



[www.perspectivesweb.com](http://www.perspectivesweb.com)

*MEMO*

February 18, 2010

To: Philip Kegelmeyer, Kristin Glass, Rich Colbaugh, Tim Trucano  
From: Ann Miksovic and Mark Huey  
Subject: Commercial Activities in Web Forecasting

This report gives background information on commercial activities related to prediction and forecasting based on web data. We use the informal term “web data” to mean text collected from a variety of on-line sources with explicit or implicit links between the text units (blogs, discussion forums, Facebook, Twitter, news items, etc.), much of which lacks formal structure. In theory, this information can be seen as capturing the point of view of a given community or population, and this understanding may be helpful in terms of forecasting.

Our initial survey of activities has led us to organize this report into two primary sections.

1. In the first chapter we deal primarily with companies focused on mining online “conversation” to better understand collective opinions. This category is generally referred to as Sentiment Analysis (SA). It is the largest body of activity that we identified as relevant to NGC’s interests in web forecasting, and therefore forms the bulk of this report. The goal of most companies in this category is to add insight that is then used as a factor in decision-making.
2. In the second chapter we identify some activities that go beyond distilling sentiment from web sources and into predicting future events. There is much less information available on this topic. We believe this to be the case mostly because the problem is harder (and distilling sentiment is in itself a challenge), and in part because successful efforts are likely to be extremely valuable. For instance, a system that could reliably predict stock price movement, even for a narrow subset of scenarios, would not likely see much public description (since its trading advantage would depend on exclusivity).

While conducting this research, Perspectives looked for attributes including:

- Use of unstructured web data as the major source for analysis
- The extent to which the work was at a large scale (millions or billions of records)
- Whether there appeared to be novel algorithms or interesting technology underlying marketing claims.
- An element of prediction or forecasting or correlation of information and events.
- Whether explicit linkage was made to counterterrorism and national security problems.

## Table of Contents

<b>I. Sentiment Analysis .....</b>	<b>4</b>
A. Background .....	4
B. Market Size .....	5
C. Overview and Challenges .....	5
D. Contrarian Viewpoints .....	7
E. Looking Forward .....	8
F. Sentiment Analysis for Brand and Reputation Monitoring .....	9
1. <i>Jodange</i> .....	14
2. <i>Crimson Hexagon</i> .....	18
3. <i>Sysomos</i> .....	22
4. <i>NEC Laboratories, America</i> .....	24
5. <i>BlogPulse (a Nielsen Company)</i> .....	26
6. <i>WiseWindow</i> .....	29
7. <i>Visible Technologies</i> .....	31
8. <i>J.D. Power &amp; Associates Web Intelligence</i> .....	33
9. <i>Dow Jones Insight</i> .....	35
10. <i>BuzzLogic</i> .....	36
11. <i>Yahoo! Research</i> .....	38
12. <i>Radian6</i> .....	39
13. <i>Newssift</i> .....	40
14. <i>Sentimine from the Parnassus Group</i> .....	41
15. <i>Nstein</i> .....	42
16. <i>PeopleBrowsr</i> .....	42
G. Sentiment Analysis Related to National Security .....	43
1. <i>SentiMetrix</i> .....	43
2. <i>MITRE Corporation</i> .....	45
3. <i>University of Arizona, AI Lab – Dark Web Terrorism Research</i> .....	48
4. <i>Applied Systems Intelligence</i> .....	50
5. <i>Alias-I, Inc.</i> .....	51
6. <i>Inxight (an SAP company) Federal Systems Group</i> .....	52
H. Other Interesting Sentiment Analysis Activities .....	53
1. <i>Web Ecology Project</i> .....	53
2. <i>CERATOPS (University of Pittsburgh, University of Utah, Cornell University)</i> .....	54
3. <i>Twittermood (Northeastern University)</i> .....	55
4. <i>EU-funded Programs</i> .....	56
I. Intellectual Property and Sentiment Analysis .....	57
<b>II. Forecasting and Prediction .....</b>	<b>59</b>
A. Applications .....	60
1. <i>Stock Market Prediction</i> .....	60
2. <i>Customer Behavior Prediction</i> .....	62
3. <i>Identifying “Influentials”</i> .....	62

4. Predicting Movie Grosses from Blog Traffic .....	66
<b>III. Content Acquisition .....</b>	<b>69</b>
A. Data Collection Companies .....	69
1. Spinn3r .....	69
2. 80legs .....	69
3. Techrigy .....	70
4. Teragram .....	70
B. Dataset Lists .....	70
<b>IV. Appendix .....</b>	<b>72</b>
A. Academic Papers .....	72
1. Multi-lingual .....	72
2. Topology .....	73
3. Topic Modeling .....	73
4. Community Detection .....	74
B. List of Companies in this Report .....	76

*Just as a spider notices strange vibrations in its web, the US Government could recognize changes to the normal state of the web of human connectivity that indicates marshalling of resources for nefarious purposes - before catastrophic events occur - given proper tools and capabilities. In particular, we examine two insights from the field of social network analysis:*

- 1) The innovation necessary to conceive of these rare events originating at the periphery of the terrorist network, and*
- 2) The organization of an event and how it generates activity in regions of the network.*

Source: "Anticipating Rare Events: Can Acts of Terror, Use of Weapons of Mass Destruction or Other High Profile Acts Be Anticipated? A Scientific Perspective on Problems, Pitfalls and Prospective Solutions," white paper, November 2008, available [here](#).

## I. Sentiment Analysis

The rise of social media web sites (Facebook, Twitter, Digg, Delicious, Flickr, blogs, discussion forums, etc.) is creating a trove of publicly-available, opinion-rich information generated by millions of people about millions of different topics (products, companies, politics, movies, etc.). Mining this information using traditional analysis methods is daunting at best: the content is messy (many different formats, many languages, little structure, terse language, etc.), the scale is large, the data never stops flowing, and existing tools and infrastructure are geared toward very different types of analysis. Automating aggregation and analysis of vast amounts of messy data (to extract sentiment or other details) is therefore extremely attractive, even if solutions are only partially effective.

The tally of information on the web includes 15,000 tweets every minute, with over 20% of them noting products or brands; more than 100,000 suggestions, tips and tricks posted to tens of thousands of expert forums daily; more than 110 million blogs; and over 20,000 online mainstream news sites, just to name a few ([source](#)).

### A. Background

[Seth Grimes](#), a long-time observer of the business analytics industry, offers a useful definition of sentiment analysis and some context on how it intersects with traditional commercial business intelligence:

... a set of algorithms and tools for identifying and extracting a) features that express attitudes or opinions, b) attributes that indicate sentiment polarity, intensity, and other characteristics, and c) the topics those sentiments and attributes apply to. Sentiment analysis may further include aggregating sentiment across sources, documents, or document sections and studying trends and correlation of sentiment with events and demographic information ([source](#)).

...  
Sentiment data is often presented in business intelligence (BI) dashboards and other familiar interfaces, but sentiment (and general text) analysis is not yet broadly done in conjunction with conventional BI, with analysis of numerical data drawn from transactional and operational systems. “Unified analysis is coming, but it’s not here yet ([source](#)).”

Perspectives found more than 140 companies engaged in some aspect of sentiment analysis. Many of them are small companies and/or start-ups. Some of these companies are layering data extraction and presentation on top of third party sentiment analysis engines. The sophistication of products varies widely. The typical focus is on a combination of keyword extraction and linguistic analysis to help understand collective opinion. The field has seen some consolidation activity in the last few years (e.g., Umbria by J.D. Power, BuzzMetrics by Nielsen, Cymfony by TNS Media). Some vendors target business users (e.g., assessing corporate reputation or the reaction to a new product launch), either directly or through marketing services partners; a few companies offer services for the general public. Standouts include:

- Startups such as Sentimetrix, Galaxy Advisors, Jodange and Crimson Hexagon.
- Companies with text analytics products (most of these firms incorporate some degree of sentiment analysis) such as Attensity, Clarabridge, Lexalytics, Scout Labs, SPSS and Temis.
- Marketing and branding firms such as Biz360 (which analyzes news media and consumer opinion information from traditional and social media sources to “help businesses make better decisions”), Nielsen BuzzMetrics (“a service for brand metrics, consumer insights and real-time market intelligence”), and TNS Cymfony (“a market influence analytics company”). Additionally, there are a slew of companies using data from Twitter to perform some level of real-time sentiment analysis. These companies offer (as a *New York Times* article gently describes) “lightweight tools.” These tools are sometimes little more than search engines, and are

sometimes referred to as “Twitter apps,” but the area may have a lot of promise. Companies include: Twazzup, Twendz, TipTop Technologies, Tweet Sentiments, and Tweetfeel.

## B. Market Size

We have been unable to locate credible estimates for the size of the sentiment analysis market. But it is not large. Text analytics, which has some overlap but is probably a good deal larger, is estimated at less than \$500M worldwide by one informed observer ([source](#)). We did find an estimate of £60M for Online Reputation Monitoring tools and services in the UK. The US software market might be extrapolated at perhaps 5x this figure. Of course, market size analysis is notoriously imprecise, and estimates are sometimes generous; applying a discount of 50% to figures would not be unreasonable. The conclusion here from the data is that commercial spending on sentiment analysis is more than a trivial sum, but that the sector is still very much in the category of niche software.

## C. Overview and Challenges

The opinions listed below (from respected academic researchers in the field and leading industry experts) are representative of what was found in our research regarding the main challenges of sentiment analysis.

Creating systems that can process subjective data effectively requires, according to Bo Pang of Yahoo! Research and Lillian Lee of Cornell University, “overcoming a number of novel challenges.” In their important 2008 paper, “[Opinion Mining and Sentiment Analysis](#),” they list challenges for review- or opinion-search applications (paraphrased below).

- 1) In general, query classification is a difficult problem - If the application is integrated into a general search engine, one needs to determine whether the user is, in fact, looking for subjective material. This may or may not be a difficult problem, depending on whether the query is in some type of structured format or not (e.g., perhaps the application would provide a checkbox to the user so that he or she could indicate directly that reviews are what is desired).
- 2) Simultaneously or subsequently determining which documents, or portions of documents, contain review-like or opinionated material has its challenges - blogs, for instance, can vary widely in content, style presentation and even level of grammaticality, while sites like Amazon.com have review-oriented text in a relatively stereotyped format from which to draw.
- 3) Once the documents are in hand, the problem then becomes identifying the overall sentiment expressed. Free-form text is much harder for computers to analyze - for example, care must be taken to attribute the views expressed inside quoted text to the correct entity. Conversely, user reviews posted to Yahoo! Movies, for instance, are easier to evaluate because users must specify grades for pre-defined sets of characteristics.
- 4) Lastly, the system will need to present the sentiment information in some kind of summary fashion, whether it be visual or textual and can involve some of the following actions: Aggregation of votes that may be registered on different scales; selective highlighting of some opinions; representation of points of disagreement and points of consensus; identification of communities of opinion holders; and accounting for different levels of authority among opinion holders (full text paper available [here](#)).

Seth Grimes, an expert in business intelligence and decision support systems, sums up additional challenges of sentiment analysis:

The challenge [of sentiment analysis] stems from the huge variability and subtlety of spoken and written language: meaning that humans readily grasp from context is very difficult for computers to detect. How can software reliably discern facts and feelings in light of not only abbreviations, bad spelling, and fractured grammar, but also sarcasm, irony, slang, idiom, and, well,

personality? How is a computer to understand? ... Text analytics can extend reach, lower costs, and improve reaction time in dealing with important enterprise information, including sentiment, that is locked in a variety of forms of human communications. Workers have limited capacity and they're (relatively) expensive, so we use computers for what they're good at: processing large volumes of data fast. Yet accuracy is a serious concern, and there is wide variation in the suitability of various available tools to the task. It is important to know what you can expect in order to create an approach that works given your information sources and goals ([source](#)).

Another problem with analyzing sentiment is **opinion spam**. In his paper, "[Sentiment Analysis and Subjectivity](#)," (full text [here](#)). University of Illinois researcher, Bing Liu says:

... In the context of opinions, we have a similar spam problem. Due to the explosive growth of the user-generated content, it has become a common practice for people to find and to read opinions on the Web for many purposes. For example, a person plans to buy a camera. Most probably, he/she will go to a merchant or review site (e.g., amazon.com) to read the reviews of some cameras. If he/she find that most reviews are positive about a camera, he/she is very likely to buy the camera. However, if most reviews are negative, he/she will almost certainly choose another camera. Positive opinions can result in significant financial gains and/or fames for organizations and individuals. This, unfortunately, also gives good incentives for opinion spam, which refers to human activities (e.g., write spam reviews) that try to deliberately mislead readers or automated opinion mining systems by giving undeserving positive opinions to some target objects in order to promote the objects and/or by giving unjust or false negative opinions to some other objects to damage their reputations. Such opinions are also called fake opinions or bogus opinions. They have become an intense discussion topic in blogs and forums, and also in press.(e.g., <http://travel.nytimes.com/2006/02/07/business/07guides.html>), which show that review spam has become a problem. We can predict that as opinions on the Web are increasingly used in practice by consumers and organizations, the problem of detecting spam opinions will become more and more critical.

The level of **granularity** in sentiment analysis is also important, says author Nancy Kho,

... If sentiment is assigned at a document level - that is, each tweet or blog post is assigned a positive, neutral or negative sentiment - how does the hypothetical tweet "I love Marriott's bathrooms but the beds are lumpy" get classified? Marcel LeBrun, CEO of Radian6 ... cautions, "Ratings need to be assigned on a subject level at a minimum; a solution that assigns them at a document level is going to miss something ([source](#))."

Of sentiment analysis in general, Bing Liu feels that ...

... all the sentiment analysis tasks are very challenging. Our understanding and knowledge of the problem and its solution are still limited. The main reason is that it is a natural language processing task, and natural language processing has no easy problems. Another reason may be due to our popular ways of doing research. We probably relied too much on machine learning algorithms. Some of the most effective machine learning algorithms, e.g., support vector machines and conditional random fields, produce no human understandable results such that although they may achieve improved accuracy, we know little about how and why apart from some superficial knowledge gained in the manual feature engineering process.

A noteworthy article, "[The challenge is still the accuracy of sentiment prediction and solving the associated problems](#)" has Bing Liu commenting on the many specific technical problems to be solved. He outlines automated vs. manual approaches and the challenges of accuracy and its associated problems -- multiple entities in a single document, limited datasets, lack of full-text understanding (available [here](#)).

No single approach today provides the desired accuracy. Thus, in practice, it is important to have novel ideas that can cleverly combine existing methods to accurately classify or predict sentiments and emotions. I expect some innovative techniques to come out in the next few years. As

sentiment analysis is not a single problem, every technical problem needs to be solved satisfactorily. Otherwise, the compound error can be rather bad.

Along the same lines, an analyst with **Forrester Research** says, regarding accuracy:

"The reality is automated sentiment is a bit of a misnomer."... The technology isn't at a place today when automated sentiment is always 100 percent accurate. In fact, Vital says in talking to clients who have deployed some form of sentiment analysis, accuracy rests at about 50 percent. There are ways in which users can increase the accuracy and results of their automated sentiment analysis. For instance, playing around with taxonomies and auto discovery can yield better results. "In the near term," Vital says. "Human intervention will still be necessary."

## D. Contrarian Viewpoints

There is a lot of "buzz" surrounding sentiment analysis, and as with many fields perceived as trendy, there is a subcurrent of backlash against SA. A number of pundits have noted that the obstacles may be more than simply technical hurdles, that sentiment analysis is not merely difficult: it is fundamentally unachievable, or even misguided. The argument here, in the words of various commentators, is that:

- There are millions of manners to express an opinion. Only human brains can process the way people express their views. Because most of the time, digital influencers use images, cultural references, that are not directly explaining a pro or against position, but that more slightly drive the readers to getting the points.
- We don't live in a binary world, in which people would be only pro-something or against-something: in opinion surveys, there are more grads, more details, more levels of understanding. Summarizing social media opinions into this pattern is not only wrong, but is a professional mistake: you give fake overviews, fake trends, fake insights, that can lead to marketing disasters. Moreover if you use "neutral" just to sort all the conversations that don't match your sentiment automatic filtering, you're just wrong! Neutral MEANS something, it does not mean "NOT UNDERSTOOD."
- Sentiment analysis diminishes the richness of social web. If you declare that 20% of online conversations say that this specific product is great, well, you don't tell where the conversations take place, if some of them are more important to consider than others (or more influent), what's the point of contact between all these guys, what makes people very different talk in the same field of interest. The recommendations that you pre-install through this sentiment analysis are not optimal, as it does not cross unexpected data, whereas social media monitoring is all about that: challenging what you already know.
- Sentiment analysis breaks the market credibility: the more you sell wrong studies, the more you destroy a daily work with diverse clients to capture all the issues, all the things necessary to do "good" social media ([source](#))
- We are measuring the WRONG thing. It's about the effect, not the content of the message. What you really want to measure is not whether a message is positive or negative, but what influence it has on the people who read it. We spend so much time worried about the mindset of the vocal few that we ignore how their message really changes the decisions of the many. We need to understand:
  1. The human language is complex, but so are people
  2. A positive plus a negative does not mean neutral
  3. Analysis doesn't consider "degree" of sentiment
  4. Sentiment makes no room for personal authority. The always angry forum troll is not going to have as powerful an effect with their negative sentiment, as the mildly disappointed and very respected super fan.
  5. Sentiment does not indicate action ([source article](#))

- Automated sentiment analysis... [is] a blunt tool at best that's still many years away from fulfilling its potential. I'm not entirely against sentiment analysis-70% accuracy is better than 0%-but I continue to be concerned that businesses are lulled into a false sense of security by it. After all, would you walk into a coal mine with a bird that has a 30% chance of getting it wrong about dangerous gas levels? I know I wouldn't. The most accurate sentiment analysis continues to be of the human variety. Only you can determine-with 100% accuracy-if a blog post about your company is to the benefit or detriment of your reputation. Until technology improves, we'll have to continue acting as our own canaries (source [article](#)).
- Predictive capability may be limited: So the benefit of social media is not so much in its predictive ability -- with this complex environment and consumers' serendipitous reaction to events, predictability is virtually impossible. Social media's benefit is more in its ability to keep the marketer in tune with consumer moods in real-time, or, as I like to say, "at the speed of the market." ([source](#))

## E. Looking Forward

While there are clearly challenges for the field and some of these may ultimately prove insurmountable, at the moment there are many companies investing in developing sentiment analysis technology. Jeffrey Catlin, CEO of **Lexalytics**, notes that "sentiment analysis has come a long way in the last four years. In certain domains, and under certain uses, it's a very dependable technology." He points out that **"If we're right 75 to 80% of the time, we don't care about any single story."** ([source](#)) The field is probably not at that point yet, but may be getting close in certain circumstances. **Nstein**, a text analytics/content management vendor, claims that their nSentiment annotator, "when trained with appropriate corpus, can achieve a precision and recall score between 60% to 70%." According to Seth Grimes, "these are *good* numbers when it comes to attitudinal information." **Attensity's**, marketing VP says that "getting beyond sentiment to actionable information, to 'cause,' is what our customers want. But first, you've got to get sentiment right." ([source](#))

University of Illinois researcher Bing Liu is "optimistic that there will be novel automated techniques coming out in the next few years to make this technology practical for large scale applications." ([source](#))

One important driver for improved sentiment analysis is that it offers a way for search engines to improve their results. This category may ultimately be the major driver for the field. A few observations from respected commentators:

- "Sentiment analysis is no short-term hot trend," Richard MacManus writes in *ReadWriteWeb*. "It will eventually become a key feature of search engines, which will integrate the aggregate sentiment of the crowd into search results." ([source](#))
- "I see sentiment analysis becoming a standard feature of search engines," said [Seth] Grimes, who suggests that such algorithms could begin to influence both general-purpose Web searching and more specialized searches in areas like e-commerce, travel reservations and movie reviews. (*New York Times* [article](#), "Mining the Web for Feelings, Not Facts," August 2009)
- As search engines begin to incorporate more and more opinion data into their results, the distinction between fact and opinion may start blurring to the point where, as David Byrne once put it, "facts all come with points of view." (NYT)
- Bo Pang (Yahoo!) envisions a search engine that fine-tunes results for users based on sentiment. For example, it might influence the ordering of search results for certain kinds of queries like "best hotel in San Antonio." (NYT)

## F. Sentiment Analysis for Brand and Reputation Monitoring

Businesses are using Sentiment Analysis (SA) tools to monitor the “buzz” on the web to help improve brands, products and reputations. Nstein’s VP of Strategy and Client Services, Scot Wheeler, describes that: “reputation can be hurt by unaddressed negative responses to products and service, including poor customer support or any perceived affront to an audience, while a new product announcement or a viral campaign or launch of user groups may create highly positive coverage in influential publications and blogs will help build reputation ([source](#)).”

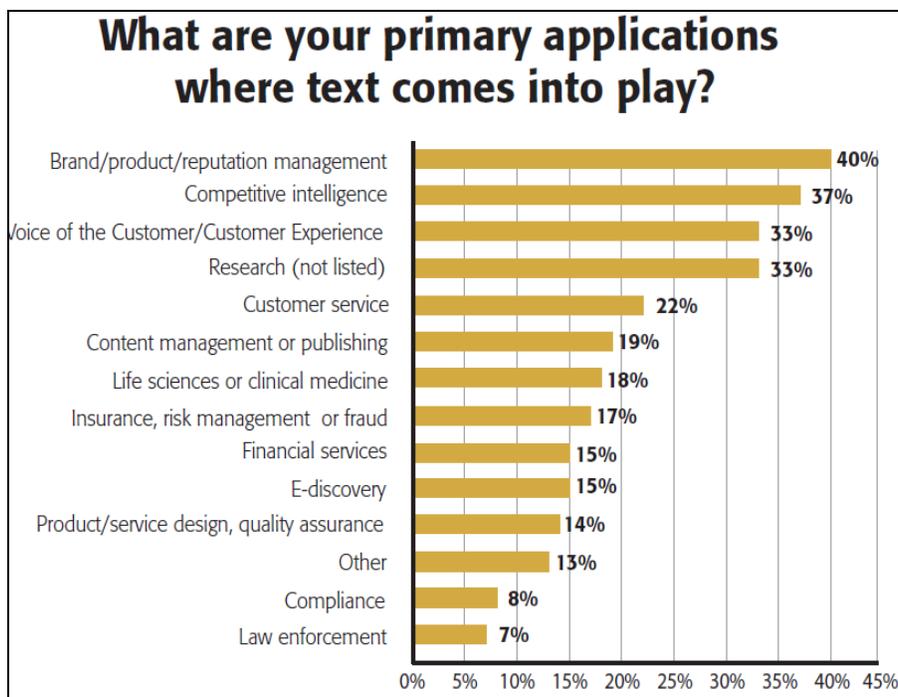
A 2007 study by Jupiter Research found that 30% of frequent social networkers trust their peers’ opinions when making a major purchase decision, compared to the 10% who trust advertisements.

([Source](#))

**Applications and Reputation Management Tools:** One commentator notes that there are a variety of applications for reputation management tools (see [blog](#) posting), including listening platforms; reputation or online reputation management tools; brand defense tools; social media monitoring or buzz tracking software; and consumer generated media tracking. The commentator distinguishes seven types of products and companies:

- Category 1 – Wide scope analytical and reporting tools for all aspects of monitoring customer opinions and campaign effectiveness (Examples: Radian6, Visible Technologies’ Trupulse, Scout Labs, SEER, Nielsen BuzzMetrics, Techrigy/Alterian’s SM2, Feedback Ferret, Autonomy’s Interwoven, ListenLogic)
- Category 2 – Blog based influence assessment tools, designed to gain access to influential customers/commentators (Examples: Nielsen BlogPulse, Market Sentinel Ltd. (UK), BuzzLogic)
- Category 3 – PR and media management tools for reputation management and assessing opinion forming influence (Examples: TNS Cymfony, Marchex Reputation Management [in development], Reputica, Jodange)
- Category 4 – Social media tracking and intervention including free tools (Examples: Converseon, Who’s Talkin, Social Mention, Trackur, ViralHeat, NetBase Consumer Insights)
- Category 5 – Fraud protection, security and threat detection (Examples: Envisional, KnowEm, Reputrace, Mark Monitor, FiltrBox)
- Category 6 – News media tracking (Examples: NewsSift, NewsLive)
- Category 7 – Social media within sales management - for identifying B2B [business to business] prospects (Examples: Inside View and Newstwit)

Below is a 2009 chart put together by Seth Grimes’ company, Alta Plana, showing the breadth of usage for text analytics tools ([source](#)).



**Evaluation of Tools:** Parameters needed when evaluating Online Reputation Monitoring tools include, according to this commentator:

- Efficient filtering of queries based on language, country, source, date, topic
- Depth of coverage
- Real-time monitoring
- Duplicate elimination
- Smart Sentiment analysis (learning with time)
- Ability to modify sentiment analysis
- Sentiment plus (point in time, trend over time, compared with competitor, overlay with another issue etc.)
- Ability to associate timeline with events
- Identification, ranking and monitoring of influencers on multiple networks
- Ranking based on “social popularity” or social engagement (PostRank)
- Comparison to competitive information
- Identification of entity and events within the conversation
- Easy, dynamic ability to chart and graph analysis of queries
- Self-service set up of queries
- Multiple User Management
- Backward trending (and not just for 30 days!)
- Real-time threshold monitoring
- Integration with CRM system
- Analysis services by experts

I tend to put a heavy emphasis on the last point as so far from my experience, monitoring tools are only used as a first step to decipher the whole conversation and identify influencers while the real

value lies in the results of the analysis services offered afterward. Despite great progress in technology, we still need a human with social science skills to make some sense out of all this ([source](#)).

