

# SOCIAL MEDIA ANALYTICS

THE NEXT GENERATION  
OF ANALYTICS-BASED  
MARKETING SEEKS INSIGHTS  
FROM BLOGS.

By Rick Lawrence,  
Prem Melville,  
Claudia Perlich,  
Vikas Sindhvani,  
Steve Meliksetian,  
Pei-Yun Hsueh  
and  
Yan Liu



© Photographer: Saniphoto | Agency: Dreamstime.com

**Over the past decade,** the Internet has created new channels and enormous opportunities for companies to reach customers, advertise products and transact business. In this well-established business model, companies fully control their own Web-based reputation via the content appearing on their Web sites. The advent of Web 2.0, with its emphasis on information sharing and user collaboration, is fundamentally altering this landscape.

Increasingly, the focal point of discussion of all aspects of a company's product portfolio is moving from individual company Web sites to collaborative sites, blogs and forums – collectively known as *social media*. In this new media essentially anyone can post comments and opinions about companies and their products, which may influence the perceptions and purchase behavior of a large number of potential buyers. This is of obvious concern to marketing organizations – not only is the spread of negative information difficult to control, but it can be very difficult to even detect it in the large space of blogs, forums and social networking sites.

One rough measure of the potential magnitude of the impact can be gleaned from recent studies about use of the Internet. In July 2009, a survey conducted by Universal

McCann [1] concluded that 31.7 percent of the more than 200 million bloggers worldwide blog about opinions on products and brands, and that 71 percent of all active Internet users read blogs. The 2009 Nielsen Global Online Consumer Survey [2] of 25,000 Internet users in 50 countries has found that 70 percent of consumers trust opinions posted online by other consumers. One piece of good news for brand marketers is that a similar number (70 percent) of consumers trust brand Web sites where the message can be carefully crafted. But consumers are just as likely to be influenced by what they read in social media.

While the expansion of user-generated content in the blogosphere poses major challenges to traditional marketing, it also opens huge opportunities for marketing organizations to differentiate their strategy by leveraging social media to their advantage. This requires not only new thinking, but also new, automated analytics-based capabilities that are now defining the emerging discipline of *social media analytics*. From a technology perspective, social media analytics touches a number of disciplines including social network analysis, machine learning, data mining, information retrieval and natural language processing. This article discusses at a high level some of the ideas

that utilize these analytics capabilities to provide marketing insights from blogs.

## Social Media Analytics for Marketing

FROM A MARKETING and market intelligence perspective, blogs are a very important form of social media because they provide access to previously inaccessible information such as specific customer insights and opinions. Social media analytics can address several interesting questions by providing algorithms and approaches for the automated analysis of blogs and related social media:

1. Given the massive size of the blogosphere, how can we identify the subset of blogs and forums that are discussing not only a specific product, but also higher-level concepts that are in some way relevant to this product?
2. What sentiment is expressed about a product or concept in a blog or forum?
3. Who are the most authoritative or influential bloggers in this relevant subspace?
4. What are the novel emerging topics of discussion hidden in the constant chatter in the blogosphere?

A typical blog or micro-blog has one author (the blogger) and consists of multiple entries or posts. It is useful to think of a blog in a three-dimensional space defined by the first three metrics above: relevance, sentiment and authority. While the first two dimensions, relevance and sentiment, are specific to a given post or even smaller section of text (“snippet”), the notion of authority is most naturally assigned at the blog level. A blogger’s authority can also depend on the specific topic. Emerging topics are a property of the blogosphere at large and require analysis across many blogs.

All three dimensions are important and they need to be considered in a unified view in order to provide marketing insight. One way to provide such a view is to determine the relevance and sentiment of each post, and characterize the overall relevance and sentiment of the blog as a simple statistic over individual posts.

