

Sysomos Inc.

Business Intelligence for
Social Media

contact@sysomos.com

www.sysomos.com

(866) 483 3338

120 Baldwin St, Suite 3

Toronto, ON

M5T 1L6

Whitepaper

April 2009

Social Media Meets Business Intelligence

Vue Magazine & Canadian
Marketing Association
Report on Social Media
Analytics

Dr. Nick Koudas





ABSTRACT



User generated content (e.g., web logs or blogs, social networks, micro-blogging such as twitter, wikis, collaborative tagging, news, music sites, podcast and video sharing sites) are proliferating at unprecedented rates. The numbers quantifying user participation are astonishing; more than half a billion individuals in social networking sites worldwide, in excess of 100 million blogs, millions of users utilizing micro-blogging services, etc. In aggregate such services generate very large amounts of data on a daily basis. Commonly the word social media is used to refer to this information collective, primarily contributed by individuals online.



user generated content:

- blogs
- social networks
- wikis
- podcasts
- microblogs
- news
- video

Historically corporate databases have been accumulating information about any aspect of a business from employment records to corporate supplies and sales records. Business Intelligence is a mature set of technologies offering deeper understanding of such data. For example such technologies inform analysts and executives not only that the volume of sales in a region is increasing or decreasing but also point to the reasons behind such changes.

In this article we argue that business intelligence technologies and functionalities can be of great service to social media as well. This wealth of information contains invaluable knowledge and intelligence for marketers, public relations agencies, advertisers, political campaigns, law enforcement agencies among many others.





INTRODUCTION – FROM DATA TO INTELLIGENCE

Efficient collection, archiving and cleaning of such collective knowledge provides the foundation for robust analytics from which business intelligence may be extracted. Such analytics can enable one to contrast events across multiple time periods and identify when and where, and most importantly for what reasons, perceptions or acceptance of brands and products are changing.



Analytical Capacities

1. Demographics
2. Sentiment
3. Crisis Management
4. Influencer Identification
5. Measurement
6. Comparison
7. Engagement



1-2 For example we can identify the target demographic (e.g., males in eastern US of ages 19-29) that have extremely positive opinions about a new tech gadget. Similarly, we may locate the target demographic with the most negative opinions about the same gadget and automatically discover the reasons behind such sentiment (a lot of disappointment around the screen size and graphics in the new gadget).

3 By tapping into the information collective generated by users on a daily basis, one can identify crisis events for brands, products, individuals or the general public, as they happen and react effectively.

4 One can effectively identify how ideas or rumors spread across the globe and identify key influential sources that primarily contribute to such spread (e.g., the crisis situation for a brand started when a particular posting appeared on a blog with wide readership, escalated into several online forums and subsequently made it to wide circulation news sources; the top 10 blogs and forums actively writing about the crisis with the highest readership can be subsequently identified).

5 Gain insight on the impact of an advertising campaign around the globe as a function of time (positive sentiment towards a brand increased steadily in a particular demographic following a particular online ad campaign with the video ad receiving the most positive engagement online).

6 Obtain feedback on how key competitive products and brands fare comparatively at several locations around the world or in aggregate, across select demographic groups (professional males in their 30's prefer product A to B in the US, but the situation is different in Canada).

7 Identify key communities or individuals online to engage with (a technology centered community with specific focus on wifi enabled gadgets). Taking such desired functionality one step further, it is possible to contrast such information with key performance indicators (e.g., volume of sales as a function of time) and understand correlations and influence of social media on the bottom line.

It is evident that the ability to make sense of the information in social media offers tremendous value to marketing, public relations and brand conscious corporations as they are now able to obtain continuous feedback from their



“New media, social media and internet connectivity has become increasingly important... this has led to fundamental shifts”



customers, understand issues, react to them as well as efficiently engage with consumers. Such ability is becoming increasingly important.

A variety of studies point to changing trends in the ways people communicate and exchange information, the ways people choose to be informed, the types of preferred media for their entertainment. New media, social media, and internet connectivity has become increasingly important. As a result, this has led to fundamental shifts in the way people research products to obtain information and the way in which opinions are expressed. Given the volume of information involved it is evident that any attempt to manually make sense of this information collective is at best extremely slow and expensive, and more likely destined to fail.

Technological advances enable the effective processing of very large volumes of information in real time. Hence we outline the challenges associated with an attempt to provide business intelligence insights utilizing the social media collective and highlight how technology can aid the ability to collect, process, and most importantly interpret social media.

DATA COLLECTION

Social media, which consists of blogs, wikis, message boards, social networks, podcasts, microblogging, bookmarking, and online videos, is a highly diverse and *heterogenous* set of data. Collecting data from heterogeneous sources presents several challenges; especially related to dissimilar data formats, data types and non standard meta-data tags.