## Forrester Wave Report

In January 2009, Forrester published a vendor assessment [report](#) on the category of vendors with listening platforms. “The Forrester Wave™: Listening Platforms, Q1 2009,” by Suresh Vittal, includes evaluations of seven leading “listening platform vendors” on 62 criteria. According to Forrester, “Listening platforms differ from brand monitoring vendors in one fundamental way: They deliver insights to shape marketing strategy rather than simply tracking metrics.”

The vendors evaluated were Biz360, Dow Jones Insight, J.D. Power & Associates, Nielsen BuzzMetrics, Radian6, TNS Cymfony, and Visible Technologies. Forrester’s overall conclusions were that:

...Nielsen BuzzMetrics and TNS Cymfony established early leadership - thanks to their strong balance of data collection, analytics, and consulting services. Dow Jones Insight, J.D. Power & Associates (JDPA), and Visible Technologies are all Strong Performers: Dow Jones Insight for its strong data coverage, JDPA for text mining and market segmentation capabilities, and Visible Technologies for a strong technology backbone. Biz360 is also a Strong Performer with an innovative product offering - Opinions Insight. This study’s sole Contender - Radian6 - lacks the ability to identify sentiment but offers a solution with an easily customizable user interface tailored for PR teams.

A note on TNS Cymfony: Forrester say that although this company offers a strong technology platform, Orchestra, “the lack of **multilingual** natural language processing (NLP) hinders the expansion of its international presence.”

Below are vendor comparison charts for all seven companies evaluated. Perspectives profiled most of these companies and has added detail from Forrester’s report to individual vendors’ sections in this report.

**Figure 1** Evaluated Vendors: Listening Platforms’ Information And Selection Criteria

Vendor	Number of active customers	Date evaluated
Biz360	More than 50	October 2008
Dow Jones Insight	Did not disclose	October 2008
J.D. Power & Associates	More than 50	October 2008
Nielsen BuzzMetrics	More than 125	October 2008
Radian6	More than 200	October 2008
TNS Cymfony	78	October 2008
Visible Technologies	98	October 2008

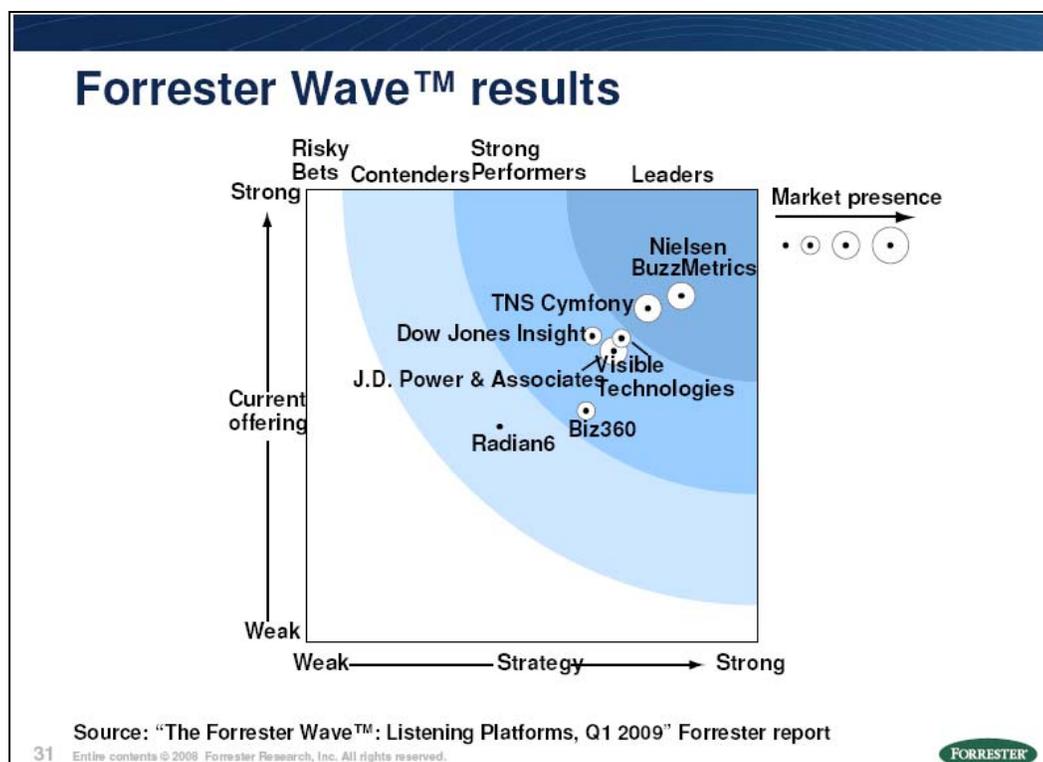
### Vendor qualification criteria

The vendor’s annual revenue has to be more than \$5 million.

The vendor has to have an installed base of at least 50 clients.

At least 75% of the vendor’s clients must be enterprise-level (defined as greater than 999 employees).

Source: Forrester Research, Inc.



**Figure 2** Forrester Wave™: Listening Platforms, Q1 '09 (Cont.)

	Forrester's Weighting	Biz360	Dow Jones Insight	J.D. Power & Associates	Nielsen BuzzMetrics	Radian6	TNS Cymfony	Visible Technologies
<b>CURRENT OFFERING</b>	50%	2.54	3.37	3.20	3.80	2.37	3.67	3.32
Background information	0%	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Data sources	25%	2.84	3.85	1.82	3.07	2.63	3.75	2.52
Text analytics	25%	3.00	3.70	4.40	4.40	1.00	4.10	4.10
Functionality	25%	2.71	3.21	3.57	3.94	3.73	3.42	3.87
Consulting and analysis services	25%	1.60	2.70	3.00	3.80	2.10	3.40	2.80
<b>STRATEGY</b>	50%	3.10	3.16	3.40	4.16	2.16	3.78	3.48
Strength of management team	30%	2.00	3.00	4.00	5.00	2.00	4.00	3.00
Corporate strategy	20%	4.20	3.00	4.00	4.40	4.20	4.00	4.00
Product strategy	40%	3.40	3.40	3.00	3.70	0.80	3.70	3.70
Cost	10%	3.00	3.00	2.00	3.00	4.00	3.00	3.00
<b>MARKET PRESENCE</b>	0%	2.44	2.70	3.94	3.92	1.68	3.16	2.80
Customers	40%	3.35	2.75	3.60	3.55	3.65	3.10	2.85
Financials	40%	1.60	3.10	4.40	4.00	0.45	3.10	2.40
Employees	20%	2.30	1.80	3.70	4.50	0.20	3.40	3.50

All scores are based on a scale of 0 (weak) to 5 (strong).

Source: Forrester Research, Inc.

(Source: "The Forrester Wave™: Listening Platforms, Q1 2009")

**A note on companies that provide sentiment analysis:** Lexalytics' VP of Marketing commented in a blog posting (see [here](#)) that the SA field is actually not expanding in terms of providers of core analytic engines (emphasis from Perspectives):

We are absolutely seeing an increase in solutions and platforms interested in integrating our sentiment analysis, but there are still only a handful of providers (us being one of them, obviously) who can offer the technology. Many of the solutions profiled on the market do a fantastic job at gathering and presenting the results, which is critical when you think about how much information floating around online is written by YOUR customers. But that doesn't necessarily mean that there are more sentiment analysis providers out there - the core analysis engines traditionally sit behind the scenes and do their thing and there are still only a few of us on the market.

Perspectives followed up to ask if Lexalytics could provide names of these providers. In a 1/6/10 email, a company representative responded "We have offered the capability as part of our core text analytics software for years now. There are several vendors that are just getting into the game because "reputation management" and social media are hot topics right now and others, like us, that have been working on it for a while," and attached a third party [report](#) that cites Lexalytics' competitors as: SAP BusinessObjects Text Analysis, SAS' Teragram, Clarabridge, Attensity Group, Nstein and TEMIS.

The remainder of this section of the report focuses on some visuals and case studies using sentiment analysis. It is a sampling of things that caught our eye in a very large set of products and services. Where available, we have listed papers and publications, intellectual property, pricing, multilingual capabilities, size, evaluations from industry experts, accuracy claims, news items and more.

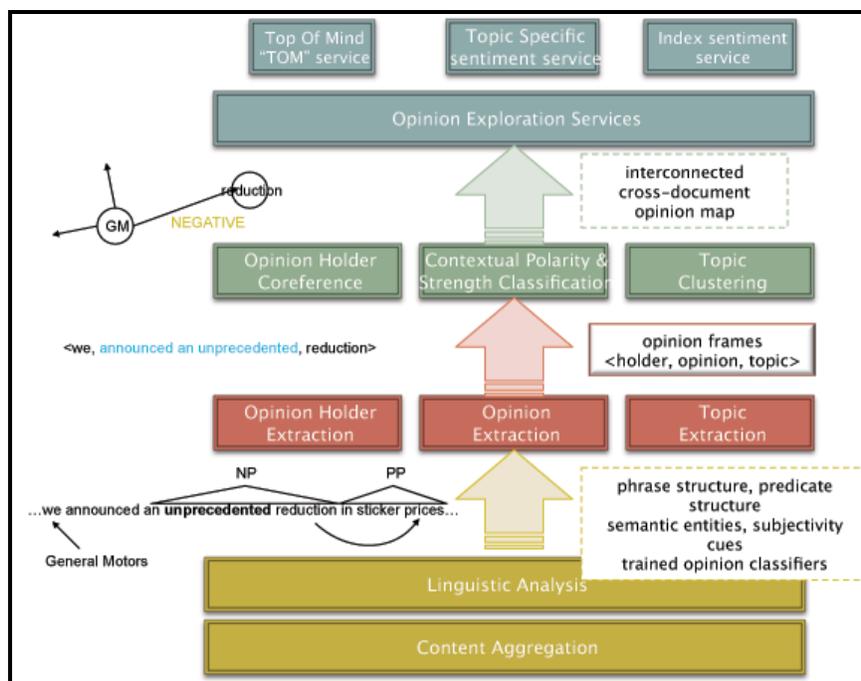
## 1. Jodange

“Predicting outcomes and planning strategy through deeper understanding of opinion holders' sentiment over time.”

**About:** Jodange’s technology is described in this *Communications of the ACM* [article](#) (April 2009):

Jodange’s sentiment analysis software grew out of a research project by Claire Cardie at Cornell University and Jan Wiebe at the University of Pittsburgh. Drawing on a body of theory in linguistics, philosophy, and computational linguistics, their team developed an algorithm that tries to determine the context of any particular statement by isolating three key data points: the topic, the opinion holder, and the opinion itself. First, the algorithm employs an entity extraction routine that locates keywords to identify particular topics and opinion holders. Next, it layers that data onto a linguistic analysis of the opinion being expressed. The resulting unit of data is a triple consisting of opinion, opinion holder, and topic. These triples are then stored in a relational database, where they can be cross-referenced across multiple documents to create what Jodange vice president of product management and marketing Pia Chong calls a “walled garden of opinion.” By connecting opinions from multiple sources about a particular topic, the application can provide users with a bird’s-eye view of a particular topic presented in a variety of different formats: straightforward lists, heat maps that show the concentration of opinions on particular topics, an opinion index that calculates positive or negative trends, or a so-called Doppler view that shows a graphical summary of opinion data. The company is currently working on a new predictive model that could use opinion data to predict future developments, such as the impact of written opinion on trends in a company’s stock price.

Jodange offers a number of tools for sentiment analysis. Jodange describes its technology and process as follows:



Jodange's Top of Mind™ product:

... goes beyond keywords and objective data to uniquely isolate and track people's opinions and sentiments about key topics over time. Imagine the ability to tap into the conversations of 40 experts on a topic of interest, find out when their opinions change and have the forensic ability to isolate their specific remarks, and to also track the change in sentiment relative to a specific market outcome. Top of Mind is an indispensable suite of opinion discovery tools empowering you to:

- Analyze who is saying what about the topics most relevant to you.
- Learn what opinion leaders and early adopters are thinking collectively or independently.
- Make better decisions by better understanding the key influencers driving your marketplace.

Below are several screenshots for the Top of Mind product.

The screenshot displays the Jodange Top of Mind Service interface. At the top, the Jodange logo and 'Top of Mind Service' are visible, along with a 'Welcome, demo | Log out' link. The main heading is 'Home'. Below this, a search bar contains the text 'alternative energy' and a 'Search' button. A search instruction box explains: 'Enter terms to search by Company Name, Topic, Opinion Holder, or keyword. If you want to search for a specific phrase: "place it in quotes". To include a specific term, add a plus sign: +term. To exclude a specific term, add a minus sign: -term.' The search results are organized into three columns: 'Opinion Holder', 'Topic', and 'Category'. Under 'Opinion Holder', there are three entries: 'California Alternative Energy and Advanced Transportation Financing Authority (4)', 'The Alternative Energy Technology Centre Inc (2)', and 'The Alternative Energy Technology Center (7)'. Under 'Topic', there are two entries: 'Alternative Energy Sources Inc (6)' and 'The Alternative Energy Technology Centre Inc (2)'. The 'Category' column shows 'No Categories found.' Below the search results, there is a 'Saved Alerts' section with an 'edit' link and one alert for 'Apple Computer'. The 'Opinion Stream' section is described as 'a personalized list of the most recent opinions based on your saved Alerts and other general published news'. It features two columns of opinion snippets. The left snippet is about Barack Obama and migration, and the right snippet is about Barack Obama's acceptance speech. A 'Media Coverage' badge with the number '2' is positioned to the right of the second snippet.



**Papers** (A list of Jodange publications with these and older papers available [here](#)):

- Yejin Choi and Claire Cardie (2008). “[Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis.](#)” [Full text [here](#)] Empirical Methods in Natural Language Processing (EMNLP).

Excerpted abstract: In this paper, we consider the task of determining the -bearing expression, considering the effect of interactions among words or constituents in light of compositional semantics. We presented a novel learning-based approach that incorporates structural inference motivated by compositional semantics into the learning procedure. Our approach can be considered as a small step toward bridging the gap between computational semantics and machine learning methods. [Worked supported by NSF and DHS.]
- Eric Breck, Yejin Choi, and Claire Cardie (2007). “[Identifying Expressions of Opinion in Context.](#)” Twentieth International Joint Conference on Artificial Intelligence (IJCAI). [Full text [here](#)]

Excerpted abstract: While traditional information extraction systems have been built to answer questions about facts, subjective information extraction systems will answer questions about feelings and opinions. A crucial step towards this goal is identifying the words and phrases that express opinions in text. ... We present an approach for identifying opinion expressions that uses conditional random fields and we evaluate the approach at the expression-level using a standard sentiment corpus. Our approach achieves expression-level performance within 5% of the human interannotator agreement. ... When predicting, a sequence of consecutive tokens tagged as In constitutes a single predicted entity.
- Yejin Choi, Eric Breck, and Claire Cardie (2006). “[Joint Extraction of Entities and Relations for Opinion Recognition.](#)” Proceedings of Empirical Methods in Natural Language Processing (EMNLP). [Full text [here](#)]

Excerpted abstract: We present an approach for the joint extraction of entities and relations in the context of opinion recognition and analysis. We identify two types of opinion-related entities – expressions of opinions and sources of opinions—along with the linking relation that exists between them. [supported by ARDA, NSF, Google and Xerox.]
- Veselin Stoyanov and Claire Cardie (2006). “[Partially Supervised Coreference Resolution for Opinion Summarization through Structured Rule Learning.](#)” Proceedings of Empirical Methods in Natural Language Processing (EMNLP). [Full text [here](#)]

Excerpted abstract: Combining fine-grained opinion information to produce opinion summaries is important for sentiment analysis applications. Toward that end, we tackle the problem of source coreference resolution - linking together source mentions that refer to the same entity. The partially supervised nature of the problem leads us to define and approach it as the novel problem of partially supervised clustering. We propose and evaluate a new algorithm for the task of source coreference resolution that outperforms competitive baselines. [supported by ARDA, NSF, Google and Xerox.]
- Veselin Stoyanov and Claire Cardie (2006). “[Toward Opinion Summarization: Linking the Sources.](#) COLING-ACL 2006 Workshop on Sentiment and Subjectivity in Text.” [Full text [here](#)]

Excerpted abstract: In this paper we describe how source coreference resolution can be transformed into standard noun phrase coreference resolution, apply a state-of-the-art coreference resolution approach to the transformed data, and evaluate on an available corpus of manually annotated opinions.

- Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan (2005). "Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns." [Full text [here](#)] Proceedings of HLT-EMNLP.

Excerpted abstract: We pursue another aspect of opinion analysis: identifying the sources of opinions, emotions, and sentiments. We view this problem as an information extraction task and adopt a hybrid approach that combines Conditional Random Fields (Lafferty et al., 2001) and a variation of AutoSlog. ... The resulting system identifies opinion sources with 79:3% precision and 59:5% recall using a head noun matching measure, and 81:2% precision and 60:6% recall using an overlap measure. [Funded by NSF and ARDA]

## 2. *Crimson Hexagon*

**About:** Founded in 2007, [Crimson Hexagon](#)'s patent pending technology is based on work conducted at [Harvard University's Institute for Quantitative Social Science](#).

Crimson Hexagon's "brand monitoring" technology, based on work by **Gary King of Harvard University** analyzes the social internet (blog posts, forum messages, Tweets, etc.) by "identifying statistical patterns in the words used to express opinions on different topics. It uses these patterns to calculate the percentage of opinion for each opinion category, as defined by the business user client. Competitive technologies, by contrast, simply count the number of mentions of different keywords or infer generic positive/negative sentiment." The product offers continuous opinion monitoring and is already in use. They report: "Advanced online opinion monitoring proposes a modern division of labor: humans provide the up-front intelligence, and machines take over for the high-volume execution."

**Product:** The company describes its three products below:

[VoxTrot](#) listening platform provides companies with actionable insight into consumer opinion of their brand, product, or market. VoxTrot technology can identify opinion from large quantities of text, whether it's an in-house content repository or the vast blogosphere.

Today, VoxTrot users access a private, customized dashboard providing two views: VoxTrot Buzz shows the volume of mentions and positive/negative sentiment on topics of your choosing, across the internet and social media sources. VoxTrot Buzz also provides our proprietary VoxScore metric, evaluating your topic's perception across social media, based on number of mentions, their credibility, and reach.

VoxTrot Opinion helps enterprise customers drill deeper into relevant opinion, by letting companies explore:

- Why is my sentiment trending up or down?
- Is my messaging resonating with the right audiences?
- What are the dominant opinions about specific product features?
- What do people like most and least about competitive products?

The VoxTrot listening platform keeps your company connected to the voice of the customer using an innovative, automated approach – avoiding the high cost and risks associated with periodic manual reporting.

The underlying research for VoxTrot is described in this [paper](#).



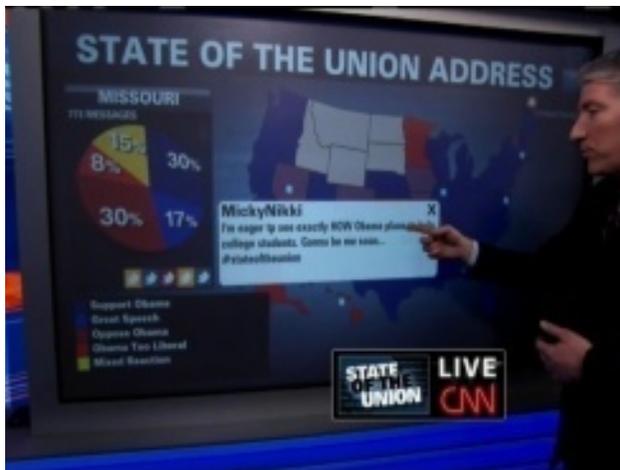
**Sentiment accuracy claim:** “Most companies claim 75-80% accuracy with NLP, Crimson claims **97% accuracy with its algorithm** ([source](#)).” In a text of CH’s system, this [source](#) said: “The accuracy and reliability of Crimson Hexagon is truly astounding. Equally remarkable is the fact that the technology developed by Gary King’s group parses every word in a given text.”

**Multilingual capabilities:** The claim is that “...the technology can parse any language, say Farsi, just as long as the sample blogs provided are in Farsi ([source](#))”. This is because the algorithms are language-independent. Says one reviewer of the technology, “You train the system in a specific language and it performs its magic against the data set regardless of the source language... Information is not translated on the fly, however. If you have trained a monitor in Hebrew it will not find mentions in English ([source](#)).”

**Cost:** The core platform costs \$25K per year, plus charges for each monitor. Its Buzz product is free, however, and its basic keyword search capabilities “seem solid,” according to this [article](#).

**Crimson Hexagon’s technology in the news:** In a recent article, the *Huffington Post* described how CNN’s John King used Crimson Hexagon’s software technology in the “Magic Wall” to analyze almost 150,000 Twitter responses to President Obama’s State of the Union speech. King drilled down to show state-by-state reactions and highlighted sample tweets from each state, as well as showed a macro-view of Twitter users’ responses to the speech ([article and video here](#)).

Melyssa Plunkett-Gomez, an executive with Crimson Hexagon... added that recent analysis the company did on the public opinion and Afghanistan came within two percentage points of a CBS poll on the same topics. She said the company hopes to be able to drill-down even further in future analyses.



**Employees:** 10

**Investors:** Crimson Hexagon is backed by Golden Seeds, Beacon Angels, and the Angel Investor Forum in Connecticut.

**Other:** Crimson Hexagon has an exclusive licensing agreement with Harvard University's Office of Technology Development for the technology ([source](#)); King's academic [website](#) has a wealth of content of interest here.

A review of Crimson Hexagon's work is covered in Patrick Meier's [blog](#) on "Conflict Early Warning and Early Response." Meier concludes:

Crimson Hexagon is truly pioneering a fundamental shift in the paradigm of textual analysis. Instead of trying to find the needle in the haystack as it were, the technology seeks to characterize the hay stack with astonishing reliability such that any changes in the hay stack (amount of hay, density, structure) can be immediately picked up by the parser in real time.

### More on Gary King

Gary King of Harvard University has been working in the field of International Relations for some time, and one key thrust of this work is rare events and web data. King's approach to automated content analysis, which he describes as more interested in characterizing the haystack than finding the needle in it," has been recently published, as well as an approach for rare events prediction and a dataset on international events:

- "A Method of Automated Nonparametric Content Analysis for Social Science", Daniel J. Hopkins and Gary King. *American Journal of Political Science* 54, 1 (January 2010): 229–247 [[Full text paper](#)]

The increasing availability of digitized text presents enormous opportunities for social scientists. Yet hand coding many blogs, speeches, government records, newspapers, or other sources of unstructured text is infeasible. Although computer scientists have methods for automated content analysis, most are optimized to classify individual documents, whereas social scientists instead want generalizations about the population of documents, such as the proportion in a given category. Unfortunately, even a method with a high percent of individual documents correctly classified can be hugely biased when estimating category proportions. By directly optimizing for this social science goal, we develop a method that gives approximately unbiased estimates of category proportions even when the optimal classifier performs poorly. We illustrate with diverse data sets, including the daily expressed opinions of thousands of people about the U.S. presidency. We also make available software that implements our methods and large corpora of text for further analysis.

- “Methods to evaluate automated information extraction systems when coding rare events, the success of one such system, along with considerable data.” Gary King and Will Lowe. An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design, *International Organization*, Vol. 57, No. 03 (July, 2003): pp. 617-642. [[Full text paper](#)]

Despite widespread recognition that aggregated summary statistics on international conflict and cooperation miss most of the complex interactions among nations, the vast majority of scholars continue to employ annual, quarterly, or occasionally monthly observations. Daily events data, coded from some of the huge volume of news stories produced by journalists, have not been used much for the last two decades. We offer some reason to change this practice, which we feel should lead to considerably increased use of these data. We address advances in event categorization schemes and software programs that automatically produce data by “reading” news stories without human coders. We design a method that makes it feasible for the first time to evaluate these programs when they are applied in areas with the particular characteristics of international conflict and cooperation data, namely event categories with highly unequal prevalences, and where rare events (such as highly conflictual actions) are of special interest. We use this rare events design to evaluate one existing program, and find it to be as good as trained human coders, but obviously far less expensive to use. For large scale data collections, the program dominates human coding. Our new evaluative method should be of use in international relations, as well as more generally in the field of computational linguistics, for evaluating other automated information extraction tools. We believe that the data created by programs similar to the one we evaluated should see dramatically increased use in international relations research. To facilitate this process, we are releasing with this article data on 4.3 million international events, covering the entire world for the last decade.

**DATA: 10 Million International Dyadic Events**, conflict and cooperation in international relations, 1990-2004, as evaluated by King and Lowe (2003), automatically coded from Reuters news reports. (Website: [Events](#) | Abstract: [HTML](#))

King’s work is probably of interest to the NGC team, and some additional information on his research group’s work is supplied below:

Gary King is the Albert J. Weatherhead III University Professor at Harvard University and also serves as Director of the [Institute for Quantitative Social Science](#). King and his research group develop and apply empirical methods in many areas of social science research. *His work is categorized here by type and research area: [descriptions of each category are on the [source page](#); only selected descriptions are shown here – links lead to extensive lists of research papers and presentations]*

- [Rare Events](#): How to save 99% of your data collection costs; bias corrections for logistic regression in estimating probabilities and causal effects in rare events data; estimating base probabilities or any quantity from case-control data; automated coding of events.
- [Content Analysis](#): Automated methods of extracting information from text documents and for correcting for errors by human coders.
- [Ecological Inference](#) (Inferring Individual Behavior from Group-Level Data): The original methods that incorporate both unit-level deterministic bounds and cross-unit statistical information, methods for 2x2 and larger tables, Bayesian model averaging, applications to elections, EI/Ezl software.
- [International Conflict](#): Evidence that the causes of conflict, theorized to be important but often found to be small or ephemeral, are indeed tiny for the vast majority of dyads, but they are large, stable, and replicable wherever the ex ante probability of conflict is large. Also methods for coding, analyzing, and forecasting international conflict and state failure.