Figure 1 captures such a blog-centric view along these dimensions from a prototype tool at IBM Research. Here, we are interested in a high-level view of blogs relevant to the broad topic of “social collaboration.” Relevance is shown on the y-axis, and sentiment is on the x-axis – both metrics are computed at the post level and aggregated to the blog level. Each circle represents a blog, and the size of the circle reflects the blogger’s authority. The output of the model can be interpreted as the probability that the post is positive in tone. We are most interested in extremes of sentiment, so we naturally look for authoritative blogs in the upper-left and upper-right quadrants to find the most relevant blogs with non-neutral sentiment. Such a view allows marketing people to quickly identify blogs of interest, and to drill down to obtain more specific understanding of the potential marketing impact.

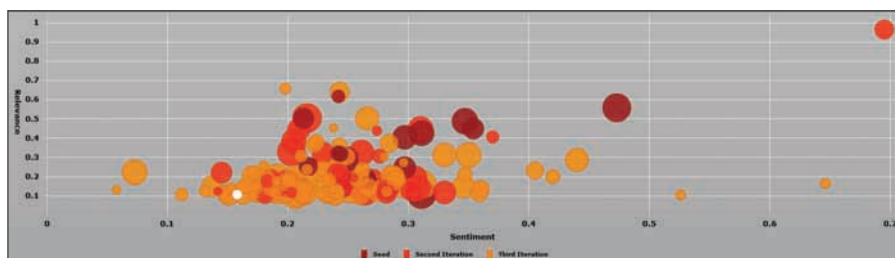


Figure 1: Relevance, authority and sentiment at the blog level.

## Finding the Relevant Blogs

OUR FIRST OBJECTIVE is to filter the vast blogosphere from millions to the thousands of blogs most relevant to the topic of interest. In the simplest case, the “topic” can be a specific product, and the objective is therefore to identify all blogs discussing this product and perhaps competing products as well. More generally, one would like to cast a wider net and include blogs that are discussing higher-level concepts related to the market addressed by the product(s) of interest. For example, IBM offers a social networking product called Lotus Connections, but marketing experts may wish to follow all discussions touching on the concept of collaborative software as a means to understand emerging trends in this space.

The distinction between tracking a specific product and tracking a broader concept impacts the methodology used to find the relevant set of blogs. If the interest is only in a specific product, it is straightforward to identify blogs (e.g. by using a blog search engine) containing references to the product. Such an approach is less effective for broad topics because discussions that touch on such a topic (e.g. “collaborative software”) may not specifically contain these keywords. In practice, it is reasonable to ask marketing experts to identify a small set of “seed” blogs that are highly relevant to the topic at hand. One approach is to use these labeled blogs to build a straightforward text classification model to identify other relevant blogs.

Relevant blogs are likely to link to other relevant blogs, and an alternative approach to text classification is to exploit the structure of the blog cross-reference graph. One simple approach is to start with the small set of expert-identified seed blogs, add all the blogs they link to and then repeat this process for several iterations (degrees of separation). This snowball sampling procedure was used to identify the blogs shown in Figure 1; note that the second (and third) iteration of this process identified a number of relevant blogs not included in the seed population.

Discussions about broad concepts like “collaboration software” tend to be tightly connected, and hence this simple approach is likely to be more efficient than keyword search in finding these blog sub-communities. Using the graph structure also alleviates the problem when product search terms have multiple meanings, e.g. “Lotus” is a car, a flower and a software brand – it is unlikely that blogs talking about Lotus the car will reference blogs discussing Lotus Software.

