as URL or author): 30 minutes
- Total document count: 40 minutes
In order to provide powerful Business Intelligence (BI) and interactive exploration of social media data, Visible Technologies designed a platform with not only elastic linear scaling and fault tolerance, but also low-latency data access. Results from the same tests using the Visible platform are summarized below:

- Distinct author count: 45 minutes
- A single mapper task initialization: 15 – 30 seconds
- Multi-lingual, free-form text search: Not easily supported

In an abstract form, the fundamental building blocks of the Visible Intelligence (VIQ) platform are processing units, each of which is composed of a processing engine and a set of documents. The processing engine is responsible for fulfilling user requests and keeping the set of documents up-to-date. Documents are distributed to processing units based on a partitioning and replication specification.

Processing units are assigned to servers at run-time based on a distribution policy. Each server acts as an intermediate response integrator that integrates and compresses the responses from processing units and sends them back to the requester. Processing units are reassigned to other servers in case of server failure.

At run-time, an API layer determines the servers to communicate with based on a selection policy. The API layer also serves as a final response integrator to integrate the intermediate responses computed by the servers. Web services communicate with the VIQ platform via the API layer.

For near real-time data freshness, the processing units responsible for fresh data use an in-memory data store to achieve maximum throughput and
minimum latency. Currently, each in-memory processing unit can process up to 500 documents per second.

By design, all requests are fully distributed and all processing units function autonomously. A single request is processed by all the processing units concurrently to respond in low-latency. This also allows the system as a whole to scale linearly and elastically and support a large number of concurrent users.

All processing units and response integrators are able to handle updates. Updates may come from data sources (e.g., an author updated a document), from Visible’s internal system (e.g., Visible introduces a new data enrichment), or from users (e.g., an end-user tags a document or overwrites one of its attributes). Updates are managed by the following incremental response computation:

\[
\text{Response} = \text{Response}_{\text{CurrentSnapshot}} - \Delta_{\text{old}} + \Delta_{\text{new}}
\]

An update is stored as a before image ($\Delta_{\text{old}}$) and an after image ($\Delta_{\text{new}}$). All processing units and response integrators first compute a response from the current snapshot of the document set, then semantically negate the before image and integrate the after image. Most of the time, an end-user triggered document update will be processed and available to the user in less than a second.

### Social Search at Scale

VIQ has fast, federated search across the entire indexed social web, but what about search quality? On one hand, simple retrieval models leave much to be desired but on the other, “semantic” is often a euphemism for “slow.”

Visible Technologies has been able to adapt high-performance information retrieval (IR) models to social data. Most off-the-shelf IR technology is based on well-written documents and encodes assumptions about vocabularies, document lengths, punctuation, and capitalization that don’t hold true on most social data.

- **Classic IR techniques ignore punctuation and stop words.** Emoticons, hashtags, @references, and stop words can be important in social media.
- **Classic IR techniques value word repetition.** While a book chapter mentioning a search term repetitively may make it more relevant, in general this is not true for social media. (“iphone iphone iphone iphone iphone” is an actual tweet, but it’s not very useful.)
- **Classic IR techniques value shorter and purer documents.** While five mentions in a paragraph might be more relevant than five mentions in a chapter, this is generally not true for social media. (A one-word tweet “iPhone” is a perfect, pure match, but not as useful as longer documents that also mention iPhone a single time.)

In addition, Neuro-Linguistic Programming and language identification models learned from other domains are generally too brittle to apply to language routinely used in social media data. The good news is that often social data (especially micro-text) is less semantically complex than traditional documents. For example, tweets often contain only a single topic. Statistical approaches, even those using naive bag-of-word assumptions, often yield very good performance when trained on social data.

### Data Enrichments at Scale

One of the fundamental tasks in social media analytics (and text analytics in general) is to derive structure from unstructured content. In the Visible Technologies platform, its structured dimensions are derived from the content, people, and sites of the social web through its enrichment pipeline. In addition to social dimensions like language, sentiment, and influence, Visible’s statistical models derive a number of other demographic (age, gender, location), psychographic (interests, affinities, intents), and topical (brand, theme, concept) attributes. These dimensions are exposed through faceted search allowing on-demand aggregation, drill-down, and slice-and-dice across any set of dimensions.
Most of Visible Technologies’ statistical models are developed in a supervised learning framework in which models are trained offline. For these models, speed of training is not critical, but can become an encumbrance at web scale. Execution (scoring) performance, however, is critical. Time spent enriching content in the collection and ingestion pipeline is measured in microseconds and deploying new models often involves back-filling billions of social media documents. Visible generally uses fast maximum margin or stochastic gradient boosting-based models for core platform enrichments.