### 3. Sysomos

[Sysomos](#) emerged from an advanced research project started in 2005 at the University of Toronto. The company was founded by Nilesh Bansal and Nick Koudas in September 2007.

The ten-person company is led by Nick Koudas, Ph.D. an “authority in data analytics” with more than 20 patents, and more than 100 research publications in the areas of database systems, text analytics, and information mining. Koudas is also a professor of computer science at the University of Toronto.

**Current investors:** GrowthWorks Capital and Ontario Centres of Excellence ([Corporate fact sheet](#))

Sysomos mines information from social media sites such as Twitter and Facebook and close to 30 million blogs. Findings are **graphed by time, sentiment** and demographics to provide a full picture of buzz around products brands, public figures or issues ([source](#)).

Their **patent-pending core technology** consists of:

- Real-time content aggregation
- A four-step spam filtering process
- An accurate sentiment engine
- Robust geo-demographic information
- Sophisticated text analytics and language processing algorithms

#### API's:

**Data API:** Bring our social media data archive to your product. We aggregate, clean and maintain over five million conversations collected from an extensive number of social media services every day. Our APIs provide access to this vast repository of content in realtime. With each piece of data we serve, we optionally include complete authority, influence, language, and geo-demographics information.

**Charts and Trends API:** We export aggregate statistics for popularity, mentions, geo-demographics and share-of-voice for any query across our data archive via our charts and trends APIs.

**Text Analytics API:** Use our analytics functionality on your internal proprietary text databases (beyond just social media content). Create automated text summaries or extract key conversations from your data. Use our APIs to display beautiful and informative visual summaries within your software application.

**Sentiment API:** We export our sentiment analytics engine for use with your application and any text content (beyond just social media content) to access tone. Our patent-pending sentiment technology is built and optimized over the course of three years, providing you access to robust and accurate automated sentiment analytics.

**Query-by-Doc API:** Sysomos' Query-by-Doc functionality provides a unique capability to links multiple different media sources together. The API allows you to seamlessly integrate the content you own with the social media buzz. The system will automatically extract reactions and commentary related to any input text document, be it a press release, news article, email message or an internal report. ([API page](#))

#### Publications:

The following are recent articles and papers written by members of Sysomos' technical staff, research collaborators at the University of Toronto, and other research institutions:

- Michael Mathioudakis, Nick Koudas (2010). Demo proposal, Website under construction, “[TwitterMonitor: Trend Detection over the Twitter Stream](#)” [Full text [here](#)]
- Manos Papagelis, Nilesh Bansal, Nick Koudas (2009). “[Information Cascades in the Blogosphere: A Look Behind the Curtain,](#)” To appear in the 3rd International AAAI Conference on Weblogs and Social Media, ICWSM 2009. San Jose, California USA, May 17-20, 2009 [Full text [here](#)]

- Michael Mathioudakis, Nick Koudas (2009). “Efficient Identification of Starters and Followers in Social Media,” In: 12th International Conference on Extending Database Technology, EDBT 2009, St. Petersburg, Russia, Mar 23-26, 2009 [Full text [here](#)]
- Yin Yang, Nilesch Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas, Dimitris Papadias (2009). “Query by Document,” To appear in Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, WSDM 2009, Barcelona, Spain, Feb 9-12 2009. [Full text [here](#)]
- Nilesch Bansal, Sudipto Guha, Nick Koudas (2008). “Ad-Hoc Aggregations of Ranked Lists in the Presence of Hierarchies,” In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, Canada, June 9-12 2008. [Full text [here](#); slides [here](#)]
- Nilesch Bansal, Fei Chiang, Nick Koudas, Frank Wm. Tompa (2007). “Seeking Stable Clusters in the Blogosphere,” In: Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB 2007, Vienna, Austria, Sept 23-28 2007. [Full text [here](#); slides [here](#)]
- Nilesch Bansal, Nick Koudas (2007). “BlogScope: A System for Online Analysis of High Volume Text Streams,” In *Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB 2007*, Vienna, Austria, Sept 23-28 2007, Demonstration Proposal. [Full text [here](#)]
- Nilesch Bansal, Nick Koudas (2007). “Searching the Blogosphere,” In Proceedings of the 10th international Workshop on Web and Databases, WebDB 2007, [Full text [here](#); slides [here](#)] (co-located with SIGMOD) Beijing, China, June 15 2007.
- Nilesch Bansal, Nick Koudas (2007). “BlogScope: Spatio-temporal Analysis of the Blogosphere,” In Proceedings of the 16th international conference on World Wide Web, WWW 2007, Banff, Canada, May 8-12, 2007, Poster. [Full text [here](#)]

Reports and whitepapers page [here](#). Interesting stats on Twitter in this [report](#), “Inside Twitter: An In-Depth Look at the 5% of Most Active Users.”

## Intellectual Property

Title: Method and system for information discovery and text analysis

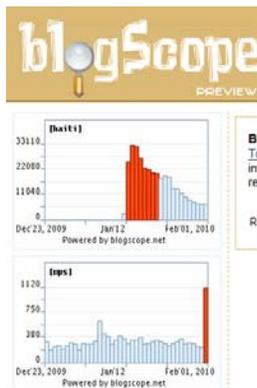
Publication #: [20090319518](#)

Publication date: Dec 24, 2009

Inventors: Koudas, Nick; Bansal, Nilesch

A method for searching text sources including temporally-ordered data objects, such as a blog, is provided including the steps of: (i) providing access to text sources, each text source including temporally-ordered data objects; (ii) obtaining or generating a search query based on terms and time intervals; (iii) obtaining or generating time data associated with the data objects; (iv) identifying data objects based on the search query; and (v) generating popularity curves based on the frequency of data objects corresponding to one or more of the search terms in the one or more time intervals. A system and computer program for text source searching is also provided.

## Sysomos / University of Toronto Tool: BlogScope



BlogScope is an analysis and visualization tool for blogosphere which is being developed as part of a research project at the University of Toronto. It is currently tracking over 41.44 million blogs with 1149.09 million posts. BlogScope can assist the user in discovering interesting information from these millions of blogs via a set of numerous unique features including popularity curves, identification of information bursts, related terms, and geographical search.

Marshall Sponder of WebMetricsGuru, an industry blog about web analytics and related topics, conducted a sentiment analysis [product comparison](#) of several social media monitoring platforms (including Radian6, Techrigy, Biz360, Brandwatch). His conclusions regarding the Sysomos product:

... the sentiment results were generally good... Sysomos, the more I work with it, the more I like - Sysomos was built by a programmer and a university- the University of Toronto - and their backend is able to splice and dice data very well - Sysomos is more like a programming think tank that grew into marketing - yes, their interface could improve - but the sentiment analysis and noise suppression are excellent.

### 4. NEC Laboratories, America

[NEC Laboratories America](#) (Information Analysis Department ([webpage](#)), Cupertino, CA) is the US-based part of NEC's global network of research laboratories. Their primary focus is on technology research and early market validation in support of NEC's core businesses globally and in the US. They have a staff of about 120 employees, with locations in Princeton, New Jersey and Cupertino, California.

NEC researchers are active at academic conferences and appear to be doing work of relevance to NGC, but we have been unable to identify any commercial activities based on the research.

#### About:

...Long-tail data also contain a lot of valuable information, such as grassroots opinions, sentiments, wisdom of crowds, etc. Such information is more valuable for the purposes of business intelligence, better decision making, and market strategies. We need new technologies to dig out such valuation information from the long tail data.

In our Cupertino office, we strive to develop cutting edge technologies to dramatically improve humans' abilities to:

- Sift through large volumes of raw video streams and noisy, low-quality Internet data to extract highly semantic, value-added information
- Summarize and visualize large volumes of Internet data to discover their overall pictures and internal structures

By developing these technologies, we want to turn raw video streams into valuable information sources, noisy, low-quality Internet data into knowledge-bases. Such technologies have great potentials in a variety of areas such as intelligent video surveillance, customer attributes and shopping behavior recognitions for retail business, targeted advertisements, grassroots wisdom discovery, market analysis, business intelligence, etc.

**Projects:** [Internet OLAP](#) (iOLAP); Paper abstract about iOLAP [here](#).

...We are developing the Internet OLAP technologies that aim to sift through millions of websites and blogs to recognize/visualize the overall picture of the underlying internet data set, and to discover value-added information such as the constituent communities, the hot topics and their evolutions, the influential people, etc. Such technologies will enable us to leap beyond the keyword-based search to a deeper understanding of semantics and structures of internet data sets, and will have huge potential for grassroots knowledge acquisition, sentiment discovery, market analysis, business strategy planning, etc.

**Intellectual Property:** Patent page [here](#) (most, if not all, of these patents topics are only relevant to NEC's other technologies)

**Publications** page [here](#).

- Yu-Ru Lin (Arizona State University), Yun Chi, Shenghuo Zhu (NEC Labs), Hari Sundaram (Arizona State University), and Belle L. Tseng (Yahoo! Research) (2009). "**Analyzing communities and their evolutions in dynamic social networks.**" *ACM Trans. Knowl. Discov. Data*, 3(2):1-31, [Full text [here](#)]

We discover communities from social network data and analyze the community evolution. These communities are inherent characteristics of human interaction in online social networks, as well as paper citation networks. Also, communities may evolve over time, due to changes to individuals' roles and social status in the network as well as changes to individuals' research interests. We present an innovative algorithm that deviates from the traditional two-step approach to analyze community evolutions. In the traditional approach, communities are first detected for each time slice, and then compared to determine correspondences. We argue that this approach is inappropriate in applications with noisy data. In this paper, we propose FacetNet for analyzing communities and their evolutions through a robust unified process. This novel framework will discover communities and capture their evolution with temporal smoothness given by historic community structures...

- Yun Chi, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, Belle L. Tseng (NEC Laboratories America) (2007), "**Structural and temporal analysis of the blogosphere through community factorization,**" In: International Conference on Knowledge Discovery and Data Mining archive, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, California, USA. [[Abstract](#)]

The blogosphere has unique structural and temporal properties since blogs are typically used as communication media among human individuals. In this paper, we propose a novel technique that captures the structure and temporal dynamics of blog communities. In our framework, a community is a set of blogs that communicate with each other triggered by some events (such as a news article). The community is represented by its structure and temporal dynamics: a community graph indicates how often one blog communicates with another, and a community intensity indicates the activity level of the community that varies over time. Our method, community factorization, extracts such communities from the blogosphere, where the communication among blogs is observed as a set of subgraphs (i.e., threads of discussion). This community extraction is formulated as a factorization problem in the framework of constrained optimization, in which the objective is to best explain the observed interactions in the blogosphere over time. We further provide a scalable algorithm for computing solutions to the constrained optimization problems. Extensive experimental studies on both synthetic and real blog data demonstrate that our technique is able to discover meaningful communities that are not detectable by traditional methods.

- Yun Chi, Belle L. Tseng, Junichi Tatemura, (NEC Labs). “**Eigen-trend: trend analysis in the blogosphere based on singular value decompositions.**” [Full text [here](#)]  
 Excerpted abstract: ...In this paper, we introduce a novel concept, coined eigen-trend, to represent the temporal trend in a group of blogs with common interests and propose two new techniques for extracting eigen-trends in blogs. First, we propose a trend analysis technique based on the singular value decomposition. Extracted eigen-trends provide new insights into multiple trends on the same keyword. Second, we propose another trend analysis technique based on a higher-order singular value decomposition. This analyzes the blogosphere as a dynamic graph structure and extracts eigen-trends that reflect the structural changes of the blogosphere over time. ...
- Yu-Ru Lin Hari Sundaram Yun Chi Jun Tatemura Belle Tseng, (NEC Laboratories America Arizona State University), “**Discovery of Blog Communities based on Mutual Awareness,**” WWW 2006 Workshop. [Full text [here](#)]  
 Excerpted abstract: ... We focus on extracting communities based on two key insights - (a) communities form due to individual blogger actions that are mutually observable; (b) semantics of the hyperlink structure are different from traditional web analysis problems.
- Arun Qamra, Belle Tseng , Edward Y. Chang (UCSB and NEC Labs), “**Mining blog stories using community-based and temporal clustering,**” Proceedings of the 15th ACM international conference on Information and knowledge management, November 6-11, 2006, Arlington, Virginia, USA [[Abstract](#)].
- Tianbao Yang, Rong Jin (Michigan State University); Yun Chi, Shenghuo Zhu (NEC Laboratories America Inc.), (2009) “**Combining Link and Content for Community Detection: A Discriminative Approach,**” [[Abstract](#)]; presentation [here](#) (discusses link prediction).]  
 Excerpted abstract: In this paper, we consider the problem of combining link and content analysis for community detection from networked data. ...
- Xiaodan Song, Yun Chi, Koji Hino, and Belle Tseng (2007). “**Identifying opinion leaders in the blogosphere.**” In CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pages 971-974, New York, NY [[Abstract](#)]  
 Opinion leaders are those who bring in new information, ideas, and opinions, then disseminate them down to the masses, and thus influence the opinions and decisions of others by a fashion of word of mouth. Opinion leaders capture the most representative opinions in the social network, and consequently are important for understanding the massive and complex blogosphere. In this paper, we propose a novel algorithm called InfluenceRank to identify opinion leaders in the blogosphere. The InfluenceRank algorithm ranks blogs according to not only how important they are as compared to other blogs, but also how novel the information they can contribute to the network. Experimental results indicate that our proposed algorithm is effective in identifying influential opinion leaders.

## 5. BlogPulse (a Nielsen Company)

**BlogPulse** (A “technology showcase” for the Nielsen Company. Also found under the name Intelliseek, BlogPulse is part of the [BuzzMetrics](#) Product Family). This free dashboard is an automated trend discovery system for blogs and says it applies machine-learning and NLP techniques to discover trends in the world of blogs. It is also a blog search engine that analyzes and reports on daily activity in the blogosphere.

### BLOGPULSE STATS

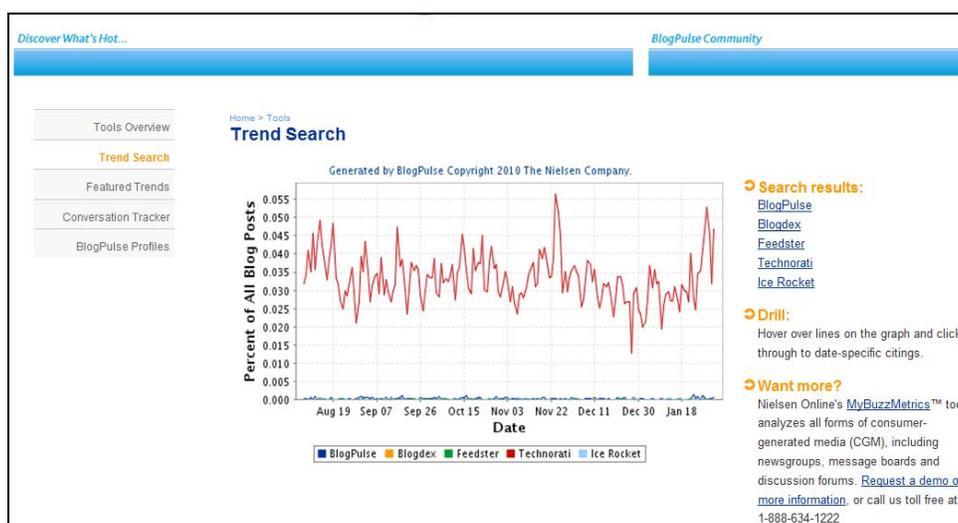
Total identified blogs: 128,861,574
New blogs in last 24 hours: 42,234
Blog posts indexed in last 24 hours: 869

Features include the following:

- A [Search Engine](#) for blogs

- A set of [Buzz-Tracking](#) tools that are applied to blog content daily to track blog activity on key issues, people, news stories, news sources, bloggers and more
- Real-world [Trends](#) as reflected through blogs [including: sports, S&T, business, health, personal and “bizarre stuff.”]
- Daily [blog stats](#) that measure activity in the world of blogging (number of blogs identified, new blogs created, number of blog posts analyzed)
- A [Trend Search](#) that allows users to create trend charts comparing buzz in the blogosphere on up to three specific topics
- A [Conversation Tracker](#) that follows and captures the discussion, or conversation, that emanates and spreads from individual blogs or individual blog posts
- [Blogger Profiles](#) that identify top-ranked blogs and analyze their blog presence, activity and relative influence in the blogging world. [See also this [Key People](#) page that represents the most prominently featured people across blog entries daily.]

Following is an example of a graph of activity for several blog portals from August 19 - January 18.



**Publications:** There are a number of scholarly publications from the Nielsen BuzzMetrics team (see listing [here](#)). Most of the publications were co-authored by **Natalie Glance** from Nielson/Intelliseek and many with researchers at Carnegie Mellon [\*\*See this interesting [case study](#) that refers to the research below. It has observations about **temporal** and **topographical** information].

- **Jure Leskovec**, Mary McGlohon, Christos Faloutsos, **Natalie Glance**, Matthew Hurst (2007). “Cascading Behavior in Large Blog Graphs, Patterns and a model,” In: Society of Applied and Industrial Mathematics: Data Mining SDM07. [Full text [here](#)]

How do blogs cite and influence each other? How do such links evolve? Does the popularity of old blog posts drop exponentially with time? These are some of the questions that we address in this work. Our goal is to build a model that generates realistic cascades, so that it can help us with link prediction and outlier detection. Blogs (weblogs) have become an important medium of information because of their timely publication, ease of use, and wide availability. In fact, they often make headlines, by discussing and discovering evidence about political events and facts. Often blogs link to one another, creating a publicly available record of how information and influence spreads through an underlying social network. Aggregating links from several blog posts creates a directed graph which we analyze to discover the patterns of information propagation in blogspace, and thereby

understand the underlying social network. Here we report some surprising findings of the blog linking and information propagation structure, after we analyzed one of the largest available datasets, with 45, 000 blogs and 1.2 million blog-postings. Our analysis also sheds light on how rumors, viruses, and ideas propagate over social and computer networks. We also present a simple model that mimics the spread of information on the blogosphere, and produces information cascades very similar to those found in real life.

- Jure Leskovec Andreas Krause Carlos Guestrin Christos Faloutsos Jeanne VanBriesen (Carnegie Mellon University); Natalie Glance (Nielsen BuzzMetrics) "[Cost-effective Outbreak Detection in Networks.](#)" [Full text [here](#)]

Given a water distribution network, where should we place sensors to quickly detect contaminants? Or, which blogs should we read to avoid missing important stories? These seemingly different problems share common structure: Outbreak detection can be modeled as selecting nodes (sensor locations, blogs) in a network, in order to detect the spreading of a virus or information as quickly as possible. We present a general methodology for near optimal sensor placement in these and related problems. We demonstrate that many realistic outbreak detection objectives (e.g., detection likelihood, population affected) exhibit the property of "submodularity". We exploit submodularity to develop an efficient algorithm that scales to large problems, achieving near optimal placements, while being 700 times faster than a simple greedy algorithm. We also derive online bounds on the quality of the placements obtained by any algorithm. Our algorithms and bounds also handle cases where nodes (sensor locations, blogs) have different costs. We evaluate our approach on several large real-world problems, including a model of a water distribution network from the EPA, and real blog data. The obtained sensor placements are provably near optimal, providing a constant fraction of the optimal solution. We show that the approach scales, achieving speedups and savings in storage of several orders of magnitude. We also show how the approach leads to deeper insights in both applications, answering multicriteria trade-off, cost-sensitivity and generalization questions.

- "[Finding Patterns in Blog Shapes and Blog Evolution](#)" Mary McGlohon, Jure Leskovec, Christos Faloutsos, (Carnegie Mellon University); Matthew Hurst, Natalie Glance (Nielsen BuzzMetrics). [Full text [here](#)]

Can we cluster blogs into types by considering their typical posting and linking behavior? How do blogs evolve over time? In this work we answer these questions, by providing several sets of blog and post features that can help distinguish between blogs. The first two sets of features focus on the topology of the cascades that the blogs are involved in, and the last set of features focuses on the temporal evolution, using chaotic and fractal ideas. We also propose to use PCA to reduce dimensionality, so that we can visualize the resulting clouds of points. We run all our proposed tools on the ICWSM dataset. Our findings are that (a) topology features can help us distinguish blogs, like 'humor' versus 'conservative' blogs (b) the temporal activity of blogs is very non-uniform and bursty but (c) surprisingly often, it is self-similar and thus can be compactly characterized by the so-called bias factor (the '80' in a recursive 80-20 distribution).

### Intellectual Property (Intelliseek/Buzzmetrics)

Title: Topical sentiments in electronically stored communications

Publication#: [7,523,085](#) (see also: [US20090164417](#), [US20060069589](#) and [WO2006039566](#))

Publication Date: April 21, 2009

Assignee: Buzzmetrics / Intelliseek

Inventors: Nigam, Kamal P.; Hurst, Matthew F.

Abstract: The present application presents methods for performing topical sentiment analysis on electronically stored communications employing fusion of polarity and topicality. The present application also provides methods for utilizing shallow NLP techniques to determine the polarity of an expression. The present application also provides a method for tuning a

domain-specific polarity lexicon for use in the polarity determination. The present application also provides methods for computing a numeric metric of the aggregate opinion about some topic expressed in a set of expressions.

### Other patents

Pub #	Date Issued	Title
<a href="#">7,600,017</a> <a href="#">7,185,065</a>	Oct 6, 2009 Feb 27, 2007	System and method for scoring electronic messages
<a href="#">7,596,552</a>	Sept 29, 2009	Method and system for extracting web data
<a href="#">7,363,243</a>	April 22, 2008	System and method for <b>predicting external events</b> from electronic posting activity
<a href="#">7,197,470</a>	March 27, 2007	System and method for collection analysis of electronic discussion methods
<a href="#">7,188,078</a> <a href="#">7,188,079</a>	March 6, 2007 March 6, 2007	System and method for collection and analysis of electronic discussion messages

### Forrester's 2009 "Listening Platforms" evaluation:

Nielsen BuzzMetrics and TNS Cymfony take top honors. Nielsen BuzzMetrics and TNS Cymfony lead the category because they offer the best balance between technology, insight delivery, and strategy. But each vendor has a distinct set of strengths – Nielsen BuzzMetrics offers a strong analytical and insight capability while TNS Cymfony excels in data collection and media coverage. Both vendors have exceptional text mining capabilities and strong strategy. Many client references commended Nielsen BuzzMetrics for the level of strategic insight it offers. As one brand marketer for a large consumer products organization stated, "Nielsen BuzzMetrics is a critical extension of my social insights function." ...

... Nielsen BuzzMetrics delivers a market leading listening platform that includes sophisticated sentiment analysis capabilities, strong international coverage, and multilingual support. It leverages its exclusive relationships with other Nielsen subsidiaries like Nielsen//NetRatings to deliver critical insights to marketing organizations. Furthermore, the re-write of the reporting and user interface - My BuzzMetrics - extends its market leadership. Clients are positive about the strong vertical practices. However, to extend its lead, Nielsen must build on its consulting practice to support segmentation and innovation.

**Number of Active Customers:** > 125 (October 2008 [source](#)), including CNN, Toyota, Nike, Walt Disney, NFL (2008 [source](#)).

## 6. WiseWindow

**About:** [WiseWindow](#), founded in 2007, provides web measurement technology which mines and analyzes "millions" of opinions expressed in social media each day, identifies only those that relate to a given company or product, and refines those opinions into actionable reports. The measurement, branded as Mass Opinion Business Intelligence (MOBI), purportedly can discover things like total share of opinion, how it changes over time and how it correlates with share of market – measurements that have never been available through market research. Using 10 standard syndicated reports that track what customers think, what they want, who they follow and what they'll buy, WiseWindow says it provides "relevant, actionable decision support to senior executives, marketers and market researchers."

MOBI uses patent-pending technologies developed by CTO Rajiv Dulepet in deep website crawling, auto-classification of opinions, relevance recognition and statistical natural language applications.

*ReadWrite Enterprise* describes the company as:

...using artificial intelligence technology, web crawlers and the processing power of the cloud to get real-time results for enterprise customers. For example, this means that companies may leverage the social Web to make sales forecasts and gauge the opinions of mass society to immediately understand the current opinions about its brand or those of competitors.

WiseWindow calls the product Mass Opinion Business Intelligence, describing it as a service that goes beyond keyword search and click-throughs to predict market movement.

According to WiseWindow, sentiment analysis has failed as a strategic research tool. When matching words, the context is lost. People use words differently to describe their sentiments. The mass amount of data available makes the process overwhelming.

Instead, the WiseWindow web crawler will search for comments and other opinions across thousands of sites that are not blocked by privacy restrictions. The artificial intelligence trains itself to look for a particular topic. It brings back all related opinions. The information is then distilled for the client or made available through a web portal where the data can be analyzed.

Recently, WiseWindow worked with a client from the film industry. WiseWindow used its technology to research 400 films, generating 4.5 million comments from 70,000 sites. They distilled the data to learn what is hot and what is not.