An important consideration is to avoid *crawling* [computer program that browses the Internet in a methodical, automated manner] the parts of the relevant blog sub-universe that are irrelevant from a marketing perspective. A practical solution is focused snowball sampling [3], which explicitly focuses Web

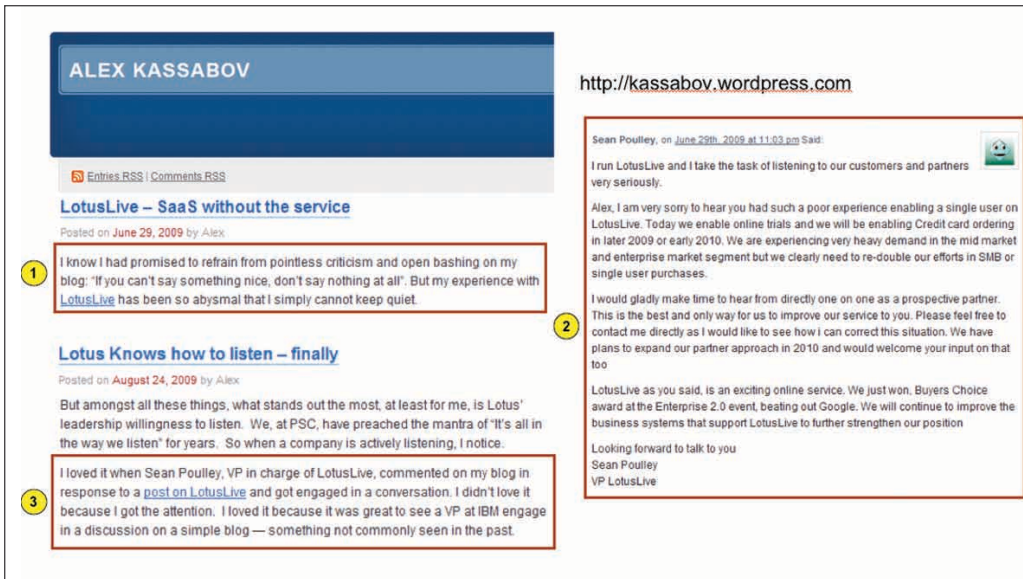


Figure 2: Identifying and addressing negative sentiment.

belief of the sentiment associated with it. It is possible to learn from such labeled words in conjunction with labeled documents. Furthermore, the selection of words and documents to be labeled can be made algorithmically.

Such an approach is known as active dual supervision [5], and it can greatly reduce the effort required to label examples in a new domain. Even though there are expressions of sentiment that are domain-specific, there is still a large amount of overlap in how positive and

crawling using only links deemed to be relevant by a text classifier. There are many other opportunities to apply large-scale versions of classification models [4] that exploit both graph structure and text content.

negative emotion is conveyed across domains. This enables the use of transfer learning to adapt a classifier trained in one domain to a new domain with little to no labeled data in the target domain [6].

**Sentiment Detection**

WHILE THE RELEVANCE MODEL can help to limit the universe to thousands rather than a half million blogs, this is still a volume far beyond close scrutiny by a small marketing staff. But which ones of all the relevant blogs should you read? And are the most relevant indeed the most crucial? An immediate concern from a marketing perspective is to detect strong sentiment, particularly strong negative sentiment. Since it is impossible to read all blog posts relevant to a particular topic, there is strong motivation to develop automated capabilities to characterize the tone and sentiment in these discussions.

**Measuring Influence and Authority**

WHILE RELEVANCE AND SENTIMENT provide two essential filters, it is unlikely that each and every relevant blog with negative sentiment warrants an action. An important consideration is how much does the opinion of one blogger actually matter? A well-known riddle asks, "If a tree falls in a forest and no one is around to hear it, does it make a sound?" This ultimately translates into the question of how influential is the blog in question – is anybody actually listening (or reading) and is it likely that these opinions will influence other individuals?

Sentiment detection models are crucial in order to identify blog posts that may require swift marketing action. Figure 2 illustrates such a situation identified by an IBM social media analytics tool. Sentiment models are able to detect a negative post (1), resulting in a rapid response (2) from a product executive. The quick response leads to a very positive statement (3) from the original blogger. This exchange illustrates the role social media analytics can play in allowing marketing to identify and address negative sentiment before it can cause more brand damage.