In other scenarios, training speed can be crucial. Visible is experimenting with online learning from analyst behavior and feedback as well as topic modeling in rapidly changing situations. Learning what constitutes “actionable” content or modeling the probability of successful interventions should happen through real-time feedback. Visible uses fast online techniques based on stochastic gradient descent for classification/prediction and based on latent Dirichlet allocation for topic modeling.

Analytics at Scale

Visible Technologies analytics supporting search and BI on a broad set of data enrichments offer the flexibility for unconstrained and rapid data exploration at scale. However, discovery and insight features are also critical in order to make sense of the continuous tsunami of social data. How can we scale up opinion mining (e.g., text mining, data mining, graph mining) techniques to distill insights in near real-time? While some algorithms like the eigenvector centrality measures used to calculate social influence propagation can be scaled up with relative ease, classic clustering and summarization techniques cannot.

Visible developed very fast, distributed summarization techniques that extract statistically relevant summaries (words, phrases, themes) integrated with the BI capabilities of our platform. The social media posts belonging to any segment are summarized on-the-fly as the analyst aggregates, drills, slices, and dices. These summaries are based on the principles of reducing information entropy and maximizing marginal relevance.

The Value Proposition

While modern search engines have done a good job of taming web scale data, the same level of agility has not previously been available for BI, analytics, and discovery at scale. By tackling these challenges, we can elevate social media analytics on Fast Data beyond simply reading through emerging content. We can elevate social media analytics on big data beyond retrospective analysis of events that are no longer actionable.

Visible Technologies enables analysts to understand and react to emerging events now, not just through search but through state-of-the-art analytics and BI, and on emerging, connecting, and branching topics within the context of historical trends. Users can explore the conversation, pivot through consumer sentiment, dive into the themes and drivers, and quantify local reactions in different channels and around the world.

Visible Technologies, an IQT portfolio company, is an enterprise ready social media solution offering a combination of software and services to harness business value from social communities. To learn more, visit www.visibletechnologies.com.
The IQT Quarterly examines trends and advances in technology. IQT has made a number of investments in social media technologies, and several companies in the IQT portfolio are garnering attention for their unique solutions.

**Recorded Future**
Recorded Future was recently featured in prominent news outlets such as Wired, Forbes, and the New York Times for its innovative approach to data analysis. The company’s technology sifts through vast amounts of web data to develop predictions about possible future events. By analyzing currently available information, Recorded Future’s software places events on a timeline known as a “temporal index” that is then used to make predictions. It is used for data monitoring and prediction in financial services and trading, brand management, and competitive intelligence. Recorded Future is based in Cambridge, MA and has been a member of the IQT portfolio since January 2010. [www.recordedfuture.com](http://www.recordedfuture.com)

**Geosemble**
Geosemble, an IQT portfolio company since March 2009, has been expanding its line of product offerings to provide geographic content search and discovery to the government and private markets. Geosemble’s technology identifies and extracts geographic data from a wide variety of sources. In 2011, the company issued new updates to its core software GeoXray, including the PlaceGenius feature, which allows users to extract key geographic information from unstructured sources in addition to traditional data sets. Geosemble is headquartered in El Segundo, CA. [www.geosemble.com](http://www.geosemble.com).

**GeoIQ**
IQT portfolio company GeoIQ was recently recognized as a finalist in the first annual LOCAL awards. The awards, hosted by leaders from the location-based services industry, are given to top innovators in a variety of categories, including location-based advertising and branding, commerce, and mapping. Based in Arlington, VA, GeoIQ was named as a finalist for Best Location-Based Data Platform for its suite of geographic intelligence products. GeoIQ tools allow users to visualize geographic information on a map using public or proprietary data. GeoIQ has been a member of the IQT portfolio since May 2007. [www.geoiq.com](http://www.geoiq.com).

**NetBase**
NetBase is an IQT portfolio company whose software analyzes social media to gauge consumer sentiment about a variety of products and topics. It has proven to be a useful tool for brand intelligence, positioning NetBase as a trusted market research and analysis tool and gaining them a number of high profile clients. The technology is increasingly being used by business journalists, and has been featured in recent articles in the Wall Street Journal, New York Times, and Computer World. NetBase has been a member of the IQT portfolio since November 2010, and is based in Mountain View, CA. [www.netbase.com](http://www.netbase.com).