As another example, WiseWindow did research for the film, *Marley and Me*, starring Jennifer Anniston and Owen Wilson. The pre-release promotions featured Luke Wilson. But the comments from the Web demonstrated that Anniston had greater appeal than Wilson. As a result, the trailers were changed to feature Anniston more than Wilson ... WiseWindow started developing its technology in 2007 and began working with clients last year. The company has four patents for its web crawling, auto-classifications of opinions, relevance recognition and in statistical language applications. ([source](#))

#### Patent Applications:

Pub #	Date Issued	Title
<a href="#">20090125382</a>	May 14, 2009	Quantifying a Data Source's Reputation - Methods of quantifying a reputation for a data source are presented
<a href="#">20090125381</a>	May 14, 2009	Methods for identifying documents relating to a market.
<a href="#">20090119157</a>	May 7, 2009	Systems and method of deriving a sentiment relating to a brand
<a href="#">20090119156</a>	May 7, 2009	Systems and methods of providing market analytics for a brand

**News item:** WiseWindow announced in February 2010 that company founder and CTO, Rajiv Dulepet, has been named advisor and architect for a new project funded by the **National Institute of Health** and executed by **Caltech**. The open-source project will develop a web-based bio-computational tool that allows bio-scientists and bio-computation engineers to "crunch data in the cloud" for large-scale tasks such as processing gene sequence data sets on a large cluster of computers. The new tool allows scientists to save considerable time that's now spent waiting for computations on their desktops by moving these operations to the cloud, thereby freeing up their computers for other work. The Caltech project expects to deliver a useful open-source tool for bio-scientists by mid-2010.

In 2004 and 2008, Dulepet served as a visiting scholar at the Stanford School of Management and Engineering, spearheading the development of **U.S. Presidential prediction analysis** ([press release](#)).

This [FAQ sheet](#) has a detailed explanation of the products, services and technology and more.

A list of clients [here](#), including Microsoft, Acer, Cisco, 20<sup>th</sup> Century Fox, Paramount and eHarmony.

**Multilingual capabilities:** The basic offering is in English, “although the technology can transcend Latin based languages (i.e., French, Spanish, German), but the company says that, “Capabilities can be developed and offered if a requirement presents itself (source: FAQ sheet).”

**Cost:** “...cost range generally in the tens of thousands (source: FAQ sheet).”

**Other:** See this [case study](#) that measured sentiment for the movie “Fighting” from April to May of 2009.

## 7. Visible Technologies

Visible Technologies, which has recently [announced](#) a partnership with (and investment from) In-Q-Tel, offers a suite of products that “enable real-time visibility into online social conversations regardless of where dialogue is occurring.” The company positions its truREPUTATION offering as “a best-in-class online reputation management service that provides both individuals and brands an effective way to repair, protect and proactively promote their reputation in search engine results.”

A screenshot from the company’s TruVoice product (part of the TruCast suite of products) is shown below.



**Case Studies:** [Xerox](#); [Microsoft](#)

**Size:** ~90 people

**Revenue:** ~\$20 M in 2010 ([source](#)).

**Customers:** ~100 ([source](#)), including Microsoft, Xerox, Autodesk and Boost.

**Multi-lingual capabilities:** Unknown.

**Cost:** \$200K for monitoring one to two brands, to several million dollars to monitor many, per year (2007 [source](#)).

### Intellectual Property:

Title: Systems and methods for consumer-generated media reputation management

Publication #: [WO/2007/143314](#) / [US20070294281](#)

Publication date: 12/13/2007

Inventors: Ward, Miles; Webber; Jim; Graziano, Dean Michael

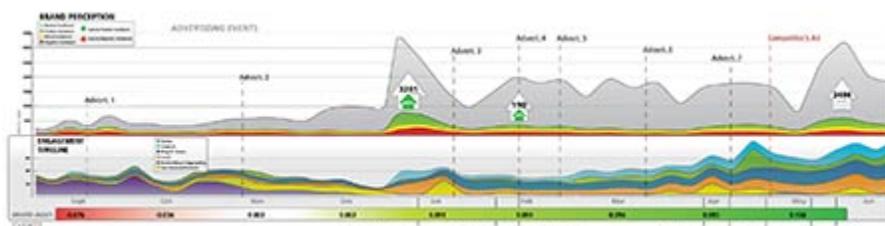
TruCast is a method for management, by way of gathering, storing, analyzing, tracking, sorting, determining the relevance of, visualizing, and responding to all available consumer generated media. Some examples of consumer generated media include web logs or "blogs", mobile phone blogs or "mo-blogs", forums, electronic discussion messages, Usenet, message boards, BBS emulating services, product review and discussion web sites, online retail sites that support customer comments, social networks, media repositories, and digital libraries. Any web hosted system for the persistent public storage of human commentary is a potential target for this method. The system is comprised of a coordinated software and hardware system designed to perform management, collection, storage, analysis, workflow, visualization, and response tasks upon this media. This system permits a unified interface to manage, target, and accelerate interactions within this space, facilitating public relations, marketing, advertising, consumer outreach, political debate, and other modes of directed discourse.

**Papers/Publications:** No scientific papers were found for this company. VT provides PR-oriented product whitepapers which contain very little content (see [here](#)).

**Forrester's 2009 "Listening Platforms" evaluation:** Visible Technologies was cited as being a "strong performer." The findings were based on vendors' current offering, strategy and market presence. The report stated that, "Visible Technologies delivers the most technology-focused solution set but needs to improve its strategic services." Additionally, the Forrester report says:

Visible Technologies, a new entrant to the listening platform category, offers an impressive technology stack. Visible Technologies delivers strong reporting capabilities and is the vendor best positioned to extend social media analysis data and attributes into CRM systems. Visible Technologies is a good fit for the self-sufficient and socially aware marketer. But the vendor must bulk up its consulting and strategic insight delivery capabilities to play a bigger role in a firm's social strategy.

Description of the SA product: ...For definition, we consider tone to be the measurement of the



overall tenor of a piece of content, and sentiment to be the specific measure of the Positives, Negatives, Mixed emotions related to a specific topic of interest. Our systems first topically decompose a blog post or article so that we understand the subjects of

conversation and particularly the ones that our clients are interested in studying. This is an area that is not talked about as often, but just as important to the overall measurement. The capacity to understand sentiment is most critical when you can be sure that the sentiment you are measuring is related to your product/company/issue ([source](#)).

## 8. J.D. Power & Associates Web Intelligence

J.D. Power, a respected provider of market research, provides “web intelligence” products that monitor blogger postings (Power has this functionality because the company purchased Umbria, a blog monitoring company). Analysis includes assessment of sentiment and also **classifies blogger postings by their demographics, which are assessed on the basis of “speech pattern” markers**. They describe their approach in a brochure as:

Consider the type of analysis that’s available today. Let’s start at the basic level of monitoring “buzz,” or the number of mentions on a topic. While informative, it’s difficult to make business decisions based purely on volume of discussion when you don’t know who’s talking or what they’re saying. Trying to make decisions with this level of analysis is like trying to fix your car without opening the hood. You know there’s a noise, and you can make some educated guesses, but not much more.

Next let’s layer in positive and negative sentiment. That’s more interesting, because if you see a sudden spike in negative sentiment, at least you know you should delve into the issue more.

Finally, adding the layer of demographics gives a level of information and specificity that allows you to make some real decisions. You know who’s saying what and why. J.D. Power is unique in its ability to use speech patterns to identify whether authors are likely to be male or female; Gen Y, Gen X or Baby Boomer. The development of this proprietary system enables the transformation of unstructured but highly valuable online discussions into actionable market intelligence. J.D. Power uses a combination of natural language processing methodologies, machine learning algorithms and business rules to produce its analysis.

J.D. Power Web Intelligence uses this patent-pending technology to make social media research more valuable than our competition. Our algorithms collect, de-spam, and categorize millions of publicly available blog posts daily. ([Source](#))

The company provides a number of interesting case studies about how such analysis (sentiment + demographics) has been used by businesses (for the entire list, see this [link](#)). One example is shown below.

**Situation:** A successful dessert brand launched a narrow low-carb product line in response to consumer interest in the recent diet trend. Concerned that this trend was a short-lived fad, the company asked Umbria to help them better understand the way consumers really felt about low-carb products and the diet, in general.

**Research:** Umbria developed a Market Tracker Report to monitor the way bloggers discussed the low-carb diet in relation to other popular diet programs, such as Atkins, Weight Watchers, and the South Beach Diet.

**Recommendation:** Paying particular attention to the sentiment of the blog postings we identified, Umbria tracked the rise in the popularity of low-carb diets, but also determined a point in time when negative buzz overtook positive comments about the low-carb diet. Convinced that consumer interest in the trend was quickly waning, our client decided not to invest in any additional low-carb line extensions, and later ceased the line extension.

## Intellectual Property:

Title: Automatic Sentiment Analysis of Surveys

Publication #: [US20090306967](#)

Publication date: December 10, 2009

Inventors: Nicolov, Nicolas; Tuohig, William Allen; Wolniewicz, Richard Hansen

In one aspect, the invention provides apparatuses and methods for determining the sentiment expressed in answers to survey questions. Advantageously, the sentiment may be automatically determined using natural language processing. In another aspect, the invention provides apparatuses and methods for analyzing the sentiment of survey respondents and presenting the information as actionable data.

Title: Tribe or group-based analysis of social media including generating intelligence from a tribe's weblogs or blogs

Publication #: [US20080215607](#)

Publication date: September 4, 2008

Assignee: Umbria

Inventors: Kaushansky, Howard; Kremer, Ted V.; Nicolov, Nicolas; Tuohig, William A.; Wolniewicz, Richard Hansen

A computer-based method for generating intelligence from social media data, such as blog data, that is publicly available on the Internet. A server is provided that runs a tribe analysis tool, and the method includes accessing a set of the social media data with the tribe analysis tool. The social media data is associated with a plurality of network users or authors. The method continues with operating the tribe analysis tool to identify members of a tribe from the authors by processing the set of social media data to determine the authors having associated portions of the social media data that satisfies tribe membership criteria. Common interests for the identified members of the tribe are determined by processing the social media data associated with the tribe authors. A report is generated for the tribe that includes information related to the set of common interests and additional generated tribe-based intelligence.

**Papers/Publications:** A list of Umbria scientists' papers and publications are below and on this [web page](#) with abstracts.

- Salvetti, F. and Nicolov, N. (2006), "[Weblog Classification for Fast Splog Filtering: A URL Language Model Segmentation Approach](#)," To appear in Proceedings of HLT-NAACL 2006: Human Language Technology Conference, New York City, NY. [[Abstract](#)]
- Salvetti, F. and Srinivasan, S (2005), "[Local Flow Betweenness Centrality for Clustering Community Graphs](#)." In Proceedings of WINE 2005, 1st Workshop on Internet and Network Economics, Hong Kong. Springer, Lecture Notes in Computer Science. [[Abstract](#)]
- Salvetti, F., Reichenbach, C. and Lewis, S, (2005) "[Opinion Polarity Identification of Movie Reviews](#)." To appear in Computing Attitude and Affect in Text, Springer, Dordrecht, The Netherlands, 2005. [[Full Text](#)]
- Boguraev, B. and Nicolov, N. (2005). "[Current Trends and Techniques in Temporal Analysis](#)", EUROLAN'2005. [[Abstract](#)]
- Salvetti, F. and Srinivasan, S. (2005), "[Information Flow using Edge Stress Factor](#)," WWW 2005, Special interest tracks and posters, Chiba, Japan, 2005. [[Abstract](#)]
- Florian, R., Hassan, H., Ittycheriah, A., Jing, H., Kambhatla, N., Luo, X., Nicolov, N. and Roukos, S. (2004). "[A Statistical Model for Multilingual Entity Detection and Tracking](#)," HLT-NAACL 2004: Human Language Technology conference / Annual meeting of the North American chapter of the Association for Computational Linguistics, Boston, Mass., 2004 [Full text [here](#)]

- Salvetti, F., Reichenbach, C. and Lewis, S (2004). "[Impact of Lexical Filtering on Overall Opinion Polarity Identification](#)," Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text, 2004. [[Abstract](#)]

**Multilingual capabilities:** The Forrester report (info follows) mentions that JD Powers provides "limited multilingual support," but Perspectives found no specific information on what that means, other than what was in a 2004 paper listed above – researchers performed entity detection experiments in Arabic, Chinese and English texts.

#### **Forrester's 2009 "Listening Platforms" evaluation:**

J.D. Power & Associates, which acquired Umbria in 2008, offers good coverage of social media sources but continues to offer limited: 1) coverage of mainstream media like print, radio, and TV; 2) multilingual support; and 3) user interface support for the advanced analytical marketer. Still, J.D. Power & Associates impressed us with its grasp of sentiment analysis and the use of tribe-analysis techniques to drive market segmentation projects. Client references are generally positive, but relationships are one-off exercises rather than sustained partnerships. J.D. Power & Associates is a good fit for consumer product companies looking to integrate existing segmentation efforts and social analysis.

## **9. Dow Jones Insight**

[Dow Jones Insight](#) reportedly takes in more than 1.5 million articles in **23 languages** from press release wires, newspapers, magazines, radio and TV transcripts, Web sites, blogs and message boards ([source](#)). According to the 2009 Forrester report, the company:

... has made great strides in its data coverage capabilities since the previous Brand Monitoring Wave. Automated sentiment analysis coupled with strong multilingual capabilities makes Dow Jones a good choice for PR professionals and marketing communications specialists for large global organizations. On the flip side, Dow Jones Insight has room for improvement in the creation and delivery of strategy and insight. Client references were generally PR professionals using the platform for tracking and monitoring purposes. The next step for Dow Jones Insight? Extend its visibility beyond PR and deliver strategic insights to the entire marketing organization. (Case studies [here](#).)

**Cost:** \$50K-\$350K (2008 [source](#)).

[Dow Jones Economic Sentiment Indicator](#): While Dow Jones ESI is looking at news items, and not raw "commentary" from blogs, this too claims to **predict** economic status.

The Dow Jones Economic Sentiment Indicator aims to predict the health of the U.S. economy by analyzing the broad coverage of 15 major daily newspapers in the U.S....

The ESI represents one of the most comprehensive and far-reaching examinations of media coverage as an economic indicator. The ESI's back-testing to 1990 shows that the ESI clearly highlighted the risk that the U.S. economy was sliding into recession in 2001 and 2008 and suggests the indicator can help predict economic turning points as much as seven months in advance of other indicators...

The Dow Jones Economic Sentiment Indicator is calculated using a proprietary algorithm through Dow Jones Insight, a media tracking and analysis tool.

2006 [whitepaper](#), "[Beyond the Numbers: An Analysis of Optimistic and Pessimistic Language in Earnings Press Releases](#)."

## 10. BuzzLogic

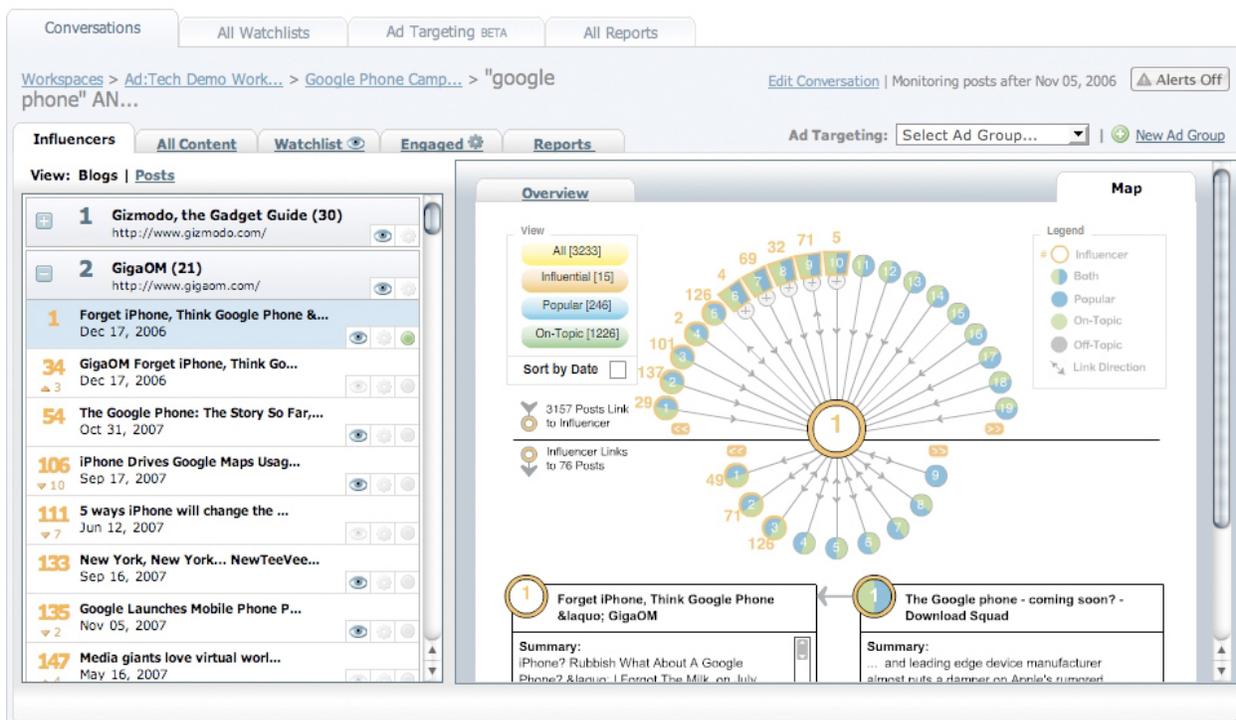
**About:** BuzzLogic's ad targeting platform identifies the quality blogs and other peer sites responsible for shaping consumer perception and buying behavior. The platform layers in third-party audience measurement and intent data. BuzzLogic says it enables top media companies to benefit from its platform by gaining access to relevant, conversationally-targeted campaigns and new revenue. "With more than 78 million monthly uniques, BuzzLogic represents access to the largest pool of trusted blog content on the Web, connecting advertisers to millions of consumers passionate about everything from gaming and gadgets to parenthood and politics."

### Software:

**BuzzLogic's** software provides an automated analysis of the influence of sites and individual items (such as blog posts). The analysis is based on message volume, on-topic frequency, link analysis in context, and - when possible - audience reach. The company differentiates between relevant, contextual links and lower-value links, such as blogrolls and template items, to improve its analysis. The map interface below illustrates the connections into and out of an item, allowing users to explore the links among sites. Blogs and posts are ranked for influence, and influencer profiles explain the links that contribute to the rankings ([source](#)).

BuzzLogic influence calculation is completely automated. Customers need only to plug in conversation "queries" (or keyword search terms) to guide BuzzLogic's algorithm in uncovering the influencers on that topic. BuzzLogic's Conversation Index is built on a number of trust factors to keep out spam and "splogs." To qualify, a blog must be linked to by a certain number of other trusted blogs within our index. This ensures the content we present to our customers is relevant and high quality. BuzzLogic also has a system for removing duplicate content, and a number of other capabilities, such as content sectioning, that enables us to zone in on specific areas within a blog or social media site - such as comments, title and the body of the post - and deliver the most relevant analysis possible.

The data is presented via a dashboard (see below) that illustrates lists of top influencers, as well as a view that displays individual posts in order of influence. BuzzLogic also offers a Social Map - a visual depiction of the linking activity around any given post, complete with which sites are linking, which sites are being linked out to ([source](#)).



### Papers/Publications:

Whitepaper: "Five Steps to Optimizing Paid Media with Social Intelligence" (download [here](#)).

### Intellectual Property:

Pub #	Date Issued	Title
<a href="#">20090119173</a>	May 7, 2009	System and Method For Advertisement Targeting of Conversations in Social Media
<a href="#">20070214097</a>	Sept 29, 2009	Social analytics system and method for analyzing conversations in social media

**Clients include:** Walmart, HBO, Starbucks, Microsoft, American Express, CBS Television, razorfish, etc. (more listed on this [page](#)); case studies [here](#).

**Multilingual capabilities:** BuzzLogic indexes content from **19 languages** and plans support for additional languages ([source](#)).

**Employees:** 27

**Cost:** BuzzLogic is a subscription-based service starting at \$1,000 per month ([source](#)).

## 11. Yahoo! Research

[Yahoo! Research](#) has long been active in the field of sentiment analysis, and so we include a brief discussion of some of their activities.

### Project:

[Graph partitioning](#): The design of efficient algorithms for graph clustering is a long-standing and active research topic in computer science. Yahoo! researcher Kevin Lang, together with Reid Anderson, Fan Chung, Anirban Dasgupta, Jure Leskovec, Michael Mahoney, Lorenzo Orecchia and Satish Rao, is leading the effort in finding the clusters of Yahoo!'s sponsored search spending graph. Initially, using expensive partitioning algorithms, the team at Yahoo! Research discovered that even when the graph is large, containing 100 million nodes, the clusters that are typically found are much smaller, with a size range of one hundred to one million nodes - implying that the time it takes to calculate the clusters should be much faster. ([“Statistical Properties of Community Structure in Large Social and Information Networks”](#)). This discovery led the team to study a new class of improved algorithms, namely “local partitioning” algorithms, which have the unusual property that their run time is proportional to the size of their resulting clusters (the output), and not the size of the original graph (the input). ([“Local Graph Partitioning using PageRank Vectors”](#)). The conceptual shift from global to local partitioning algorithms has permitted much larger graphs to be processed than before,” said Lang. “We now have the world’s most scalable graph partitioning algorithms that are especially applicable to large complex networks.”

**Yahoo! Researchers.** Yahoo! has many well-respected researchers on its staff working in the sentiment analysis and social networking fields including:

- **Bo Pang** is a researcher at Yahoo! and is the well-known in the sentiment analysis community. She co- wrote the book-length survey, [“Opinion Mining and Sentiment Analysis,”](#) mentioned earlier in the report ([here](#)). A list of her recent publications, projects and news is available [here](#).
- **Belle Tseng**, formerly of NEC Labs, is currently senior manager at Yahoo! Several of her papers are listed in the NEC Laboratories section of the report [here](#).
- **Ravi Kumar** has co-authored several interesting papers in the social network field (see a partial list [here](#)), including:

[“Structure and evolution of online social networks.”](#) [Full text [here](#)]

In this paper, we consider the evolution of structure within large online social networks. We present a series of measurements of two such networks, together comprising in excess of five million people and ten million friendship links, annotated with metadata capturing the time of every event in the life of the network. Our measurements expose a surprising segmentation of these networks into three regions: singletons who do not participate in the network; isolated communities which overwhelmingly display star structure; and a giant component anchored by a well-connected core region which persists even in the absence of stars. We present a simple model of network growth which captures these aspects of component structure. The model follows our experimental results, characterizing users as either passive members of the network; inviters who encourage offline friends and acquaintances to migrate online; and linkers who fully participate in the social evolution of the network.

[“Structure and evolution of blogspace.”](#) [[Abstract](#)]

A critical look at more than one million bloggers and the individual entries of some 25,000 blogs reveals blogger demographics, friendships, and activity patterns over time.



**Number of active customers:** More than 200

**2009 Forrester report** observations about Radian6:

Radian6 lacks insight and strategy delivery capabilities beyond PR. Radian6 is an emerging vendor in this category, but it's currently limited by a focus on the Public Relations function only. Radian6 focuses on keyword analysis but offers no capabilities for text analysis and natural language processing. This hinders Radian6's ability to offer sentiment analysis. Radian6 offers a strong user interface and monitoring setup capabilities for PR teams (report [here](#)).

In the WebMetricsGuru sentiment analysis [product comparison](#) of social media monitoring platforms (mentioned above in the Sysomos section) the conclusions about Radian6's product were that:

... none of the 22 tweets that were flagged negative - were actually negative when I read them visually. ...if you really care about the accuracy of your results around sentiment - best to look at them by hand and reclassify them to what you think they are - Radian6 may not be smart enough to tell sentiment. ... Radian6 is getting better - but the problem is, as far as sentiment goes - there's too much that has to be done on a configuration level to get sentiment right - I know they will continue to improve - but they also come from a "marketing" background and the Flash Interface, that makes their product look better than the others, is also, at time, frustrating to work with.

No publications or intellectual property found for this vendor.