Influential bloggers may or may not be factual experts but nevertheless influence the opinions of others via discussions on the topic. From a marketing perspective, it is important to identify this set of bloggers, since any negative sentiment they express could spread far and wide. In addition to authorities, there are bloggers who are very well connected, who are most responsible for the spread of information in the blogosphere. When presented with a large number of posts relevant to a topic, ordering them by the blogger's influence assists in information triage, given that it is not feasible to read all posts. Figure 3 shows such a view, where we have found the most authoritative blogs relevant to the topic "social collaboration."

The main challenge in sentiment classification is that the expression of sentiment tends to be domain specific, and the set of domains to monitor change often. Thus we require sentiment classifiers that can rapidly adapt to new domains without requiring a large number of manually labeled training examples of positive and negative sentiment. Treating sentiment detection as a text classification task has made it possible to adapt to new domains, provided there are enough training examples in the target domain. However, supervision for a sentiment classifier can be provided not only by labeling documents (e.g. blog posts), but also by labeling words. For instance, labeling a word such as "atrocious" as negative is one way to express our prior

Since reliable blog readership information is difficult to obtain, the links between blogs are commonly used instead to determine a blog's authority. For instance, Technorati (www.technorati.com/) assigns an authority score to a blog based on the number of blogs linking to the Web site in the last six months. Similarly, Blogpulse (www.blogpulse.com/) ranks blogs based on the number of times it is cited by other bloggers over the last 30 days. Given that we have a network of directed edges indicating the links between posts/blogs, we can apply more complex measures of prestige from social network analysis. For instance, the author-

ity of a blog can be characterized based on the number and authority of other blogs that link to it, using the well-known PageRank algorithm [7], while the influence of a blog can be captured by the degree to which the blog contributes to the flow of information between other bloggers, determined by Flow Betweenness [8]. Page Rank and Flow Betweenness were used to compute the authority and influence shown in Figure 3. The sentiment shown here was estimated via a transfer-learning model mentioned above.

An interesting question is how to quantitatively measure the utility of different measures of authority and influence. For the purposes of social media marketing, we are interested in bloggers who influence the thinking, and, subsequently, the content blogged by others. If a blogger is indeed influential, we would expect his ideas to propagate to other blogs. Based on this, we propose objectively comparing candidates influence measures on the task of predicting user content generation. If a measure of influence helps you select a blog that more accurately predicts future discussion in the blogosphere than a randomly selected blog, then there is some value to such an influence measure.

### Detecting Emerging Topics

RELEVANCE, AUTHORITY AND SENTIMENT provide a useful way for us to focus attention on the blog posts that we

URL (144)	Relevance	Authority	Y	Influence	Sentiment
http://blogs.newsgator.com/daily/	50	98		97	Slightly Positive
http://www.web-strategist.com/blog	10	98		99	Very Positive
http://enterprise2blog.com	32	96		95	Slightly Positive
http://blogs.zdnet.com/hinchcliffe	55	96		97	Very Positive
http://www.hyperic.com/blog	10	95		51	Slightly Positive
http://beth.typepad.com/beths_blog/	15	94		98	Very Positive
http://pistachioconsulting.com	13	90		90	Slightly Positive
http://www.projectsplaces.com/	13	87		0	Slightly Positive
http://blogs.zdnet.com/collaboration	48	87		95	Very Positive
http://ross.typepad.com/blog/	11	87		88	Slightly Positive
http://blogs.zdnet.com/feeds	15	86		96	Slightly Positive

Figure 3: Finding the most authoritative blogs.

should read. As we read these posts, we naturally synthesize this information to identify important, higher-level concepts and trends that summarize the discussions. This is perhaps the ultimate objective of social media analytics: to automate the human-intensive process of detecting and summarizing patterns that are emerging in the relevant sub-space of the blogosphere. NLP approaches can identify commonly occurring collation of consecutive words like “Barack” and “Obama,” but such occurrences may not be particularly interesting if the phrase is mentioned frequently. Of greater interest are phrases that occur much more frequently in the past day relative to the past week – such an approach is more likely to identify phrases like “healthcare reform” that capture topics that are emerging from the background discussion.