An additional challenge is related to data source discovery. Blogs (especially those not residing on hosted services such as blogspot or livejournal) need to be discovered the moment they are created. If this data is to be used effectively, new information from blogs (new blog posts), social networks and the rest of the social media collective, should be collected the moment they are publicly available. Collectively, multiple million opinions are posted online by consumers on a daily basis. Given the rapid evolution of social media services and the ever increasing volume of data, robust and exhaustive data collection is a fundamental first step in enabling business intelligence on social media.

DATA CLEANING AND SPAM REMOVAL



A large fraction of the information in user generated content is spam. According to some statistics, more than half of the content hosted at blogspot (a blog hosting service provided by Google) is spam. Spam is primarily created for search engine optimization purposes but also for malicious advertising and phishing attacks. Although there are many ways to create spam, the one that is more pronounced is malicious content. Spammers inject content unrelated to, say, a blog post in order to cause the post to be relevant to many possible search queries.



-
-
-

“The presence of spam is harmful to any attempt to understand content from social media”

-
-
-

Then the actual (spam) post commonly contains links to several sites (advertising malicious content).

The presence of spam is harmful to any attempt to understand content from social media. Spam content introduces noise and clutters any emerging discussion or themes around topics of interest. It is imperative to identify and remove spam before processing social media content. A variety of techniques to remove spam exist. Such techniques aid to a certain extent further processing of content to facilitate automated comprehension; as spam filtering techniques evolve, identifying and removing spam is an ongoing battle for providers of business intelligence solutions on social media.

ANALYTICS

The key to unleashing the power of social media is gaining enhanced understanding of its content. A starting point would be to obtain simple metrics on the volume of mentions of entities of interest in online content. Having spam free content is imperative in order for such counts of mentions to be meaningful. This functionality will allow to observe increase or decrease in the volume of mentions (or buzz) around an entity of interest and indeed being able to compare several entities of interest in terms of mentions online (being related products, competitors, etc). Several solutions exist along these lines including some free solutions.



We argue that it is possible to obtain understanding of social media far superior than what is presently available in the form of counts of mentions and content clippings. An increase or a decrease to the volume of mentions of an entity probably corresponds to some event of interest. Being able to understand such events requires significant effort if the volume is high and/or if multiple attributes/features/properties are commonly associated with the entity of interest. The first powerful functionality is the ability to understand discussion topics and identify the main conversations around an entity of interest. This will enable us to further focus on what is actually important and refine our search for information (e.g., there is a lot of active discussion around the 'screen size' of a specific tech gadget as well as a different discussion thread around its 'wifi' capabilities; based on this information we can quickly focus on each discussion separately).

In many cases the volume of chatter/discussion around a specific topic will be vast, easily surpassing thousands of mentions of the brand of interest. In that case powerful summarization features can readily distil what is important in the discussion, saving time and cost associated with reading all the content. Recent advancements in summarization technology enables very fast summaries of large document collections. The basic idea behind summarization is to transform documents into points into a high dimensional 'document space' and subsequently reduce the dimension of this space, by essentially keeping only subspaces that contain the most information. Commonly diverse and polar opinions exist about topics, products and their features. The ability to understand the sentiment of opinions towards



specific topics or entities of interest is imperative. This will enable power grouping and summarization of discussion around positive and negative sentiment enabling further topical analysis across sentiments.

• • •

“The ability to understand the sentiment of opinions towards specific topics or entities of interest is imperative”

• • •

Social media is a truly global phenomenon; content exists in all languages. Support for search as well as advanced analytics across all languages is a key requirement. The requirement becomes a necessity for global corporations to track issues across the planet.

Given the diversity of social media sources, the ability to identify influential and authoritative individuals in each media source is very important. Although several simple measures of authority exist today (e.g., number of in links to a blog) we argue that the accumulated history of the online activity of individuals enables much richer forms of authority and influence to emerge, especially when they are coupled with specific business objectives. Availability of technical tools with ability to understand varied business objectives in order to conduct influencer search is therefore of critical importance.

Finally, across social media sources, several topical communities emerge. Supporting efficient identification of such communities across topics as they emerge and evolve is an additional key requirement. Given the fragmentation of social media today across multiple heterogeneous sources the ability to understand and associate communities together across sources is a pressing need.

GEOGRAPHY AND DEMOGRAPHICS



Social media content is contributed by users, individuals residing in different places of the world as well as belonging to different demographic groups, having diverse interests, backgrounds and professions. Capturing information about the location as well as demographic information about the individuals contributing information offers a way to obtain understanding of social media in various geographies, diverse demographic or interest groups.

Such information offers enhanced understanding of particular interest to those aiming to comprehend geographies or demographics. Coupled with the ability to factor time in the analysis, it offers a powerful capability to understand how perceptions and opinions change temporally.



CONCLUSION



Social media are empowering individuals and are redefining the media landscape. We argued that technology can empower marketers and public relations specialists to gain enhanced understanding of social media, offering functionality far beyond what is available today in the market place. The second generation of social media analytics is here as a result of technology breakthroughs aligned to meet business needs. This new generation of social media analysis platforms offers a convergence of business intelligence and social media.