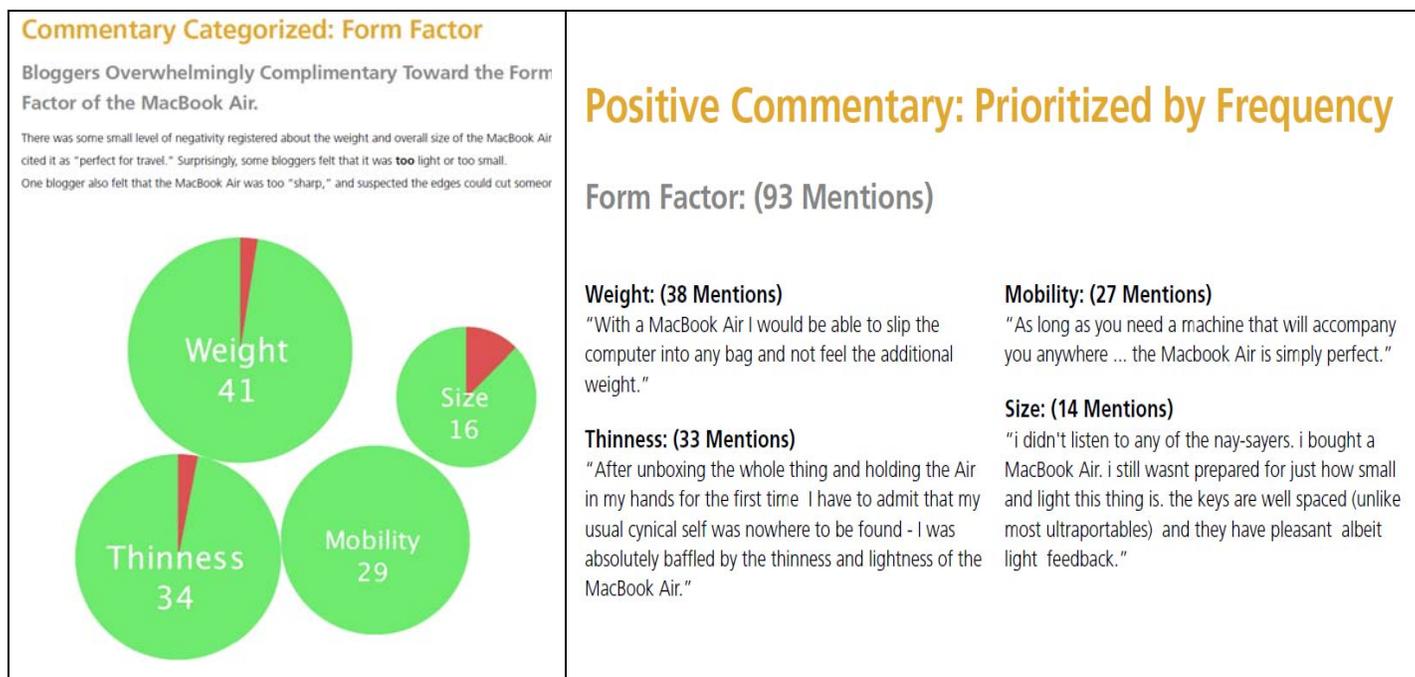
### 13. Newssift

Newssift is an intelligent search tool from the Financial Times Group. Their goal is to make search more useful and valuable by enhancing meaning, identifying relationships and **sorting articles on sentiment to allow the user to "find what's encouraging or troubling about an issue, company, or product."** Sentiment is assigned to each article and users can view articles on the basis of that classification. It is unclear if the sentiment analysis is sophisticated; but the product is interesting for the way it consolidates information into a dashboard. Below is a screen capture of the results of searching Newssift on "sentiment analysis" (note the sentiment graphic on the left side).

The screenshot shows the Newssift search interface. At the top, there's a search bar with the term "Sentiment Analysis" and a search button. Below the search bar, there are several filters: Business Topic (Monetary Volatility, Equipment And It Solutions, Product Lifecycle, Advertising And Marketing, Public Relations), Organization (Google Inc. (GOOG.NSQ), S & P Company, Inc., International Business Machines Corporation (IBM.NYQ), Apple Computer, Inc. (AAPL)), Place (United Kingdom, New York, China, Germany, San Francisco, Seattle, Europe), Person (Bernie Schaeffer, Robert Prechter, Andrea Kramer, Barack Obama, Jean-Michel Texier, Robert Fisk, Jocelynn Drake), and Theme (Sentiment Analysis, Natural Language, Text Mining, Unstructured Data, Technical Analysis, Opinion Mining, Sentiment Analysis Tools). Below the filters, there's a sentiment graphic showing a pie chart with three segments: Positive (140), Neutral (68), and Negative (17). The main content area displays "1-10 of 225 Articles" and a list of articles, including "Nstein's Sentiment Analysis Key Component for evolve24" and "Japan's new finance minister in U-turn".

## 14. Sentimine from the Parnassus Group

The [Parnassus Group](#) offers Sentimine, which assesses sentiment and tone of the blogosphere for products, services, companies, or individuals. A sample [report](#) on Apple's Macbook Air laptop is shown below. It categorizes sentiment for specific product features.



A sample of Sentimine's real time tracking can be found [here](#).



## 15. Nstein

Nstein Technologies Inc. offers a sentiment analysis engine. A recently announced partner is evolve24, a business analytics and research firm. evolve24 uses Nstein's sentiment analysis for reputation risk measurement. evolve24 analyzes traditional and social media to determine businesses' overall information landscape and provides quantitative metrics around perception, reputation and risk that let clients understand the key areas of impact in their marketing, communications and management efforts. According to the company, Nstein's Sentiment Analysis module allows evolve24 to quickly determine the tone of an article, which offers a key insight into the firm's representation through media. **"We have estimated that a person at peak performance can 'tone' 200 articles a day - or 25 an hour,"** Wheeler said. **"With Nstein, we can literally score tens of thousands of pieces of content a day for each of our clients - all of them broken down by brand or topic - using a single person to manually test and verify results (source)."**

## 16. PeopleBrowsr

PeopleBrowsr is a data mining and real-time search engine that looks into digital conversations and engages across multiple social networks simultaneously. With filtering, a user can retrieve 'memes and themes' that are important across any Web 2.0 service. "We've built a set of applications sitting on the data mine to monitor your brand, identify your audience, analyze tweets sentiment, filter the buzz, manage feedback, share accounts, run campaigns, track keywords, build widgets and engage across multiple social networks simultaneously." See this [page](#) for a list of their product.

PeopleBrowsr offers many use cases on their [website](#); a full perusal is very interesting. A few screenshots exemplify some of their work under the category of sentiment analysis ([source](#)). For the reader's convenience, we reproduce a screenshot from their "hot stocks" dashboard:

Powered by PeopleBrowsr™

Celebrity | Airlines | Stocks | Subscribe | Buy Research Report

**Stocks Hotlist** [buy now >](#)

What are Twitter users saying about the stock markets?  
PeopleBrowsr analyzes all the tweets mentioning \$Stocks and rates them as positive or negative.

Stocks HotList Sentiment Report [Show All Tweets](#) | [Hide All Tweets](#) | [Show Live Stats](#)

Stock Name	Positive Tweets	Negative Tweets	Total Mentions	Last Trade	Day Change	Vote & Tweet
Amazon.com, Inc. \$amzn	20%	3%	1178	122.07	0.00%	<a href="#">Get Tweets</a>
Apple Inc. \$aapl	13%	3%	879	197.37	0.00%	<a href="#">Get Tweets</a>
Microsoft Corpora \$msft	12%	2%	758	28.59	0.00%	<a href="#">Get Tweets</a>
S&P DEP RECEIPTS \$spy	11%	8%	522	106.42	0.00%	<a href="#">Get Tweets</a>
ENDEAVOR INTL COR \$end	2%	2%	448	1.08	0.00%	<a href="#">Get Tweets</a>
EURUSD \$eurusd	26%	5%	441			<a href="#">Get Tweets</a>
GBPUSD \$gbpusd	14%	9%	398			<a href="#">Get Tweets</a>
S&P 500 \$sp500	12%	10%	385		0.00%	<a href="#">Get Tweets</a>
GOLDMAN SACHS GRP \$gs	12%	6%	383	178.61	0.00%	<a href="#">Get Tweets</a>
Google Inc. \$goog	8%	1%	334	548.29	0.00%	<a href="#">Get Tweets</a>
HITACHI LTD ADR \$hit	7%	2%	266	32.97	0.00%	<a href="#">Get Tweets</a>
BioCryst Pharmace \$bcry	19%	5%	249	10.55	0.00%	<a href="#">Get Tweets</a>
Baidu, Inc. \$bidu	20%	6%	241	383.66	0.00%	<a href="#">Get Tweets</a>
BK OF AMERICA CP \$bac	8%	6%	214	15.45	0.00%	<a href="#">Get Tweets</a>

Showing all Stock Tweets

[@macheterosforvr](#) 233ema/hr @ 9,763.17 a very important level to watch! IF not, lookout under!

[@RobTheStreet](#) Wednesday's early headlines: <http://tinyurl.com/yk4pbsb> \$BAC \$COP \$IP \$Q \$JNY \$WLP \$SAP web

[@TXR\\_miyabin](#) <http://twitpic.com/n944g> - [\\$BE47](#)-\$= \$NB>=t192s

[@vital\\_sign](#) @howlingeverett @afin83 - just make sure you sell it in your presentation - I mean micro\$oft manage to make piles of shit look good :P

[@kbunky1](#) \$wlp needs to be investigated by congress for the big profits they made on the back of sick men and women and kids health bill will stop it web

## G. Sentiment Analysis Related to National Security

Below are brief profiles of a selection of organizations that are clearly related to counterterrorism or other national security activities.

### 1. SentiMetrix

SentiMetrix, a four-person spinoff from the **University of Maryland Institute for Advanced Computer Sciences (UMIACS)**, was established in 2006. The group has developed and patented a sentiment analysis scoring software program, [SentiGrade™](#). SentiMetrix claims that no competitors offer similar software for multilingual, contextually based, fine-grained scoring. The group positions itself as:

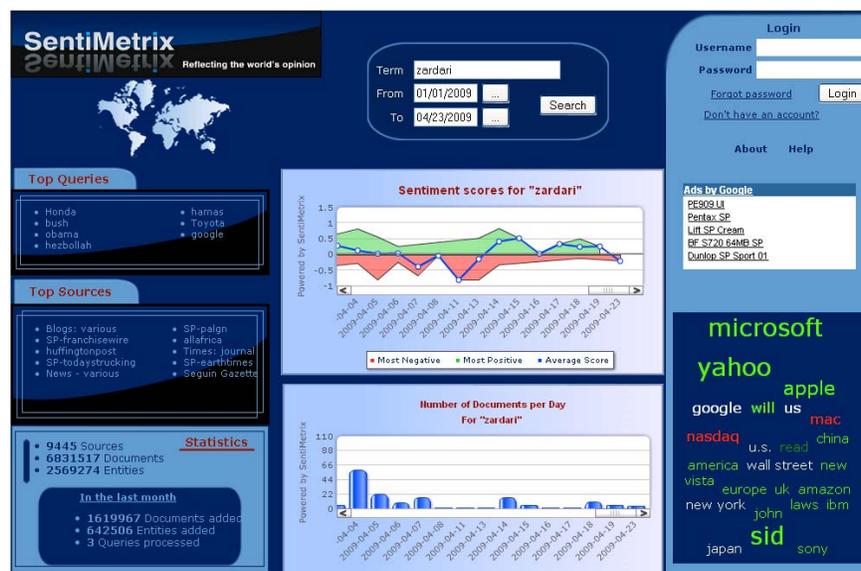
...an ideal partner for US government agencies and their trusted subcontractors. SentiMetrix has chosen to maintain a low profile in the commercial sector while it quietly works to ensure its products are the best fit for the US defense industry. Our intent is to work with the US government and its friends and allies to aid in their efforts to develop predictive analytical tools and political scenario development. (Source: [Opinion Analysis and National Security - a SentiMetrix white paper](#))

SentiGrade is based on the [OASYS technology](#), developed at the UMIACS, and SentiMetrix holds the exclusive license to commercialize OASYS.

The software analyzes opinions expressed in online textual media in nine languages [not including Farsi, according to this [whitepaper](#)] and is ideally suited for use in the government and commercial market sectors where “multilingual, highly accurate and real-time analysis is critical.” Identifying sentiment changes over time toward a determined object is perhaps the most critical step in the art and science of predictive analysis. SentiGrade provides access to data collected from a variety of sources from around the world, including over 9000 news/media outlets and a million top blogs. SentiMetrix is a pioneer in the rapidly evolving world of text analytics and whose sentiment analysis scoring methods and multilingual capabilities are superior to commercial market applications. ([source](#))

Features include:

- Real time scoring of sentiment pertaining to a specified object
- [Languages] Sentiment scoring capability is available in nine different strategically important languages: English, Spanish, French, German, Italian, Russian, Chinese, Arabic and Korean
- Implied sentiment is derived from contextual analysis in its language of origin. Fine grained scoring of sentiment that does not crudely classify sentiment as positive/neutral/negative, but provides a continuum of scores between -1 (“maximally negative”) to +1 (“maximally positive”) allowing detection of relatively small shifts in opinion
- Determined by two different analysis groups (commercial and government) to be as accurate as human analysis in less time and with fewer emotional “biases.”



SentiMetrix claims it can address national security problems including:

- Identifying terrorist sympathizers: To identify possible terrorists or sympathizers, analysts may use the SentiMetrix tool set to search the internet for possible subject targets and note changes in sentiment toward aforementioned targets. By being able to search multiple blog posts and documents simultaneously, in multiple languages, and to see a graphical representation of huge volumes of information, an information analyst is given the fastest window onto relevance of “chatter” pertaining to an object of interest. Once a “spike” or “drop” is noted, the analyst can quickly retrieve the document and information set responsible for the sentiment change. The software can also be used with surveillance data logs in a textual format to formulate sentiment scores over time.
- Identifying opinions towards the US: The US Defense Department and intelligence community can monitor the sentiment towards the US and US government officials and institutions in the foreign press and in foreign blog postings. This will help the US government shape social communications both with that country’s leaders, as well as with interested parties in that country who may hold differing views.
- Identifying opinions towards “friends” of the US: The US DoD is involved in major operations in many parts of the world. The ability to check on strength of sentiment towards its allies is critical - if rhetoric towards allied leaders turns increasingly negative, the ability to recognize this in real time and mitigate it through other moves may be critical in saving lives ([source](#)).

The company says that SentiGrade has been independently tested for accuracy both by the University of Maryland and by SAIC (March 2008), the latter on behalf of federal sponsors. In the University of Maryland’s tests of SentiGrade, a set of human evaluators scored the sentiment of randomly selected news articles on various topics on the same -1 to +1 scale. SentiGrade also computed sentiment scores for the same test sample. The correlation between the human evaluators and SentiGrade was 0.59 on a -1 to +1 scale (-1 denotes a 100% inverse correlation and +1 denotes the fact that the two scores are exactly identical). The experiments also computed the average correlation between any two of the human evaluators – this average correlation between the humans was 0.57. Thus, says the company, SentiGrade, “**performs within the range of scoring of an ordinary human evaluator.**” In the second set of accuracy testing done by SAIC, tests were conducted on Burma blogs, Amazon reviews, and newspaper articles. According to the draft report, one of the strengths of SentiGrade was that “Sentiment scores generally track with actual sentiment of document.” They also asserted that “OASYS [the core technology in SentiGrade] is suitable for transition. The software met the parameters of the testing and usability.” (Source: [Start Making Sense-of Online Opinions: a SentiMetrix white paper](#))

In another whitepaper, Sentimetrix says that: "The SentiMetrix technology has been independently evaluated by SAIC on behalf of the Intelligence Community and **found to be suitable for transition to the IC. There is no known competitor with similar credentials in the government sector** where we believe SentiMetrix have the greatest affinity... SentiMetrix products represent the cutting edge of opinion analysis tools. Our software can and should be used as a first line of defense in developing a proactive strategy for government and security applications that necessitate high volume, real-time, contextual, multilingual opinion analysis. This is perhaps the most important building block of a comprehensive strategy for developing predictive analysis tools already sought by our government." (Source: [Opinion Analysis and National Security: A SentiMetrix White Paper](#))

Papers describing the OASYS technology:

- C. Cesarano, B. Dorr, A. Picariello, D. Reforgiato, A. Sagoff, V.S. Subrahmanian (2006). "[OASYS: An Opinion Analysis System](#)". In Proceedings of AAAI-2006 Spring Symposium on Computational Approach to Analyzing Weblogs, March 2006. [Full text [here](#)]  

In a world where conflicts increasingly are not between two standing national armies but between governments and opposing ideological, factional, terrorist, and insurgent groups, computer tools that could help predict the future plans and attacks of such groups could play an important role in military strategy.
- F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato and V.S. Subrahmanian (2007) "[Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone.](#)" Accepted to ICWSM 2007. [Full text [here](#)]
- Carmine Cesarano, Antonio Picariello, Diego Reforgiato and V.S. Subrahmanian (2007). "[The OASYS 2.0 Opinion Analysis System,](#)" Accepted as DEMO to ICWSM 2007. [Full text [here](#)]

## 2. MITRE Corporation

MITRE is an FFRDC that works on behalf of a range of sponsors, especially from the Air Force and various USG agencies. We identified three interesting research areas that MITRE names as key investments for 2010:

**Measuring and Guiding Engagement [SASCA and FABTAC]:** Twenty first century societies are driven by a world of opinion and pressure from many complex and distinct publics. It is vital for U.S. interests to be able to operate in this environment effectively often in direct and active competition by adversaries for hearts and minds. ... In our research program, MITRE is developing novel methods for effectively measuring the impact of engagement and searching for early indicators that could be used to guide future engagement events. Key investments in FY10 include Sentiment Analysis for Strategic Communication Assessment (SASCA), Forum and Blog Threaded Comment Analysis (FABTAC), the exploration of the ability of social media (e.g., Twitter) to serve as a proxy for traditional polls, and foreign language blog analysis.

**Public Opinion Polling by Proxy (POP/P).** The exploration of the ability of social media (e.g., Twitter) to serve as a proxy for traditional polls

...

**Enhancing Intelligence Analysis:** Intelligence Analysis is the process of taking known information about situations and entities of importance, characterizing the known and, with appropriate statements of probability, the future actions in those situations by those entities. The descriptions are drawn from what may only be available in the form of deliberately deceptive information; the analyst must correlate the similarities among deceptions and extract a common truth. Virtually all of MITRE's customers have a need to apply these approaches to incomplete data sets for prediction at some point. MITRE is investing in research efforts that are targeted at improving our ability to execute Intelligence Analysis of all types. We are particularly interested in analysis of external threats to the nation. Our current research focuses on methods that help analysts deal

more effectively with large data sets within particular scientific domains, cultural differences, and language challenges (sources [here](#) and [here](#)).

Following are slides outlining these research areas from a July 2009 “Smart Power Systems” briefing [presentation](#) (more detail on its Smart Power corporate initiative [here](#) and [here](#)). This framework includes use of sentiment analysis.

## Sentiment Analysis for Strategic Communication Assessment (SASCA)



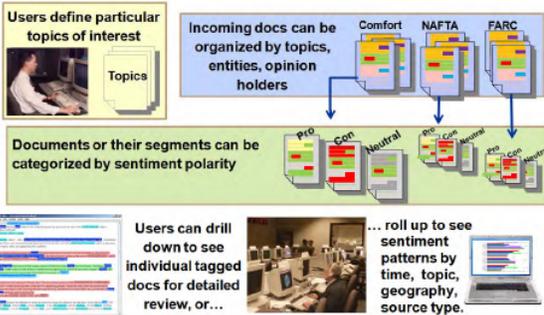

### Challenge

- Strategic Communication (SC) is increasingly important tool for nation’s global interests
- COCOMs are struggling with how to assess impact of their SC activities
  - Current methods rely on polling – expensive, delayed, short term, revealing
- Views of regional players are “hidden in plain sight” amidst tidal wave of regional media

### Objectives

- Increase fidelity, timeliness and media coverage of strategic communication assessment (SCA)
- Enable COCOM to baseline and continuously monitor and evaluate SCA
- Visualize results by time, geography, topic, demography.

### SASCA Notional Prototype



The diagram illustrates the SASCA Notional Prototype workflow. It starts with 'Users define particular topics of interest' (e.g., Comfort, NAFTA, FARC). 'Incoming docs can be organized by topics, entities, opinion holders'. Documents are then categorized by sentiment polarity (Pro, Con, Neutral). Users can drill down to see individual tagged docs for detailed review, or roll up to see sentiment patterns by time, topic, geography, and source type. The MITRE SMART POWER logo is visible at the bottom left of the slide.

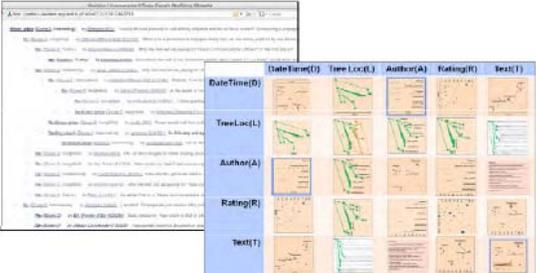
NGC: Commercial Activities in Web Forecasting

PERSPECTIVES

## Forum and Blog Threaded Comment Analysis (FABTAC)



Dr. Christine Doran

<h3>Objectives</h3> <ul style="list-style-type: none"> <li>Collect and analyze forum comment threads, in particular those representing opinions on issues of interest to MITRE sponsors</li> <li>Leverage structure inherent to the data type: hierarchical, time-stamped, associated with an author, community-rated</li> <li>We will build interactive visualizations and a textual summary for each thread</li> </ul>	<h3>Challenges</h3> <ul style="list-style-type: none"> <li>Forum and blog comments offer unparalleled insights into current public sentiment, but their content is hard to consume</li> <li>Populations of interest may not have access to new media</li> </ul>
<h3>Approach</h3> <ul style="list-style-type: none"> <li>First pass: summary of surface features</li> <li>Assess with volunteers working on a typical but proxy analytic scenario</li> <li>Second pass: add deeper linguistic features and repeat assessment</li> </ul>	



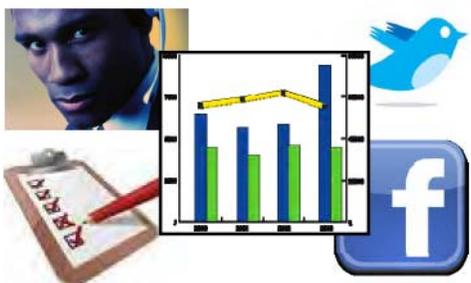

11

© 2009 The MITRE Corporation. All rights reserved

## Public Opinion Polling by Proxy (POP/P)



Dr. John Henderson

<h3>Objectives</h3> <ul style="list-style-type: none"> <li>Establish the feasibility and methodology for calibrating implicit polling performed via new social media to well-understood demographics from traditional polling sources.</li> <li style="border: 2px solid red; padding: 2px;">Develop forecasting software to predict poll outcomes for new polls in new areas.</li> </ul>	<h3>Challenges</h3> <ul style="list-style-type: none"> <li>Public opinion may not be a primary indicator of foreign government actions for some governments.</li> <li>Some poll questions may be so precise that their associated responses cannot be found in social media at a statistically significant rate.</li> </ul>
<h3>Approach</h3> <ul style="list-style-type: none"> <li>Collect from public discourse available in social media and align with known poll results from the same timeframe.</li> <li>Establish latent demographics in the new media.</li> <li>Reweight over- and under-represented categories to ensure forecasted poll responses match desired demographics.</li> </ul>	

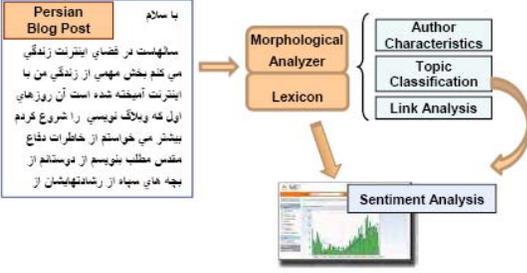



12

© 2009 The MITRE Corporation. All rights reserved



## Exploring Soft Power in Weblogistan

<h3>Objectives</h3> <ul style="list-style-type: none"> <li>■ Develop technology for analyzing Persian language blogs</li> <li>■ Build foundational tools that, combined with existing technology, will give us access to understanding Persian language blogs through:             <ul style="list-style-type: none"> <li>• Topic classification</li> <li>• Author characteristics</li> <li>• Sentiment Analysis</li> </ul> </li> </ul>	<h3>Challenges</h3> <p style="text-align: right; margin-bottom: 0;">Dr. Karine Megerdooian</p> <ul style="list-style-type: none"> <li>■ Citizen media like blogs are a powerful source for understanding culture and beliefs in closed societies but there are several issues:             <ul style="list-style-type: none"> <li>■ Critical shortage of Persian (Farsi/Dari) language speakers</li> <li>■ Current sponsor work is done manually</li> <li>■ Existing Persian language processing systems fail to successfully analyze blogs</li> </ul> </li> </ul>
<div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> <p>Persian Blog Post</p> <p>با سلام سالمهات در قضای اینترنت زندگیا می گذم بخش مهمی از زندگیا من با اینترنت آمیخته شده است آن روزهای اول که وبلاگ نویسی را شروع کردم بیشتر می خواستم از خاطرات دفاع مقدس مطلب بنویسم از دوستم از جبهه های سپاه از رشادتهایشان از</p> </div> 	<h3>Impact</h3> <ul style="list-style-type: none"> <li>■ Enabling Persian language technology for blog analysis will provide actionable intelligence for strategic communication policy in the region</li> <li>■ Enhance analysts' capabilities by detecting influential bloggers and credible posts</li> <li>■ Position MITRE to better respond to immediate sponsor needs</li> </ul>





13  
 © 2009 The MITRE Corporation. All rights reserved.

### 3. University of Arizona, AI Lab – Dark Web Terrorism Research

As part of its Dark Web Terrorism project ([website](#)), the Artificial Intelligence Lab has for several years collected international jihadist forums. These online discussion sites are dedicated to topics relating primarily to Islamic ideology and theology. The Lab now provides search access to these forums through its Dark Web Forum Portal, and in its beta form, the portal provides access to 14 forums, which together comprise nearly 4,000,000 messages. The Portal also provides statistical analysis, download, translation and social network visualization functions for each selected forum.

**Research goal:** The AI Lab Dark Web project is a long-term scientific research program that aims to study and understand the international terrorism (Jihadist) phenomena via a computational, data-centric approach. We aim to collect "ALL" web content generated by international terrorist groups, including web sites, forums, chat rooms, blogs, social networking sites, videos, virtual world, etc.

We have developed various multilingual data mining, text mining, and web mining techniques to perform link analysis, content analysis, web metrics (technical sophistication) analysis, sentiment analysis, authorship analysis, and video analysis in our research.