At an even higher level of analysis, document clustering and topic modeling techniques can be used to identify collections of posts expressing cohesive patterns of discussion. Such mod-



## Are you “suite” on INFORMS journals?

Sign up today! Select Pubs OnLine Suite when renewing online, or mark Pubs OnLine Suite on the back of your renewal notice.

Subscribe to all 12 INFORMS Journals in our Pubs OnLine Suite. This is the perfect time to upgrade your personal library.

- Regular member price \$99
- Student member price \$50
- Added subscriber benefit: Access all journal issues back to vol. 1 no. 1

<http://pubsonline.informs.org>



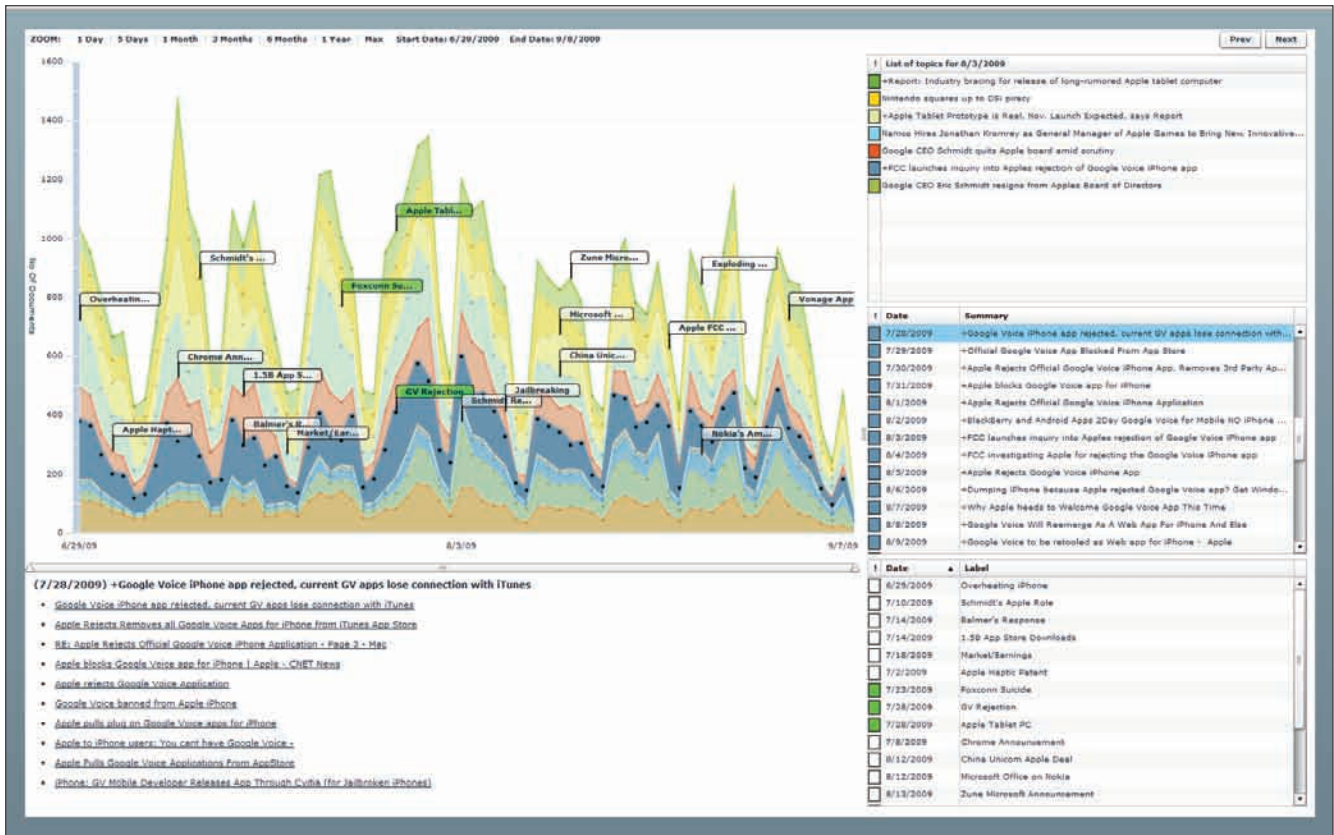


Figure 4: Tracking the iPhone/Google Voice discussion.

els can be extended with notions of temporal continuity to provide a view of how dominant themes evolved over time, possibly also incorporating feedback from a user on which themes to track or discard from the analysis. Figure 4 gives a view of the user interface for interacting with such a model. For each day, the dominant topics can be seen (top right panel). The model exposes the most relevant posts for each topic (bottom panel). The evolving discussion around each topic can be tracked over time (middle right panel) and their relative strengths can be visualized (stacked trend chart). The user can annotate the view as appropriate (white pins in the trend chart) and drive the model towards topics of particular interest (green pins in the trend chart). This particular example identifies and tracks the public discussion following Apple's decision to reject the Google Voice iPhone app in July 2009.

**Conclusion**

THE INITIAL E-COMMERCE PHASE of the Internet in the late 1990s gave rise to a new set of analytics to better understand consumer preferences and hence create targeted marketing campaigns. Similarly, the advent of social media is giving rise to huge opportunities for innovative analytics needed to understand emerging trends and themes in the rapidly expanding social media. From a technical perspective, social media analytics creates even greater challenges than the first wave of internet marketing – in particular, the need to characterize authority and sentiment relative to a specific topic in a network of bloggers gives rise to a number of interesting problems in machine learning. And the demand for this technology is