The group receives funding from a variety of government agencies, including DHS, DTRA, AFRL, and NSF (more detail [here](#)). An example of one of their proposed NSF projects: (PI is Hsinchun Chen, the director of the AI Lab), "[Developing a Dark Web Collection and Infrastructure for Computational and Social Sciences](#)." (See this NSF [article](#) about the project)

*ReadWriteWeb* covered the Lab's research in a 2008 article called, "[Spidering the 'Dark Web'](#)." The article gives a comprehensive overview of the project including such topics as: where terrorists are found on the web; the types of tools the Lab uses in their research including: **website monitoring, forum**

**spidering, multi-media spidering**, social network analysis, content analysis, **web metrics analysis**, authorship analysis and writeprint (a technique created by the Lab which automatically extracts thousands of **multilingual**, structural, and semantic features to determine who is creating 'anonymous' content online), video analysis and IEDs in dark web analysis, etc and their views on privacy. Some interesting facts from the article:

So far, the Dark Web has determined the following:

- Forums: 300 terrorist forums found, some with more than 30,000 members; nearly 1,000,000 messages posted.
- Blogs, social networking sites, and virtual worlds: Many transient sites have been identified before they disappear; more than 30 (self-proclaimed) terrorist or extremist groups in virtual world sites, though they have yet been unable to determine who is just "playing terrorist" vs. who is for real.
- Videos and multimedia content: 1,000,000 images and 15,000 videos from web sites and specialty multimedia file-hosting third-party servers; more than 50% of videos are related to Improvised Explosive Devices.

From a forum [poster](#), the AI Lab's Sentiment Analyzer:

## 2. Search | AZ Forum Portal, Sentiment Analyzer

### AZ Forum Portal Overview

The Arizona Dark Web Forum Portal is a centralized repository of Dark Web forum communications. The goal of the portal is to provide the functionality to browse, search, and analyze the Dark Web forum collections. The current portal contains more than 2.5 million messages across 7 forums.

### Major Functionality

- The Arizona Dark Web Forum Portal provides functionality to aid in the analysis of Dark Web forums within a centralized repository.
  - Forum-level statistics and graphical depiction of posting activity over the forum lifespan
  - Facility to browse forum communications by member, thread, or time period
  - Keyword search of thread titles or message bodies
  - Embedded [Google Translation](#) to facilitate multilingual analysis

### AZ Sentiment Analyzer Overview

The Arizona Sentiment Analyzer is a web-based portal for the analysis of opinions and emotions expressed in Dark Web forum communications, to enhance the understanding of forum communities and participants. For the security analyst, the portal eases the acquisition and consumption of knowledge discovered in sentiment and affect analyses of Dark Web forum communications.

### Major Functionality

- Forum-level statistics characterize sentiment and affect expression of the community, for easy comparison among forums
- Analysis across various dimensions, featuring impactful visualizations
- Temporal analysis by thread or member to identify sentiments and affects expressed at specific time periods
- Search functionality for threads titles or members
- Sentiment and affect regression analysis




Papers from U-Arizona on sentiment analysis (see their publications page [here](#)):

- Abbasi, A., Chen, H. and Salem A (2008). "[Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums](#)," ACM Transactions on Information Systems, 26(3): Article 12. [Full text [paper](#)]

Abstract: ... In this study the use of sentiment analysis methodologies is proposed for classification of web forum opinions in multiple languages. The utility of stylistic and syntactic features is evaluated for sentiment classification of English and Arabic content. Specific feature extraction components are integrated to account for the linguistic characteristics of Arabic. The Entropy Weighted Genetic Algorithm (EWGA) is also developed, which is a hybridized genetic algorithm that incorporates the information gain heuristic for feature selection. EWGA is designed to improve performance and get a better assessment of the key features. The proposed features and techniques are evaluated on a benchmark movie review data set and U.S. and Middle Eastern web forum postings.

The experimental results using EWGA with SVM indicate high performance levels, with accuracy over 95% on the benchmark data set and over 93% for both the U.S. and Middle Eastern forums. Stylistic features significantly enhanced performance across all test beds while EWGA also outperformed other feature selection methods, indicating the utility of these features and techniques for document level classification of sentiments.

- Chen and the Dark Web Team (2008). "[Sentiment and Affect Analysis of Dark Web Forums: Measuring Radicalization on the Internet](#)," in Proceedings of the IEEE International Intelligence and Security Informatics Conference (Taipei, Taiwan, July 17-20, 2008).

#### 4. *Applied Systems Intelligence*

An October 2, 2001 *New Scientist* reported in the [article](#), "[Intelligence analysis software could predict attacks](#)," on a software program that [Applied Systems Intelligence](#) (ASI) was developing at the time called **Knowledge Aided Retrieval in Activity, or KARNAC**.

Intelligence analysis software being developed in the US could be used to predict future terrorist attacks, claims the research company making it. When complete, they say it will be capable of sifting through and analysing existing databases of information, both public and private, and spotting suspicious patterns of activity. ... "We're trying to predict these events before they even happen," he says. The software is called Knowledge Aided Retrieval in Activity Context (KARNAC) and uses "profiles" of different categories of terrorist attacks to seek out key components of possible events. ... Bagdonis says the information for KARNAC would come from both structured and unstructured databases. The former includes gun registrations, driver's licences and criminal records, while the latter would include the internet and newspapers, journals and county records. So, for example, the system might send an alert if someone tried to buy materials that could be used in bomb making, and booked a large truck and a hotel room near a government office. This may seem unlikely, but it is the kind of information that was in fact available on databases before Timothy McVeigh detonated his bomb in Oklahoma City. "These small pieces of information don't have much of an impact on their own, but collectively they can be very important," says Bagdonis. ... Although ASI are reluctant to explain precisely how KARNAC works, Bagdonis admits that reliability is an issue. "I can't claim that this is going to work 100 per cent without a glitch," he says. But the data KARNAC is drawing attention to in tests is the same information that FBI agents have identified as important after an event, he says. Nonetheless, in gaining acceptance, KARNAC may have an even greater obstacle - the realisation since the 11 September that even very smart technology can be rendered impotent by terrorists intent on carrying out previously unimaginable atrocities.

It is hard to say what became of this software because the ASI website no longer makes mention of the product. Their current product is the [CrimeLink Investigative Analysis Software](#). The software webpage does not discuss sentiment analysis using web media per se... Description:

CrimeLink™ is a visual investigative analysis tool designed to assist analysts and investigators with compiling large amounts of seemingly independent and unrelated data into orderly, understandable, and comprehensible graphical analytical products.

By automating time-proven analytical techniques such as Association Matrixes, Link Diagrams, Time Event Charts, Pattern Analysis Wheels, and Telephone Toll Charts, CrimeLink™ is the perfect solution for understanding and solving complex issues involving Counter Terrorism, Organized Crime, Military Operations, Drugs, Financial Fraud, Insurance Fraud, Major Investigations, Retail, and any other set of intricate or conspiratorial activities or social networks.

Key features: Primary aid in predictive analysis. Determine your adversary's intentions and modus operandi. Automates Time-Proven analytical techniques. Visualize and quickly understand complex investigations. Auto-generates all charts from the information - No drawing involved!

ASI describes another one of its software solution (like KARNAC, it is no longer available at its website), Analyst's Associate (paper [here](#)). The company does not specifically say from where they pull their data:

The Analyst's Associate is a revolutionary solution using artificial intelligence to provide automatic detection and tracking of possible criminal/terrorist acts. Analyst's Associate's purpose is to help analysts to weave together seemingly random, disparate pieces of data that are indicative of covert criminal/terrorist activities. It does this by a method predictive analysis of events or activities based on known data and/or present conditions. Core capabilities of the Analyst's Associate include automatic event detection, activity tracking, directed retrieval from widely distributed data sources, and an alert mechanism that notifies users of possible criminal/terrorist activities. This information is automatically presented to the analysts so they can quickly react to the situation. In essence, Analyst's Associate core technology based on causal models, proactively researches, compares, anticipates and predicts future events based upon available information.

Other interesting ASI Whitepapers are available [here](#).

## 5. *Alias-I, Inc.*

**About:** [Alias-i, Inc.](#) was founded by Breck Baldwin 1999 under the name Baldwin Language Technologies. The original funding source was a DARPA research grant under the **Translingual Information Detection, Extraction and Summarization (TIDES) program**. In early 2002, Baldwin Language Technologies began doing business as Alias-i. The next two years brought continued DARPA and DoD funding, primarily to develop the [ThreatTracker](#), advanced information access application designed around the needs of analysts working through a large daily data feed. (now under the Nielson Buzzmetrics name and currently used for company, brand or product protection; product information [here](#)) information extraction application for defense intelligence analysts. In 2004, Alias-i received SBIRs from the U.S. National Library of Medicine (NLM) and NIH. The focus of the grant is extraction of biological entities in biomedical research literature. This grant funded the extensive biomedical natural language processing (NLP) tools and models now found in their software LingPipe.

**Tool:** Their language processing software for text analytics, text data mining and search is called [LingPipe](#). LingPipe's information extraction and data mining tools:

- Track mentions of entities (e.g. people or proteins);
- Link entity mentions to database entries;
- Uncover relations between entities and actions;
- Classify text passages by language, character encoding, genre, topic, or **sentiment**;
- Correct spelling with respect to a text collection;
- Cluster documents by implicit topic and discover significant trends over time; and
- Provide part-of-speech tagging and phrase chunking.

Some other items that may be of interest:

- Alias-i provides a [sentiment analysis tutorial](#) using Cornell researchers Lillian Lee and Bo Pang's movie review dataset. "The high-level idea is to use LingPipe's language classification framework to do two classification tasks: separating subjective from objective sentences, and separating positive from negative movie reviews. In the third section, we show how to build a hierarchical classifier by composing these models."
- See their blog [here](#); Alias-I has an extensive list of their competitors w/summaries on this [page](#).

## Papers / Intellectual property:

- Carpenter, Bob (2007). “[LingPipe for 99.99% Recall of Gene Mentions.](#)” *Proceedings of the 2nd BioCreative workshop*. Valencia, Spain. [Full text [here](#)]
- Carpenter, Bob (2006). “[Character language models for Chinese word segmentation and named entity recognition.](#)” *Proceedings of the 5th ACL Chinese Special Interest Group (SIGHan)*. Sydney, Australia. [Full text [here](#).]
- Carpenter, Bob (2005). “[Scaling High-Order Character Language Models to Gigabytes.](#)” In *Proceedings of the Association for Computational Linguistics Workshop on Software*. Ann Arbor. [Full text [here](#)]
- Carpenter, Bob (2004). “[Phrasal Queries with LingPipe and Lucene.](#)” In *Proceedings of the 13th Meeting of the Text Retrieval Conference (TREC)*. Gaithersburg, Maryland. [Full text [here](#)]
- Carpenter, Bob (2004). “[Orthographic variation with Lucene.](#)” In O. Gospodnetic and E. Hatcher, *Lucene in Action*. Manning Press.

List of third party papers, courses using LingPipe and a patent application [here](#).

## 6. Inxight (an SAP company) Federal Systems Group

**About:** [Inxight Federal Systems](#). Inxight Software, Inc. provides enterprise software solutions for information discovery. Using Inxight solutions, the company says organizations can access and analyze unstructured, semi-structured and structured text to extract key information to enable business intelligence. Inxight claims to be the “only company that provides a complete, scalable solution enabling information discovery in more than **30 languages.**” Customers include enterprise companies such as Novartis, Procter & Gamble and Thomson, multiple U.S. and foreign government agencies, including the DoD, Defense, Defense Intelligence Agency, DHS and Commonwealth Secretariat, and software OEMs such as SAP, SAS, Oracle and IBM. The company has offices throughout the United States and Europe.

**Product:** Access Data From Internal and External Sources Simultaneously: [Inxight SmartDiscovery™ Analysis Server](#) (see also this [overview](#)) is a federated search, clustering and alert solution, allowing users to extract data from a variety of different sources, including classified cable feeds, HUMINT, finished intelligence reports, Message Traffic, open source blogs and newsgroups, as well as Google Enterprise Search, Oracle Secure Enterprise Search, and other search indexes. Users cluster and filter their federated search results by the most relevant people, companies, places, concepts, weapons, vehicles and other things or entities mentioned in them. Users can also cluster and filter results by subject area or by source. Its entity extraction module can be used to extract custom entities, relations and events such as chemical compound names or formulae, phrases for **sentiment analysis** and medication adverse effects. Inxight’s [Thingfinder](#) and other products may also be of interest.

Person	FAWIZ AL (RABBATI), RABBATI, JAN ANTON KRACZEWSKI (AL-KIELBASA), KRACZEWSKI
Vehicle	TEN 1-TON TRUCKS, FOUR-DOOR 1984 GREEN SUBARU
Person_Common	SMUGGLERS, ELECTRICIAN
Measure	10-15 KILOMETERS, 150 CENTIMETERS (CM)
City	KHASON
Weapon	SOPHISTICATED BOMBS
Buy Artifact	FAWIZ AL (RABBATI) PURCHASED TEN 1-TON TRUCKS (NFI)
Travel across Border	SMUGGLERS TO CROSS THE BORDER APPROXIMATELY 10-15 KILOMETERS OUTSIDE OF KHASON
Recruit	RABBATI RECRUITED JAN ANTON KRACZEWSKI ((AL-KIELBASA))
Person Appearance: Age	KRACZEWSKI IS APPROXIMATELY 53 YEARS OLD
Person Appearance: Height	KRACZEWSKI IS 180 CENTIMETERS (CM) TALL
Person Attributes: Vehicle	HE (KRACZEWSKI) DRIVES A FOUR-DOOR 1984 GREEN SUBARU
Make Artifact	KRACZEWSKI USES HIS BACKGROUND AS AN ELECTRICIAN TO CREATE SOPHISTICATED BOMBS

**Applications** include (More detail [here](#)):

- Data Mining and Filtering

- **Link Analysis**

Another Homeland Security organization using Inxight SmartDiscovery also relies on message traffic information to support counter terrorism. This organization receives hundreds of messages per day on topics relating to national security. Relying on the tools provided by SmartDiscovery, this organization is able to process and analyze its vast amount of message traffic to help discover links between certain events, people and situations. For obvious reasons, policy makers and government officials need this information in a timely fashion. SmartDiscovery provides immediate response, automating a previously time-consuming manual process and enabling better, faster decisions to be made.

- Data Navigation
- Information Extraction
- Data Analysis

Inxight ThingFinder is being used by one government/intelligence organization to identify key differentiating factors of a criminal's personal history: facial features, tattoos, scars, piercings, etc. The organization is also using ThingFinder to determine criminal information including places of residence, past crimes committed, sentencing records, and many other key attributes. When individual criminal information is stored and made readily available to nation-wide law enforcement offices, felons - as they cross state lines - can be more easily found and detained. Another government organization is using Inxight VizServer with StarTree to visually display inference analysis results. Both explicit and implicit relationships are determined using various additional software tools including Inxight SmartDiscovery. In this application, information regarding terrorists and the organizations they belong to can be extracted, correlated and then visually displayed. When an organization and its members are listed, other organizations that these individuals belong to are quickly associated and visually displayed. This capability is critical to pinpointing people and activities that may pose a threat to the American population.

**Customers** include: Commonwealth Secretariat, DIA, European Patent Office, IRS, NASA, National Cancer Institute, United Combatant Command, Scottish Enterprise, US Air Force, Army, Navy, USDA, DoD, DHS, DOJ, DEA ([source](#)).

## H. Other Interesting Sentiment Analysis Activities

### 1. *Web Ecology Project*

**About:** The [Web Ecology Project](#) is an interdisciplinary research group based in Boston, Massachusetts focusing on using **large scale data mining to analyze the system-wide flows of culture and community online**. "In addition to the task of understanding culture on the web through quantitative research and rigorous experimentation – we're attempting to build a science around community management and social media. To that end, we're involved in building tools and conducting research that enable planners to launch data-driven campaigns backed by network science."

The WEP conducted a recent sentiment analysis study, "[Detecting Sadness in 140 Characters: Sentiment Analysis and Mourning Michael Jackson on Twitter](#)" (authors are Elsa Kim and Sam Gilbert, with Michael J. Edwards and Erhardt Graeff), August 18, 2009. **This study analyzed Twitter data using the ANEW dataset** (read the full text report [here](#)).

Executive Summary: Michael Jackson’s death created an emotional outpouring of unprecedented magnitude on Twitter. In this report, we examine 1,860,427 tweets about Jackson’s death in order to test various methods of sentiment analysis and gain insights into how people express emotion on Twitter.

Key findings:

- At its peak, the conversation about Michael Jackson’s death on Twitter proceeded at a rate of 78 tweets per second.
- Users tweeting about Jackson’s death tend to use far more words associated with negative emotions than are found in ‘everyday’ tweets.
- Roughly 3/4 of tweets about Jackson’s death that use the word “sad” actually express sadness, suggesting that sentiment analysis based on word usage is fairly accurate.
- That said, there is extensive disagreement between human coders about the emotional content of tweets, even for emotions that we might expect would be clear (like sadness).
- Tweets expressing personal, emotional sadness about the Jackson’s death showed strong agreement among coders while commentary on the auxiliary social effects of Jackson’s death showed strong disagreement.
- We argue that this pattern in the “understandability” of certain types of communication across Twitter is due to the way the platform structures the expression of its users.

Other Web Ecology studies include:

- [Afghanistan and its Election on Twitter: The Macro Picture \(preview\)](#) September 11, 2009 - By Erhardt Graeff with Seth Woodworth
- [The Influentials: New Approaches for Analyzing Influence on Twitter](#) September 2, 2009 - By Alex Leavitt with Evan Burchard, David Fisher, & Sam Gilbert [The Influentials \(pdf\)](#). [10 Days of Influence Tracked by Density of Responses \(2993.27 KB jpg\)](#). See also this blog article, “[Wow – Mashable is more influential than CNN](#),” describing the project, [here](#).
- [Reimagining Internet Studies: A Web Ecology Perspective](#), August 10, 2009 - Contributing scholars (in alphabetical order) were: Jonathan Beilin, Bill Bushey, Patrick Davison, Sam Gilbert, Erhardt Graeff, Tim Hwang, Sawyer Jackson, Elsa Kim, Alex Leavitt, AJ Mazur, Dharmishta Rood, Mike Rugnetta, Frank Tobia, and Seth Woodworth. [Reimagining Internet Studies](#) (pdf)
- [The Iranian Election on Twitter: The First Eighteen Days](#) June 26, 2009 - Contributing researchers: Jonathan Beilin, Matt Blake, Mac Cowell, David Fisher, Sam Gilbert, Russell Hanson, Tim Hwang, Alex Leavitt, Greg Marra, Rob Mason, Colin McSwiggen, Dharmishta Rood, Aaron Shaw, Frank Tobia, and Seth Woodworth. [The Iran Election on Twitter](#) (pdf)

## **2. CERATOPS (University of Pittsburgh, University of Utah, Cornell University)**

CERATOPS, the Center for the Extraction and Summarization of Events and Opinions in Text (website [here](#); publications page [here](#)) says it is “dedicated to developing accurate and robust techniques for extracting and summarizing information about events and beliefs from free text.” The goals of their research effort are threefold.

- (1) We will create easily trainable learning algorithms that can automatically create domain-specific patterns to identify facts and relations associated with relevant events, such as infectious disease outbreaks.
- (2) We will develop trainable learning algorithms that can distinguish factual assertions from subjective (non-factual) assertions, identify beliefs that are held by an entity, and assess the intensity, polarity, and motivation and attitude types of those beliefs.
- (3) We will create methods for understanding event and belief progressions over time.

CERATOPS is a University Affiliate Center (UAC) to the [LLNL's] Discrete Sciences Institute ([fact sheet](#)). The project is **funded by the Department of Homeland Security** as part of the UAC program to fund basic research and education, and to foster research collaborations among universities and the National Laboratories.

Universities participating in this program include (more project detail [here](#)):

- **University of Pittsburgh:** *Project:* Detecting Opinion and Sentiment Types in the News and on the Web to Improve Automatic Question Answering and Information Extraction. This research group has been studying the expression of different types of opinions, in the news and the web to support automatic question answering (QA). (**Jan Wiebe** is in charge of the U Pitt's CERATOPs project; see her homepage [here](#)).
- **University of Utah:** The [Natural Language Processing](#) group at the University of Utah has developed several new methods for extracting factual information from unstructured text. Their new IE techniques have achieved good results in the domains of Latin American terrorism, using the MUC-4 corpus, and infectious disease outbreaks, using a text collection of ProMed-mail articles. In a second line of investigation, they have been exploring a different approach to information extraction that decouples the tasks of finding relevant regions of text and applying extraction patterns (**Ellen Riloff** heads U-Utah's CERATOPs project; see her homepage [here](#).)
- **Cornell University:** *Project:* Extraction of fine-grained opinion frames. The CERATOPs UAC focuses on the development of methods for extracting and summarizing expressions of opinion that appear in digital text. *Project:* Productization of opinion summarization software. Over the past year, there has been industry interest in CERATOPs research in the area of summarizing fine-grained opinions. Entrepreneur Larry Levy formed and financed a company, Jodange LLC, to **develop opinion summarization systems for the financial services industry**. The company is based in Yonkers, NY, has five employees, and was selected to launch its initial product in January 2008 at the DEMO08 conference. (**Claire Cardie** leads the Cornell CERATOPs projects and is also chief scientist and co-founder of **Jodange**. Cardie's homepage at Cornell is [here](#); Cornell's Natural Language Processing website is [here](#)). In this overview article about the Cornell NLP group, Claire Cardie says:

Indeed, the Cornell Natural Language Processing group has done seminal work in developing algorithms for sentiment classification and extraction problems, and its research has been widely recognized in the research community and in the scientific popular press as being, in large part, responsible for the recent huge surge of interest in the area (more [here](#)).

The *New York Times* highlighted CERATOPs in an October 2006 [article](#), "Software Being Developed to Monitor Opinions of U.S."

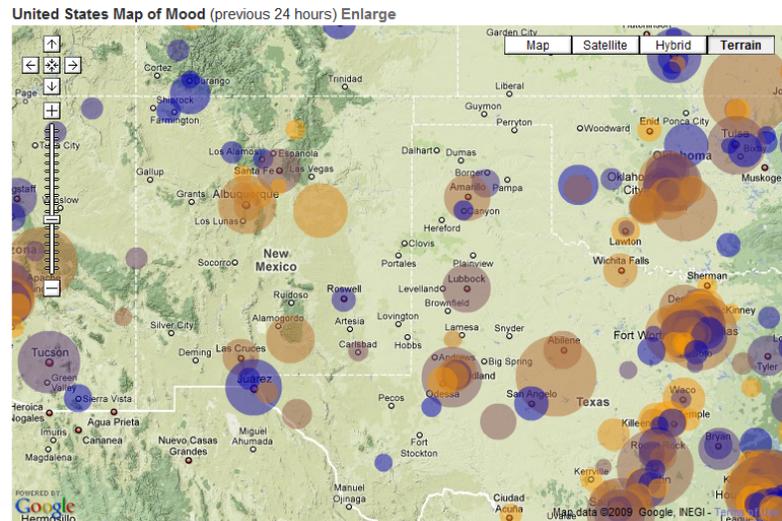
### 3. *Twittermood (Northeastern University)*

**About:** [Twittermood](#) is a [project](#) of Albert-László Barabási's lab, the **Northeastern University Center for Complex Networks Research**. The webpage was created by a group of the Lab's researchers, Yong Yeol Ahn, James Bagrow, Sune Lehmann and Alec Pawling.

**Twitter Data:** The twitter data is collected from the twitter gardenhose, which is stream containing a significant sample of all public twitter statuses. The gardenhose supplies about 1,800,000 messages per day ([source](#)).

**Mood:** We estimate the mood of all tweets for which the author has provided a location. The mood (valence) of each individual tweet is based on the Affective Norms for English Words (ANEW) data set, calculated as suggested in ["Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents." *Journal of Happiness Studies*, DOI: 10.1007/s10902-009-9150-9]. ANEW is a set of 1034 words, previously identified as bearing emotional weight (e.g. abuse, acceptance, accident). The mood of each word was estimated in a study at the University of Florida, where participants (college students), were shown lists of isolated words and asked to grade each word's valence, arousal, and dominance level on an integer scale of 1-9.

**Location:** The location of each tweet is based on the self-reported location of each twitterer. In case a latitude and longitude is provided, this information is used. If a written location is provided (e.g. Boston, MA), the string is geocoded using the [Google maps API](#).



**Presentation:** The size of each circle is proportional to the logarithm of the number of tweets recorded at the location at the center of the circle. The color (from blue to yellow) of each circle is a function of the mood at the center, with yellow corresponding moods above average and blue below average. Circles are drawn at locations with at least 5 tweets.

#### 4. EU-funded Programs

The European Union funded two programs of possible interest, referred to as INDECT and ADABTS

...[Indect will] develop computer programmes which act as "agents" to monitor and process information from web sites, discussion forums, file servers, peer-to-peer networks and even individual computers. Its main objectives include the "automatic detection of threats and abnormal behaviour or violence". Project Indect, which received nearly £10 million in funding from the European Union, involves the Police Service of Northern Ireland (PSNI) and computer scientists at York University, in addition to colleagues in nine other European countries. ... "Our focus is on novel techniques for word sense induction, entity resolution, relationship mining, social network analysis [and] sentiment analysis," it says.

A separate EU-funded research project, called Adabts - the Automatic Detection of Abnormal Behaviour and Threats in crowded Spaces - has received nearly £3 million. It is based in Sweden but partners include the UK Home Office and BAE Systems. It is seeking to develop models of "suspicious behaviour" so these can be automatically detected using CCTV and other surveillance methods. The system would analyse the pitch of people's voices, the way their bodies move and track individuals within crowds ([source](#)).