ramping up quickly as more and more marketing organizations formulate new strategies that cannot be executed without such capabilities. **IORMS**

**Rick Lawrence (ricklawr@us.ibm.com), Prem Melville, Claudia Perlich, Vikas Sindhwani, Steve Meliksetian, Pei-Yun Hsueh and Yan Liu** are machine-learning researchers in the Predictive Modeling Group within the Business Analytics and Mathematical Science organization at the IBM T.J. Watson Research Center in Yorktown Heights, N.Y.

**R E F E R E N C E S**

- 2009, "Power to the people – Social Media Tracker Wave," <http://universalmccann.bitecp.com/wave4/Wave4.pdf>, Universal McCann.
- 2009, Nielsen Global Online Consumer Survey, <http://blog.nielsen.com/nielsenwire/consumer/global-advertising-consumers-trust-real-friends-and-virtual-strangers-the-most/>, The Nielsen Company.
- Melville, P., V. Sindhwani, and R. Lawrence, 2009, "Social Media Analytics: Channeling the Power of the Blogosphere for Marketing Insight," 1st Workshop on Information in Networks, New York.
- Zhang, T., A. Popescul, and B. Dom, 2006, "Linear prediction models with graph regularization for Web-page categorization," *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, Philadelphia: ACM.
- Sindhwani, V., P. Melville, and R. Lawrence, 2009, "Uncertainty sampling and transductive experimental design for active dual supervision," International Conference on Machine Learning (ICML 2009), Montreal.
- Hu, j., Y. Liu, and R. Lawrence, 2009, "Graph-based transfer learning," 18th ACM Conference on Information and Knowledge Management (CIKM 2009), Hong Kong.
- Page, L., 2001, "Method for node ranking in a linked database," U.S. Patent 6,285,999 issued Sept. 4, 2001.
- Newman, M.E.J., 2005, "A measure of betweenness centrality based on random walks," *Social Networks*, Vol. 27, No. 1, pp. 39-54.