## I. Intellectual Property and Sentiment Analysis

According to a PriceWaterhouse [report](#) on technology trends, the number of US patents for **customer analytics / customer-centric technology** has **increased sevenfold** over the past nine years, and **125% over the past four years** in comparison to a gradual decline of 13.9% in the overall number of patents during the same period. Listed below are patents and patent applications that turned up using a keyword search at the US Patent and Patent Applications Office sites of the phrase “sentiment analysis.” What is interesting is that very few of the companies selling products in the sentiment analysis space appear to be doing so based on patentable intellectual property.

### Textmap / Stonybrook University Patent Application:

[20080270116](#) Large-Scale Sentiment Analysis

### Microsoft Patent Applications

[20080249764](#) Smart Sentiment Classifier for Product Reviews

[20080215543](#) Graph-based search leveraging sentiment analysis of user comments

[20080215571](#) Product review search

[20090037401](#) Information Retrieval and Ranking

[20090077069](#) Calculating Valence of Expressions Within Documents For Searching A Document Index

[20090083096](#) Handling product reviews

[20090106222](#) Listwise Ranking

## IBM

### Patent Applications

[20090138276](#) Privacy management system using user's policy and preference matching

[20090182554](#) Text analysis method

### Patents

[7,475,007](#) Expression extraction device, expression extraction method, and recording medium

[7,536,637](#) Method and system for the utilization of collaborative and social tagging for adaptation in web portals

[7,130,777](#) Method to hierarchical pooling of opinions from multiple sources

### Yahoo! Patent Application

[20090164557](#) User vacillation detection and response

Abstract: An embodiment of the present invention automatically detects when a user is in a state of vacillation based on user on-line behavior, records relevant parameters regarding the vacillation event, and then responds accordingly. This response may include providing relevant and/or targeted information that can be used by the user to help remove the indecision. The response may also or alternatively include providing third-party businesses, such as retailers, marketers, and advertisers, with information about vacillation events and associated behaviors for a single user or groups of users so that such businesses can identify potential markets/customers or directly engage similar users to facilitate the decision-making process.

**Psydex Corporation Patent Application**

[20080208820](#) Systems and methods for performing semantic analysis of information over time and space

**Technorati, Inc.**

[20080228695](#) Techniques for analyzing and presenting information in an event-based data aggregation system

**Google**

[20080243780](#) Open profile content identification

[20090193011](#) Phrase Based Snippet Generation

**PNNL**

[20080306899](#) Methods, apparatus, and computer-readable media for analyzing conversational-type data

**Northwestern University**

[20080313130](#) Method and System for Retrieving, Selecting, and Presenting Compelling Stories from Online Sources

**University of Illinois, Urbana Champaign**

[20090048823](#) System and methods for opinion mining

## II. Forecasting and Prediction

Some companies and institutions claim to be using web data for prediction. We have found that such statements may be stretching the concept of prediction in some cases. As noted in a [JASON report](#) both **predictive** and **insight** models are used by analysts (whether they be in business or in the IC) when addressing hard problems. Insight models "build expert intuition," systematize thinking about gaps in knowledge, and/or enhance human pattern recognition. In other words, these tools assist a human being in making a more informed or better prediction or forecast, and don't actually supply a concrete prediction. There is something of a continuum between insight and prediction, and it is arguable that many of the items in this chapter are not, strictly speaking, predictive. However they are generally more predictive than the items in Chapter I.

Vendors that provide financial firms with stock market advice appear to be almost the only companies that are willing to claim that they do web forecasting or prediction (more on that below). A handful of other companies and university researchers publically state that they either conduct research on or assist clients with web prediction.

- [SentiMetrix](#) (profiled above) comments on its website that, "Human intelligence and analysis has never been more mission critical to homeland security. The staggering volume of data in the blogosphere, news sources and internal databases necessitate some automation. By tracking hot topics in real time and in multiple languages, **your organization will be ready to predict future scenarios and be ready before they happen.**" The areas in which they say they provide automated service include: supporting counterterrorism efforts, identifying opinions towards the US, US policies and initiatives, tracking opinions towards "friends" of the US, and **anticipating election results** and implications (more detail [here](#)).
- Researchers N. Godbole, M. Srinivasaiah of **Google** and S. Skiena of **SUNY** are using predictive sentiment analysis in their research. In the paper, "[Improving Movie Gross Prediction Through News Analysis](#)," SUNY's W. Zhang and S. Skiena used quantitative news data generated by Lydia, their system for large-scale news analysis, to **help people to predict movie grosses** (full text [here](#)). [TextMap](#) is an entity search engine created by these researchers that graphs sentiment analysis and is certainly worth a look. See also [TextMed](#), [TextBlog](#) and [TextBiz](#). *Note: The TextMap web page freezes up when clicking on specific entities.* See a list of Lydia/TextMap publications [here](#).
- [Gravity Technologies](#) (Budapest, Hungary), a recommendation engine, "specializes on the key area of finding the best correspondence between personal taste and other user groups taste by sophisticated mathematical algorithms. In other words, we are developing recommendation systems that substantially **improve the accuracy of the prediction** how much a user is going to like a product."

## A. Applications

### 1. Stock Market Prediction

#### Industry Activity

Financial firms have been leveraging sentiment analysis to forecast stock prices (on the theory that they fluctuate in part based on emotion). For instance, to assess "buzz" **Jodange** processes vast amounts of data to determine an aggregate "opinion momentum." Jodange experimentally analyzed all the mentions of mortgage lender, Countrywide Financial Corporation over two years, using hundreds of thousands of mentions in blogs, tweets, news articles, TV and radio. A banking news article reports that:

The experiment took only a day, and the result was that Jodange could correctly predict the direction of Countrywide's stock the next day 70% of the time... Jodange can also isolate major influencers to see, for instance, what publications have the most effect on certain stocks. It found that when the Web magazine *Seeking Alpha* opined on Countrywide, the stock reliably swung in one direction or another, but that wasn't the case with other stocks, such as Fannie Mae. Knowing the key influencers in the Websphere could help a company know who to talk to, [says Larry Levy, CEO of Jodange] ([source](#)).

Jodange is also working on a new algorithm that could use opinion data to **“predict future developments, like forecasting the impact of newspaper editorials on a company’s stock price.”** ([source](#))

Using data obtained from newswire services, leading financial websites and blogs, in addition to historical information, companies such as [Ravenpack](#), [Fin-buzz](#), [IVolatility](#), [UnderstandMarket](#), [ParSOS Finance & IT](#) (Germany), [LifeTips](#) and [SentimenTrader](#) offer tools to “provide insight into market sentiment.” About actual stock prediction, Fin-buzz makes the telling comment: “Of course it is **impossible to predict the future** – anybody saying otherwise is living in a fool's paradise – but as this example [referring to a Lloyds Banking Group sentiment chart] demonstrates it is possible to assess the probability of likely outcomes based on **key insight** ([source](#)).

The TextBiz [website](#) (see mentioned above) performs analysis of the NYSE, NASDAQ and AMEX stock market data and it is updated weekly with the latest results. The current focus is in the following areas:

**Random Walk Modeling:** The probability distribution for the individual stock prices is calculated and compared to current performance. Min/Max and Average distribution graphs are drawn for the current year, as well as for the stock history. Past years are color-coded to provide an at-a-glance view of annual performance. If you do not know what a random walk is, read a short theoretical description, plus have a look at an interactive applet.

**Default Prediction:** Currently, the Z-score method of default prediction is used to determine how “healthy” each company appears.

**Fundamental Analysis:** The financial statements of each company are used to calculate measures that show whether there is a promising investment opportunity.

Thompson Reuters also supplies some stock-related prediction services that incorporate web data.

(September 2008) Richard Brown of Thomson Reuters delivered an illuminating talk, “News, Blogs, and Full-Tick Logs: Innovative Approaches to Quantitative and **Event-Driven Trading**,” Tuesday at Gartner's Event Processing Summit. The summit and the Event Processing Technical Society symposium now underway feature many such use cases, descriptions of low-latency transformation and analysis of high-volume data and event streams as applied to diverse business problems. Brown's case study, which looked at **exploiting information from unstructured sources to support financial-market trading, was of particular interest due to its combination of events, text sources, and sentiment analysis.**

Brown's talk profiled the [Reuters NewsScope Sentiment Engine](#), which "processes a stream of Reuters English language news items, producing sentiment data for a list of customer determined target companies." You can read more on line, including about the application of **technology from Infonic to assess author sentiment**. The application assigns sentiment scores to selected words and phrases but goes beyond simple numbers, per Brown's presentation, to assess **topical relevance** — Is a given item substantively about the subject company — and novelty — How unique (and fresh) is the news? In the financial sector, it is especially important to distinguish updates from *new* news. Brown was speaking at an [event processing conference](#) because "integration of myriad of data sources requires a sophisticated event processing framework," that is, one that delivery low-latency (low-delay) processing of high-volume data streams from diverse sources. Complex event processing's *raison d'être* is the inability of traditional data warehousing and analytical methods to accommodate these needs. Thomson Reuters works with CEP vendor [Streambase](#).

Beyond scoring the author sentiment and market sentiment, "reaction based on information in the article/text," Reuters provides meta information that support **article aggregation to company, sector, and market levels** and has been studying the **trading value implied by news propagation pathways**. As in other applications of text data mining, findings are combined with independent analytical processes. Since Reuters operates in a trading domain, these include technical analysis of financial instruments and study of company and market fundamentals.

Part of the **analysis is generation of 45 predictive indices that the company**, working with quantitative research firm [AlphaSimplex](#), determined **apply to foreign-exchange and stock volatility**. The point is to assess **event-based risk**. The **violence index**, for example, represents the percentile of violence topicality relative to "comparable" historical periods. A next phase of research will study applicability to equities, for instance, how a spike in the violence index affects defense stocks (more substantive information on the Web page of principal [Andrew W. Lo, director of MIT's Laboratory for Financial Engineering](#)) ... There are additional analytics touchpoints. Reuters deploys subsidiary ClearForest's text-analytics software at client sites to support information extraction ([source](#)).

## Academic Papers

- In the paper, "[SOPS: Stock Prediction using Web Sentiment](#)," V. Sehgal and C. Song of University of Maryland introduced a "novel way to do stock market predictions based on past performance of stocks," by scanning financial message boards and extracting sentiments expressed by individual authors. They found that **they were able to predict sentiment with "high precision" (81% accuracy)** and also show that stock performance and its recent web sentiments are closely related ([here](#)). They concluded that their results "showed promising prospects for automatic stock market predictions using web sentiments."
- Researcher Uta Hellinger of U-Karlsruhe wrote the 2008 paper "[Event and Sentiment Detection in Financial Markets](#)." [Full text [here](#)]

Abstract: Today, traders in financial markets are confronted with the problem that information is distributed over diverse sources and that there is too much information available. In our work we develop methods and tools to help traders to overcome this information overload by enabling the integrated view on news from various sources, by filtering relevant news and by providing decision support for traders. Another goal of our work is to propose a formal model of the impact of news on asset prices and thus **enable** better predictions of stock prices than possible with purely text mining based approaches.

- Devendra Tayal of the Indira Ghandi Institute of Technology, in 2009 writes in "[Comparative Analysis of the Impact of Blogging and Micro-blogging on Market Performance](#)." [Full text [here](#)]

Abstract: The general perceptions about a product and the reputation of the company determine to a great extent how well the product sells. It is thus imperative that we make efforts to understand the public opinions and sentiments, as they can be a very good indicator of the product's future sales performance. In this paper, we explore the two

most common online media which have been used by the public to express such type of subjective content: Blogs and Micro-blogs. We perform comparative analysis of the predictive power of the two media to know which of these formats can prove to be a more useful representative of sentiments to an autonomous stock price prediction system.

## 2. Customer Behavior Prediction

**Teragram** (a company owned by SAS) has a large U.S. bank client that uses the SAS Sentiment Analysis Manager (SAM) to assess customer calls, emails and faxes pertaining to its loan modification program. By assessing the language used by the customer in each exchange, the bank can filter out those who don't qualify for a loan modification, and to rank which customers it must attend to first in order to prevent attrition. "It's a way to handle the sheer volume of inquiries they're getting," says SAS rep Manya Mayes. SAS says that SAM is "the industry's first system combining a statistical method for computing reviews as well as a rules-based approach that lets brand managers evaluate certain specific terms and syntaxes ([source](#))."

SAM was launched in the third quarter of 2009. "Teragram's SAM has a hybrid approach that essentially looks at using a statistical and linguistic approach to sentiment analysis," says Manya Mayes, SAS's chief text-mining strategist. "We're not aware of anyone else in the industry who's doing that." The starting price for SAM is \$160k ([source](#)). The company claims that its NLT technologies "recognize and analyze more than **30 languages** ([source](#))."

**SPSS**: Another example of a financial institution using sentiment analysis is the **Navy Federal Credit Union**, the largest credit union with \$40 billion in assets, 3.3 million members and 179 branches worldwide. SPSS' software **Predictive Analytics** Software suite (PASW) helps analyze the unstructured data in the comment fields of the credit union's frequent customer surveys. In September, Nucleus Research gave Navy Federal Credit Union a 2009 Technology ROI award, saying that the credit union recorded a 1,531 % return on its technological investment in a two-month period, with the annual benefit equating to almost \$1.5 million (SPSS press release [here](#)). See this overview [article](#) and SPSS [presentation](#) describing this case study. SPSS reports that its software processes **multilingual sentiment** event and process extraction in **seven languages** integrates with Japanese in other products, and uses translations via Language Weaver add-on options ([source](#)).

## 3. Identifying "Influentials"

"By knowing in advance who the influencers are for your brand, you'll be better prepared to manage crisis and opportunity effectively, reaching out to 20 key contacts instead of 10,000 questionable ones". ([source](#))

The influentials hypothesis argues that a small subset of individuals carry, by means of their connections, reputation, wealth, and/or other factors, outsized influence in general population decisions. The hypothesis is very attractive to industry in that companies can theoretically reduce cost of marketing by focusing primarily on this very small group of taste-makers. At some level the behavior is certainly present – for instance, anything read by Oprah Winfrey is almost

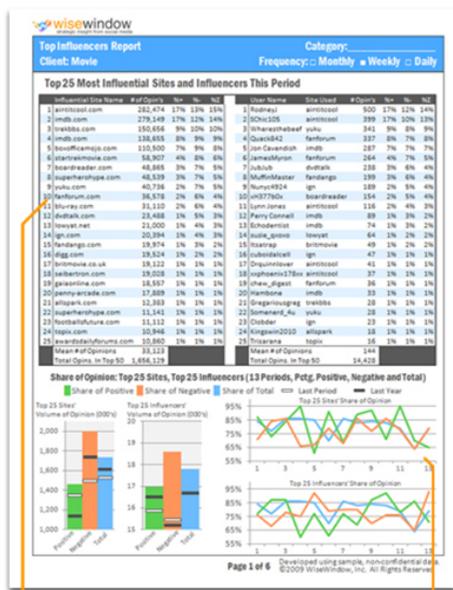
guaranteed a slot on best-seller lists – but it appears to be an open question whether the behavior is widespread. Questions aside, money is certainly being spent based on the theory. One marketing news journal notes that more than \$1 billion is spent per year on word-of-mouth campaigns targeting influentials, and estimates that this spending is growing at 36% a year (much faster than any other part of marketing and advertising) ([source](#)). While these tools are not really predictive in nature, we believe they should be of interest to the NGC.

### Companies

Providing insight on how their clients' brands are perceived by their key influencers appears to be a fairly common occurrence in the commercial industry. One company targeting identification of Influentials is:

**Sysomos**, which claims 87% sentiment analysis accuracy in their product, [MAP](#), performs Influencer ranking and produces Influencer Lists. Besides identifying influencers, Sysomos says it can map influencers – the map can be added in real time and altered while the pages automatically refresh. Because information is collected and stored atomically, extremely rich reporting can be done, including geo-demographic profiling, isolating influencers by city and state level with a high degree of confidence – this seems to be entirely missing from the capabilities of anything else [one reviewer] has worked with ([source](#)).

### Influencers Report



Lists the most influential websites and individual users whose opinions are linked to, quoted and discussed by others

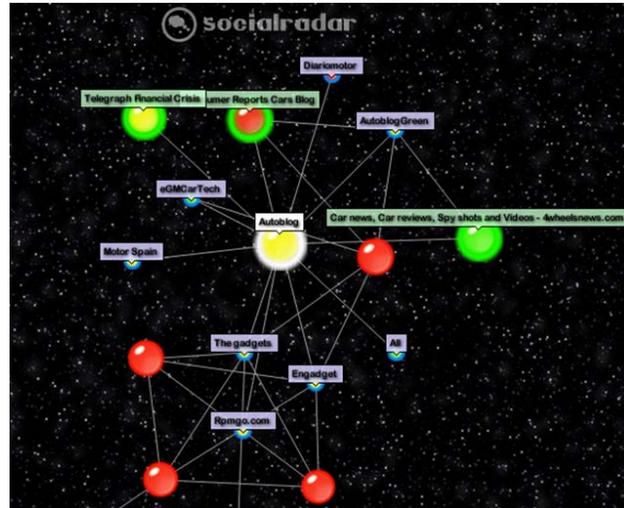
Aggregate statistics for the top 25 influencers against the rest of the category

[WiseWindow](#), provides a syndicated “Influencers: sites and users” report to its customers (see graphic on left). No other information was available about the report at the website (more about this company [above](#)).

Other examples of companies claiming to mine information about influencers include:

- **Jodange**, in addition to evaluating sentiment, also identifies **influential opinion holders** enabling “decision makers to better understand who and what is influencing their customers, competitors, and marketplace” ([source](#) [here](#); see their gadget [here](#)) “There is no longer the notion that trusted information only comes from *The New York Times*,” [says Larry Levy, Jodange CEO] “Once you get a handle on who is influencing your brand, that becomes actionable ([source](#)).”
- **Radian6** has an [Influencer widget](#). CEO of Radian6, Marcel LeBrun, estimates that Radian6 customer Dell has 8,000 to 10,000 online conversations about its brand each day, which span the spectrum of positive to negative; the company needs to understand whose opinion actually has the power to move brand perception, and keep close tabs on those. “Sentiment analysis needs to be connected to social metrics and influence analysis to make sense,” says LeBrun ([source](#)).
- **Attensity’s Cloud** (powered by Radian6) “enables real-time active monitoring of social media to identify trends, **influencers** and sentiment, providing valuable assistance in crisis detection and management, outreach measurement, and more.”
- **Clarabridge Social Media Analysis (SMA)** software solution uses “Techrigy’s warehouse of nearly 3 billion social media mentions from all social media platforms, such as blogs, Facebook, Twitter, YouTube, and MySpace,” claims to **rank influencers** with its product.
- **General Sentiment**, Stonybrook U spinoff, (and the original architects of Textmap technology) says that its product enables customers to scour data, measure sentiment, and “make this information actionable with an exclusive engaged network of **Industry Influencers**.” Example of their dashboards [here](#).
- Well-known PR company, **Ogilvy**, purportedly has “technology and methodology for **identifying and engaging influencers and activating networks of people** to share and recommend products, services and issues. We create engaging experiences designed to promote awareness, brand loyalty, advocacy and conversion.” Discussion [here](#).
- **BuzzStream** for PR and Social Media is an “integrated Social CRM and social media monitoring tool that helps you find your top **word-of-mouth influencers (journalists, bloggers, and microbloggers)** and manage all aspects of your communications with them.” See their products [here](#).

- **Infegy**'s uses their product [Social Radar](#) to identify influential people, blogs, etc. Here's one example from their [blog](#) where they identify the top influencers on Land Rover during the first few weeks in April 2009. Using Social Radar, they created the visual ecosystem below showing some of the most influential sources and how they connected.



Further reading about the science of Influentials:

- [UMBC Ebiquty Research Group](#) has an interesting 80-page slide show ([here](#)) entitled, "[The Game Theoretic Web](#)."
- See the Web Ecology paper in the section [above](#)

## Academics

Below is a list of selected papers Perspectives uncovered using a keyword search for "sentiment analysis" and "influentials."

### Papers:

- Rumi Ghosh, Kristina Lerman (USC) (2009), "[Leaders and Negotiators: An Influence-based Metric for Rank USC Information Sciences Institute](#)," Proceedings of the Third International ICWSM Conference [Full text [here](#)]

Excerpted abstract: We propose influence as a measure of the centrality of nodes in a network. Influence takes into account not only direct links but also all paths between nodes. We parametrize the influence metric by a variable  $\alpha$  that measures the strength of links. ...

- Duncan Watts (Columbia University) and Peter S Dodds (University of Vermont) (2007), "[Influentials, Networks, and Public Opinion Formation](#)," *Journal of Consumer Research*, Vol. 34 [Abstract]

Excerpted abstract: Here we examine this idea, which we call the "influentials hypothesis," using a series of computer simulations of interpersonal influence processes. Under most conditions that we consider, we find that large cascades of influence are driven not by influentials but by a critical mass of easily influenced individuals. ...

- Sebastiano A. Delre, (Bocconi University, Italy) Wander Jager, Tammo H. A. Bijmolt, (University of Groningen), Marco A. Janssen (Arizona State University). (2010) "Will It Spread or Not? The Effects of **Social Influences** and Network **Topology** on Innovation Diffusion." *Journal of Product Innovation Management* 27:2, 267-282. [[Abstract](#)]

Excerpted abstract: Innovation diffusion theory suggests that consumers differ concerning the number of contacts they have and the degree and the direction to which social influences determine their choice to adopt. To test the impacts of these factors on innovation diffusion, in particular the occurrence of hits and flops, a new agent-based model for innovation diffusion is introduced. This model departs from existing percolation models by using more realistic agents (both individual preferences and social influence) and more realistic networks (scale free with cost constraints). Furthermore, it allows consumers to weight the links they have, and it allows links to be directional. In this way this agent-based model tests the effect of VIPs who can have a relatively large impact on many consumers....

- Masahiro Kimura, Kazumi Saito, Ryohei Nakano, Hiroshi Motoda (Ryukoku University, University of Shizuoka, Chubu University, Osaka University, Japan). (2010) "Extracting **influential nodes** on a social network for information diffusion." *Data Mining and Knowledge Discovery* 20:1, 70-97, [[Abstract](#)]

We address the combinatorial optimization problem of finding the most influential nodes on a large-scale social network for two widely-used fundamental stochastic diffusion models. The past study showed that a greedy strategy can give a good approximate solution to the problem. However, a conventional greedy method faces a computational problem. We propose a method of efficiently finding a good approximate solution to the problem under the greedy algorithm on the basis of bond percolation and graph theory, and compare the proposed method with the conventional method in terms of computational complexity in order to theoretically evaluate its effectiveness. The results show that the proposed method is expected to achieve a great reduction in computational cost. We further experimentally demonstrate that the proposed method is much more efficient than the conventional method using large-scale real-world networks including blog networks.

- IBM paper (2010), "Method and System of Alert Stream Mining Based on Sentiment Analysis," [Full text [here](#)]
- Jan Kratzer and Christopher Lettl. (2009) "Distinctive Roles of Lead Users and Opinion Leaders in the Social Networks of Schoolchildren." *Journal of Consumer Research* 36:4, 646-659 [[Abstract](#)]
- Joshua Payne, Peter Dodds, Margaret Eppstein (University of Vermont). (2009) "Information cascades on degree-correlated random networks." *Physical Review E* 80:2. [[Abstract](#)]
- Jacob Goldenberg, Sangman Han, Donald R Lehmann, Jae Weon Hong (SungKyunkwan University, Korea; Columbia University, Dongseo University, Korea). (2009) "The Role of Hubs in the Adoption Process." *Journal of Marketing* 73:2, 1-13 [[Abstract](#)]
- C. de Kerchove, G. Krings, R. Lambiotte, P. Van Dooren, V. Blondel (Université catholique de Louvain, Belgium; Imperial College London, UK). (2009) "Role of second trials in cascades of information over networks." *Physical Review E* 79: [[Abstract](#)]
- James P. Gleeson (University of Limerick, Ireland). (2008) "Cascades on correlated and modular random networks." *Physical Review E* 77:4 [[Abstract](#)]
- [Presentation](#) by Duncan Watts and Peter Dodds, "The Accidental Influentials."

#### 4. Predicting Movie Grosses from Blog Traffic

Predicting movie grosses from blog traffic is something many researchers are attempting to solve. Below, are listed several papers on the topic.

- [GalaxyAdvisors](#), spinoff company of U-Cologne and MIT: “Predicting movie success and academy awards through sentiment and social network analysis.” Krauss, J.; Nann, S.; Simon, D.; Fischbach, K.; University of Cologne and Gloor, P., MIT, date of paper unknown, (Full text [here](#)).

Abstract: This paper introduces a new Web mining approach that combines social network analysis and automatic sentiment analysis. We show how weighting the forum posts of the contributors according to their network position allow us to predict trends and real world events in the movie business. To test our approach we conducted two experiments analyzing online forum discussions on the Internet movie database (IMDb) by examining the correlation of the social network structure with external metrics such as box office revenue and Oscar Awards. We find that discussion patterns on IMDb predict Academy Awards nominations and box office success. Two months before the Oscars were given we were able to correctly predict nine Oscar nominations. We also found that forum contributions correlated with box office success of 20 top grossing movies of 2006.

- **Intelliseek** (after two acquisitions, this company is now known as Nielson Buzzmetrics, see PR [here](#)) G. Mishne and **N. Glance**. (2006) “Predicting movie sales from blogger sentiment.” In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs 2006*, Palo Alto, CA (Full text [here](#))

Abstract: The volume of discussion about a product in weblogs has recently been shown to correlate with the product’s financial performance. In this paper, we study whether applying sentiment analysis methods to weblog data results in better correlation than volume only, in the domain of movies. Our main finding is that positive sentiment is indeed a better predictor for movie success when applied to a limited context around references to the movie in weblogs, posted prior to its release.

- **Stanford University**, E. Sadikov, A. Parameswaran and P. Venetis, (2009). “Blogs as Predictors of Movie Success,” *Proceedings of the Third International ICWSM Conference* [Full text [here](#)].

Abstract: Analysis of a comprehensive set of features extracted from blogs for prediction of movie sales is presented. We use correlation, clustering and time-series analysis to study which features are best predictors.

Introduction: In this work, we attempt to assess if blog data is useful for prediction of movie sales and user/critics ratings. Here are our main contributions:

- We evaluate a comprehensive list of features that deal with movie references in blogs (a total of 120 features) using the full spinn3r.com blog data set for 12 months.
- We find that aggregate counts of movie references in blogs are highly predictive of movie sales but not predictive of user and critics ratings.
- We identify the most useful features for making movie sales predictions using correlation and KL divergence as metrics and use clustering to find similarity between the features.
- We show, using time series analysis as in (Gruhl, D. et. al. 2005), that blog references generally precede movie sales by a week and thus weekly sales can be predicted from blog references in the preceding weeks.

- We confirm low correlation between blog references and first week movie sales reported by (Mishne, G. et. al. 2006) but we find that (a) budget is a better predictor for the first week; (b) subsequent weeks are much more predictive from blogs (with up to 0.86 correlation).

- **Stony Brook University.** W. Zhang and S. Skiena, architects of TextMap (2009) "[Improving Movie Gross Prediction through News Analysis](#)," [Full text [here](#)]

Abstract: Traditional movie gross predictions are based on numerical and categorical movie data. But since the 1990s, text sources such as news have been proven to carry extra and meaningful information beyond traditional quantitative finance data, and thus can be used as predictive indicators in finance. In this paper, we use the quantitative news data generated by Lydia, our system for large-scale news analysis, to help us to predict movie grosses. By analyzing two different models (regression and k-nearest neighbor models), we find models using only news data can achieve similar performance to those use numerical and categorical data from The Internet Movie Database (IMDB). Moreover, we can achieve better performance by using the combination of IMDB data and news data. Further, the improvement is statistically significant.

- **York University (Canada):** Longbing Cao, Philip S. Yu, Chengqi Zhang and Huaifeng Zhang. Book: *Data Mining for Business Applications*, Springer US 2009: Chapter on "[Blog Data Mining: The Predictive Power of Sentiments](#)." [Abstract]

In this chapter, we study the problem of mining sentiment information from online resources and investigate ways to use such information to predict product sales performance. In particular, we conduct an empirical study on using the sentiment information mined from blogs to predict movie box office performance. We propose Sentiment PLSA (S-PLSA), in which a blog entry is viewed as a document generated by a number of hidden sentiment factors. Training an S-PLSA model on the blog data enables us to obtain a succinct summary of the sentiment information embedded in the blogs. We then present ARSA, an autoregressive sentiment-aware model, to utilize the sentiment information captured by S-PLSA for predicting product sales performance. Extensive experiments were conducted on the movie data set. Experiments confirm the effectiveness and superiority of the proposed approach.

- **MIT/ University of Applied Sciences Northwestern Switzerland /University of Cologne:** Lyric Doshia, Jonas Kraussabc, Stefan Nannabc, Peter Gloor (2009). "[Predicting Movie Prices Through Dynamic Social Network Analysis](#)." *Procedia – Social and Behavioral Sciences* [Full text [here](#)]

This paper explores the effectiveness of social network analysis and sentiment analysis in predicting trends. Our research focuses on predicting the success of new movies over their first four weeks in the box office after opening. Specifically, we try to predict prices on the Hollywood Stock Exchange (HSX), a prediction market on movie gross income, and predict the ratio of gross income to production budget. When predicting movie stock values on HSX, we consider two different approaches. One approach is to predict the daily changes in prices. This means we would be predicting a mix of not only how well we think the movie will perform, but also how we think other people think the movie will perform. Our second approach is to predict the final closing price of the stock, which will be how much the movie actually grosses in the box office after four weeks. In this approach, the daily prices provide feedback with the crowd's constantly revising estimate of the final performance of the movie. Finally, we try to classify movies in three groups depending on whether they gross less than, just over, or a lot more than their production cost. For our prediction we gather three types of metrics. (1) Web Metrics are movie-rating metrics from IMDb and Rotten Tomatoes as well as box office performance data from Box Office Mojo and movie quotes from HSX. (2) SNA Metrics Web and blog betweenness represent the general buzz on the movie from the web and from bloggers. We hypothesize that they

will be useful because they are unconscious signals about a movie's popularity. (3) To determine the general sentiment about the movies, we gather posts from IMDb forums to generate Sentiment Metrics for positivity and negativity based on the discussion in the forums. Our preliminary results employing different prediction methods such as multilinear and non-linear regression combining our three types of independent variables are encouraging, as we have been able to predict final box office return at least as good as the participants in the HSX prediction market.

See also this paper by the same group of researchers. "[Predicting movie success and academy awards through sentiment and social network analysis](#)," Krauss, Jonas; Nann, Stefan; Simon, Daniel; Fischbach, Kai; University of Cologne, Gloor, Peter, MIT, [Full text [here](#)]

### III. Content Acquisition

Data acquisition is also a significant problem for doing analysis of web content. In the course of this survey we ran across a number of companies that offer solutions geared toward this issue. This information is briefly noted below. We also identified a variety of content sources.

#### A. Data Collection Companies

##### 1. Spinn3r

There are web services for indexing vast amounts of data from blogs, and other social media for the purpose of sentiment analysis. [Spinn3r](#) is one such company. They are a web service that indexes the blogosphere. “We provide raw access to every blog post being published - in real time. They provide the data, and you can focus on building your application, mashup, or search engine.” The company has recently released Spinn3r 3.1 with both Twitter firehose support and Social Media Rank. They claim to index **twenty million blogs and counting, with 100k posts per hour**. They also watch every mainstream news source, with a purported **24 hours of archived content in under an hour**.

We have researchers at Harvard, Carnegie Melon, Stanford, Caltech, University of Maryland Baltimore County, University of Washington, University of Southern California, Nanyang Technological University, University of Edinburgh, National Institute of Informatics in Japan, California Institute of Technology, University of Hannover, Stony Brook University, and on and on ([source](#)).

Basically, if you're a PhD researching blogosphere, you're probably using Spinn3r.

Cornell recently launched a [Memetracker](#) powered by Spinn3r; TextMap is another search engine that uses Spinn3r.

On their technical specs page:

##### Computed Metadata

- \* mathematically computed language classification
- \* codepage detection for all multibyte languages
- \* n-gram language detection for European languages
- \* internal spam **probability detection**
- \* content extraction or chrome/template removal
- \* inbound link count

Interesting publications can be found on this [page](#).

##### 2. 80legs

80legs purportedly leverages a grid of 50,000 servers “to search and crunch millions of Web pages within minutes.” Target customers include market researchers looking to mine public opinion on a particular product or service, lawyers searching for copyright infringement and piracy, or online ad agencies looking to do competitive analysis of where rival firms are placing their ads, says 80legs, CEO, Shion Deysarkar ([source](#)).

### 3. Techrigy

Techrigy has a proprietary Social Media Warehouse (SMW). Its product, SM2, is an alerting and analysis tool that sits on top of the SMW. The SMW is a real-time system that collects content from sources including: blogs, wikis, message boards/forums, video/photo sharing websites, blogs, microblogs, social networks, As of July 2008, the Social Media Warehouse, claims to have over 600 million pieces of content and growing. Access to the Social Media Warehouse allows SM2 to perform full historical analysis of brands as well as perform real-time alerting (see their factsheet [here](#)). Clarabridge is one of their customers.

FYI: In a product comparison Techrigy's sentiment analysis platform, according to a reviewer at WebMetricsGuru is a "total wash out – most of the results don't make sense – only one or two actually have any connection to [sentiments regarding] Social Media Week New York City – and they're not negative – at least, not negative about Social Media Week." (More detail [here](#))

### 4. Teragram

The Teragram Web Crawler is a standalone tool that enables customers to automatically download documents from the Internet.

Starting at a user-specified URL, the crawler follows the hyperlinks in the web, while repeatedly sending HTTP requests to simultaneously obtain corresponding HTML content and any URLs existent within that content... The Teragram crawler is used in a multi-threading mode. The number of threads can be specified according to your requirements. For example, specify more than 1,000 threads for large scale crawling. In this case the download speed can reach 10M bytes per second on a single machine with good bandwidth. The crawler can also be deployed in a distributed cluster environment (more detail [here](#)).

## B. Dataset Lists

Datasets for Data Mining (KDD Nuggets <http://www.kdnuggets.com/datasets/index.html>)

- [KDD Cup center](#), with all data, tasks, and results.
- [UCI KDD Database Repository](#) for large datasets used in machine learning and knowledge discovery research.
- [UCI Machine Learning Repository](#).
- [AWS \(Amazon Web Services\) Public Data Sets](#), provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications.
- [DataFerrett](#), a data mining tool that accesses and manipulates TheDataWeb, a collection of many on-line US Government datasets.
- [Delve](#), Data for Evaluating Learning in Valid Experiments
- [Enron Email Dataset](#), data from about 150 users, mostly senior management of Enron.
- [FEDSTATS](#), a comprehensive source of US statistics and more
- [FIMI repository for frequent itemset mining](#), implementations and datasets.
- [Financial Data Finder at OSU](#), a large catalog of financial data sets
- [GEO \(GEO Gene Expression Omnibus\)](#), a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.
- [Grain Market Research](#), financial data including stocks, futures, etc.
- [ICWSM-2009 dataset](#) contains 44 million blog posts made between August 1st and October 1st, 2008.

- [Infobiotics PSP \(protein structure prediction\) datasets](#), adjustable real-world family of benchmarks for testing the scalability of classification/regression methods.
- [Investor Links](#), includes financial data
- [Microsoft's TerraServer](#), aerial photographs and satellite images you can view and purchase.
- [MIT Cancer Genomics gene expression datasets and publications](#), from MIT Whitehead Center for Genome Research.
- [NASDAQ Data Store](#), provides access to market data.
- [National Government Statistical Web Sites](#), data, reports, statistical yearbooks, press releases, and more from about 70 web sites, including countries from Africa, Europe, Asia, and Latin America.
- [National Space Science Data Center](#) (NSSDC), NASA data sets from planetary exploration, space and solar physics, life sciences, astrophysics, and more.
- [PubGene\(TM\) Gene Database and Tools](#), genomic-related publications database
- [SMD: Stanford Microarray Database](#), stores raw and normalized data from microarray experiments.
- [SourceForge.net Research Data](#), includes historic and status statistics on approximately 100,000 projects and over 1 million registered users' activities at the project management web site.
- [STATOO Datasets part 1](#) and [STATOO Datasets part 2](#)
- [UCR Time Series Classification/Clustering page](#), offering datasets, papers, links, and code.
- [United States Census Bureau](#).

Also: [The info.org](#): This is a site for “large data sets and the people who love them: the scrapers and crawlers who collect them, the academics and geeks who process them, the designers and artists who visualize them. It’s a place where they can exchange tips and tricks, develop and share tools together, and begin to integrate their particular projects.”

FYI: The Netcraft Survey is one of the most widely cited measures of the Internet. “In the December 2009 survey they received responses from 233,848,493 sites” ([source](#)).

## IV. Appendix

### A. Academic Papers

In the course of this background work we ran across many academic papers that may be of interest to NGC. Though NGC team members are presumably already aware of important academic work in the various categories, the citations are included below for completeness.

#### 1. Multi-lingual

These papers were found during a keyword search using the terms, “multi-lingual and sentiment OR opinion.”

- Pant, D. and Sharma, M. K., (Kumaun University, India), Amrapali Institute (India) (2009), “Web Mining and Social Network Analysis in Cyber war, to warn about terrorist attacks” [Full text [here](#)]
- C Zhang (Jilin University, China, and Artillery Command College of Shenyang, China), D. Zeng, (University of Arizona and the Chinese Academy of Sciences), J. Li (Drexel University), F.-Y. Wang, (Chinese Academy of Sciences), Wanli Zuo (Jilin University, China) (2009). “Sentiment analysis of Chinese documents: From sentence to document level.” Journal of the American Society for Information Science and Technology archive, 60(12). [[Abstract](#)]
- Abbasi, A., H. Chen and A. Salem (University of Arizona) (2008). “Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums.” ACM Transactions on Information Systems 26(3). *ACM Transactions on Information Systems*, 26 (3): Art. No. 12 [Full text [here](#)]
- Boi y, E. and Marie-Francine Moens, (University of Leuven) (2008), “A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts” [Full text [here](#)]
- J.V. Román, S. L.-Serrano, J. C. González-Cristóbal (Universidad Carlos III de Madrid, Universidad Politécnica de Madrid) (2008), “First Experiments on Multilingual Opinion Analysis,” Proceedings of NTCIR-7 Workshop Meeting, December 16–19, 2008, Tokyo, Japan [Full text [here](#)]
- Carmen Banea and Rada Mihalcea, (U North Texas), Janyce Wiebbe, (U Pittsburgh), Samar Hassan, (U North Texas) (2008), “Multilingual Subjectivity Analysis Using Machine Translation,” Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 127–135, Honolulu. [Full text [here](#)]
- Tomohiro Fukuhara, Yoshiaki Arai et al., (U-Tokyo, U-Tokyo Denki, U-Tsukuba) (2008) “KANSHIN: A Cross-lingual Concern Analysis System using Multilingual Blog Articles” [Full text [here](#)]
- Chen, H., (University of Arizona) (2008), “Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums.” *ACM Transactions on Information Systems*, forthcoming. [Full text [here](#)]
- Bautin M., S. Skiena, L. Vijayarenu (Stony Brook University, Yahoo! Inc.), “International Sentiment Analysis for News and Blogs” [Full text [here](#)]
- Liu K. and J. Zhao, (Chinese Academy of Sciences) (2008), “NLPR at Multilingual Opinion Analysis Task in NTCIR7.” [Full text [here](#)]
- Evans, D.E. and Kando, N., (National Institute of Informatics, Tokyo, Japan) (2007): “Multi-lingual Opinion Analysis Applied to World News: A Case Study.” [Full text [here](#)]

## 2. Topology

- Andrew Arnold and William W. Cohen (Carnegie Mellon) (2009), "Information Extraction as Link Prediction: Using Curated Citation Networks to Improve Gene Detection," Association for the Advancement of Artificial Intelligence [Full text [here](#)]
- Hidehiko Ino , Mineichi Kudo , Atsuyoshi Nakamura (Hokkaido University, Japan) (2005), "Partitioning of Web graphs by community topology," Proceedings of the 14th international conference on World Wide Web, May 10-14, 2005, Chiba, Japan [Abstract]
- Qiaozhu Mei , Chao Liu , Hang Su , ChengXiang Zhai (University of Illinois at Urbana Champaign, Urbana) (2006), "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," Proceedings of the 15th international conference on World Wide Web, May 23-26, 2006, Edinburgh, Scotland [Full text [here](#)]
- Brett Adams, Dinh Phung, Svetha Venkatesh, (Curtin University of Technology, Australia) (2009). "Social reader: following social networks in the wilds of the blogosphere," Proceedings of the first SIGMM workshop on Social media, October 23-23, 2009, Beijing, China [Full text [here](#)]
- Seungyeop Han, Yong-yeol Ahn, Sue Moon, Hawoong Jeong, (KAIST), "Collaborative Blog Spam Filtering Using Adaptive Percolation Search." [Full text [here](#)]
- Kazunari Ishida, (Tokyo University of Agriculture) (2005), "Extracting Latent Weblog Communities, - A Partitioning Algorithm for Bipartite Graphs" [Full text [here](#)]
- "Why we search: visualizing and predicting user behavior," University of Washington [Full text [here](#)]

## 3. Topic Modeling

- Tae Yano, William W. Cohen, Noah A. Smith, (Carnegie Mellon University) (2009) "Predicting Response to Political Blog Posts with Topic Models," NAACL 2009. [Full text [here](#)]
- Chenghua Lin, Yulan He (University of Exeter, UK; Open University, UK) (2009), "Joint sentiment/topic model for sentiment analysis." In: Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China. [Abstract]
- Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai (University of Ill- Urbana-Champaign, Vanderbilt University) (2007). "Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs." [Full text [here](#)]
- Krisztian Balog and Maarten de Rijke (University of Amsterdam) (2007). "How to Overcome Tiredness: Estimating Topic-Mood Associations," In: ICWSM 2007 Boulder, Colorado, USA [Full text [here](#)]
- Lei Shi, Bai Sun, Liang Kong, Yan Zhang (Peking University) (2009), "Web Forum Sentiment Analysis based on Topics" CIT 2009. [Full text [here](#)]
- Yue Lu, Chengxiang Zhai (University of Illinois at Urbana-Champaign) (2009). "Opinion integration through semi-supervised topic modeling." In: Proceedings of the 17th international conference on World Wide Web [Full text [here](#)]
- Duo Zhang, Chengxiang Zhai, Jiawei Han (Tsinghua University, Carnegie Mellon, University of Illinois at Urbana-Champaign) (2009). "Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases," In: SIAM International Conference on Data Mining - SDM 2009. [Full text [here](#)]
- Duo Zhang, ChengXiang Zhai, Jiawei Han, Ashok Srivastava, Nikunj Oza (University of Illinois at Urbana-Champaign / NASA Ames Research Center). (2009) "Topic modeling for OLAP on multidimensional text databases: topic cube and its applications," *Statistical Analysis and Data Mining*, 2(5-6): 378 - 395 [Abstract]

- Ivan Titov, Ryan McDonald (University of Illinois at Urbana-Champaign, Google, Inc.) (2008) “A Joint Model of Text and Aspect Ratings for Sentiment Summarization,” In: Proceedings of ACL '08 [Full text [here](#)]
- Yi Hu, Ruzhan Lu, Yuquan Chen, Jianyong Duan (2007), “Using a Generative Model for Sentiment Analysis,” *Computational Linguistics and Chinese Language Processing*, 12(2): 107-126 [Full text [here](#)]
- Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai (University of Illinois at Urbana-Champaign) (2007) “Topic sentiment mixture: modeling facts and opinions in weblogs.” In: WWW 2007 / Track: Data Mining [Full text [here](#)]
- Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, Chengxiang Zhai (Vanderbilt University, Carnegie Mellon University, University of Illinois at Urbana-Champaign, Zhejiang University, China) (2007), “Topic sentiment mixture: modeling facets and opinions in weblogs.” In: World Wide Web Conference Series, 2007 [[Abstract](#)]
- Min Zhang, Xingyao Ye (Tsinghua University) (2008) “A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval,” In: Research and Development in Information Retrieval – SIGIR 2008 [Full text [here](#)]
- Kerstin Denecke, Mikalai Tsytsarau, Themis Palpanas, Marko Brosowski (L3S Research Center, University of Trento, Italy; IBM Research) (2009), “Topic-related sentiment analysis for discovering contradicting opinions in weblogs,” Technical Report # DISI-09-037. [Full text [here](#)]
- Julian Brooke, Matthew Hurst (University of Toronto, University of Sheffield) (2009). “Patterns in the Stream: Exploring the Interaction of Polarity Topic and Discourse in a Large Opinion Corpus.” In: TSA'09, November 6, 2009, Hong Kong, China [Full text [here](#)]
- Carnegie Mellon Course, taught by Tae Yano: “Polling Made Easy: Aggregating Political Sentiments Using Topic Modeling” [website [here](#)]

#### 4. Community Detection

- Xiaomao Yu, Yong Jiang (Tsinghua University, China) (2009) “BlogosphereExplorer: Opens a window to the blogosphere.” In WebSci '09 [Full text [here](#)]
- Yuzhou Zhang; Jianyong Wang; Yi Wang; Lizhu Zhou (Tsinghua University, Google Beijing Research) (2009) “Parallel Community Detection on Large Networks with Propinquity Dynamics.” Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. [[Abstract](#); KDD [video lecture](#)]
- Reza Zafarani and Huan Liu (Arizona State University) (2009), “Connecting Corresponding Identities across Communities.” [Full text [here](#)]
- Elham Khabiri, Chiao-Fang Hsu, James Caverlee (Texas A&M University) (2009) “Analyzing and Predicting Community Preference of Socially Generated Metadata: A Case Study on Comments in the Digg Community.” Association for the Advancement of Artificial Intelligence. [Full text [here](#)]
- Qiankun Zhao, Sourav S. Bhowmick, Xin Zheng, Kai Yi (AOL Lab, China; Nanyang Technological University, Hong Kong; Tsinghua University, China; Peiking University, China) (2008). “Characterizing and predicting community members from evolutionary and heterogeneous networks.” In: Proceedings of the 17th ACM conference on Information and knowledge management. [[Abstract](#)]
- Nan Du, Bin Wu, Xin Pei, Bai Wang, Liutong Xu, (Beijing University of Posts and Telecommunications, China) (2007). “Community Detection in Large-Scale Social Networks,” [Full text [here](#)]

### University of Maryland, Ebiquity Group.

- Anubhav Kale, Amit Karandikar, Pranam Kolari, Akshay Java, Tim Finin, Anupam Joshi (University of Maryland) "[Modeling Trust and Influence in the Blogosphere Using Link Polarity](#)," [Full text [here](#)]

There is a growing interest in social network analysis to explore how communities and individuals spread influence. We describe techniques to find "like minded" blogs based on blog-to-blog link sentiment for a particular domain. [researchers used Buzzmetrics dataset] ...The table in figure 2 depicts polarity values computed between some pairs of influential democratic and republican blogs. We present this data as a quick measure of demonstrating the potential of our work and make the following observations.

1. Trust propagation was effective in predicting the accurate polarity for DK-AT, even though our text processing did not yield the correct polarity initially.

- Akshay Java, Anupam Joshi, and Tim Finin (University of Maryland), "[Approximating the Community Structure of the Long Tail](#)," Proceedings of the Second International Conference on Weblogs and Social Media (ICWSM 2008) [Full text [here](#); presentation [here](#)]

In many social media applications, a small fraction of the members are highly linked while most are sparsely connected to the network. Such a skewed distribution is sometimes referred to as the "long tail". Popular applications like meme trackers and content aggregators mine for information from only the popular blogs located at the head of this curve. On the other hand, the long tail contains large volumes of interesting information and niches. The question we address in this work is how best to approximate the community membership of entities in the long tail using only a small percentage of the entire graph structure.... <no claims of prediction>

More papers and presentations listed [here](#).

"[Party Polarization in Congress: A Social Networks Approach](#)," a social network study from mathematicians at UC San Diego, Caltech, UNC Chapel Hill and University of Oxford. [Full text [here](#)]

We use the network science concept of modularity to measure polarization in the United States Congress. As a measure of the relationship between intra-community and extra-community ties, modularity provides a conceptually-clear measure of polarization that directly reveals both the number of relevant groups and the strength of their divisions. Moreover, unlike measures based on spatial models, modularity does not require predefined assumptions about the number of coalitions or parties, the shape of legislator utilities, or the structure of the party system. Importantly, modularity can be used to measure polarization across all Congresses, including those without a clear party divide, thereby permitting the investigation of partisan polarization across a broader range of historical contexts. Using this novel measure of polarization, we show that party influence on Congressional communities varies widely over time, especially in the Senate. We compare modularity to extant polarization measures, noting that existing methods underestimate polarization in periods in which party structures are weak, leading to artificial exaggerations of the extremeness of the recent rise in polarization. We show that modularity is a significant predictor of future majority party changes in the House and Senate and that turnover is more prevalent at medium levels of modularity. We utilize two individual-level variables, which we call 'divisiveness' and 'solidarity,' from modularity and show that they are significant predictors of reelection success for individual House members, helping to explain why partially-polarized Congresses are less stable. Our results suggest that modularity can serve as an early-warning signal of changing group dynamics, which are reflected only later by changes in formal party labels.

## B. List of Companies in this Report

For the convenience of the reader, below is a list of companies mentioned in this report.

80legs	Jodange	SentimenTrader
Alias-I	KnowEm	SentiMetrix
AlphaSimplex	Lexalytics	Sentimine (Parnassus Group)
Applied Systems Intelligence	LifeTips	Social Mention
Attensity	ListenLogic	Spinn3r
Autonomy	Marchex Rep Mgmt	SPSS
Biz360	Mark Monitor	Sysomos
BuzzLogic	Market Sentinel	Technorati
BuzzStream	Microsoft	Techrigy
Clarabridge	MITRE Corp	Temis
Converseon	NEC Laboratories, America	Teragram (SAS)
Crimson Hexagon	NetBase Consumer Insights	TextMap/Med/Blog/Biz
Dow Jones Insight	NewsLive	Thompson Reuters
Envisional	NewsSift	TipTop Technologies TNS Cymfony
Feedback Ferret	Newstwit	Trackur
FiltrBox	Nielson BuzzMetrics/BlogPulse	Twazzup
Fin-buzz	Nstein	Tweet Sentiments
Galaxy Advisors	Ogilvy	Tweetfeel
General Sentiment	ParSOS	Twendz
Google	People Browsr	Twittermood
Gravity Technologies	Psydex Corporation	UnderstandMarket
IBM	Radian6	ViralHeat
Infegy	Ravenpack	Visible Technologies
Inside View	Reputica	Who's Talkin
Inxight (SAP)	Reputrace	WiseWindow
IVolatility	Scout Labs	Yahoo! Research
J. D. Power & Assoc (Umbria)	SEER	