SNAPSHOT NEWS ANALYSIS AND VISUALISATION

CHARTING THE EVOLUTION OF NEWS ON THE NET

Mark Grundland
Functional Elegance
Mark@FunctionalElegance.com

John Snyder
Grapeshot Ltd.
info@grapeshot.com

ONLINE NEWS ANALYSIS

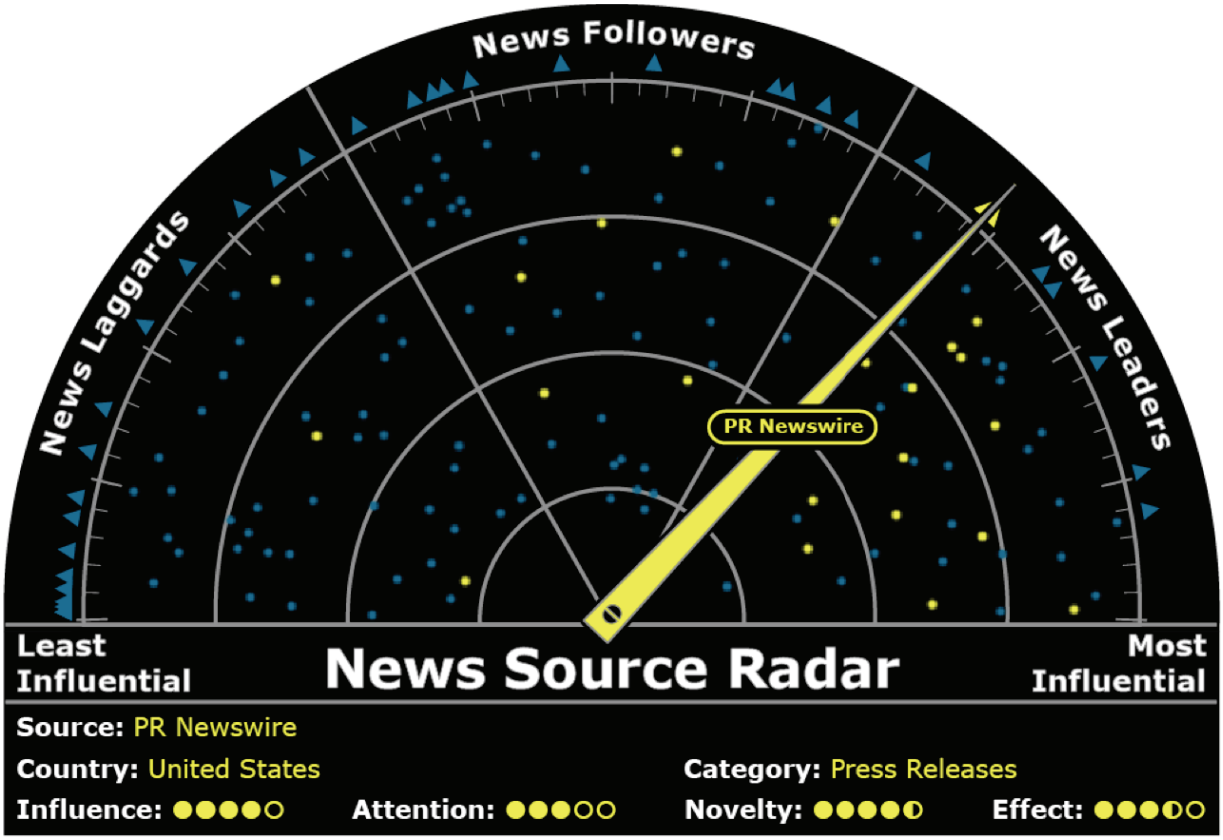
SnapShot has been developed to provide standardised, objective metrics for tracking the media coverage of any topic on a daily basis.

With over 500,000 news articles being published every day across the world, it is harder than ever just to keep up, never mind stay ahead. What if there was a simple way to see what is really going on?

IBM and Grapeshot Partnership

Developed by Grapeshot in partnership with IBM using information retrieval research from the University of Cambridge, SnapShot is an online news analysis and visualisation service that makes it easy to follow how media coverage evolves over time. Taking the traditional scrapbook of news clippings into the twenty-first century, we chart the course of the underlying forces that shape the news landscape. SnapShot helps public relations, marketing, finance, and journalism professionals to be more productive by enabling them to draw valuable insight from the current events that impact their bottom line.

By monitoring the flow of online news 24 hours a day, we enable SnapShot users to understand how the public perception of their key concerns is being influenced by the media. Our technology bridges the communication gap between skimming the headlines and tracking the trends, effectively automating the extraction of actionable knowledge from the daily deluge of information published online as news.



Novelty and Effect

By filtering news according to its relevance, novelty, effect, and influence, SnapShot can bring uncommon knowledge to your attention, giving you the competitive edge of being the first to know and the first to act. Furthermore, our news analysis technology can directly address the timeless questions that every publicist confronts day in and day out: What to say? How to say it? Who to say it to? Did it make any difference? How well has it been received?

Applying the principles of information visualisation, we can deliver the answers in an easy to use format, always up-to-date and ready to go into a report or presentation. By providing an objective, impartial perspective on the trends driving the news, our news analysis can serve to focus attention, stimulate discussion, benchmark performance, and justify expenditure. Responding to the prime concerns of a variety of industries, SnapShot technology serves to integrate business intelligence tools with online news delivery services:

- **Marketing and public relations firms:**
  Did the press release alter news coverage?
- **Investment and financial analysts:**
  Does the latest news defy market expectations?
- **Brand managers and industry watchers:**
  Which companies lead the pack on the key issues?
- **News editors and journalists:**
  What little stories are poised to make big headlines?

Modelling the Dynamics of News

At Grapeshot, we have unique expertise in modelling the statistical dynamics of news. SnapShot technology provides analysis and visualization tools applicable to both historical news archives and breaking news feeds.

To transform the qualitative content of a news wire into the quantitative metrics of a news radar, SnapShot automatically annotates news articles to describe their significance in view of the related news coverage. We track the precedents and consequences of each news article published in order to account for the factors that drive the growth and decline of its news story. We can then track the progress of each news story to work out the force of its impact on the real world. In this way, we are able to reveal what distinguishes the best from the rest, what makes a few news articles so influential that they set the agenda for everyone else. The SnapShot system consists of a SnapShot server that performs news profiling and retrieval as well as a SnapShot web client that performs news analysis and visualization.

MONITORING NEWS FEEDS

SnapShot is designed to deliver real-time performance, world wide coverage, and rich metadata. It can work with any source of XML data feed, but was designed and built throughout a three year R&D project using a comprehensive news feed from Moreover Technologies, a world leader in online news aggregation and media monitoring. SnapShot ingested news articles from global, regional, and local news providers, as well as industry publications, press releases, editorial opinions and popular blogs.

Meta-data

When indexing the news, we take into consideration not only the content of each news article but also its wider context: author, publisher, publisher reputation, time and date of publication, country of publication, news genre and channel, featured industry sectors and content topics, as well as named entities such as companies and geographical locations.

This extensive ensemble of metadata enables us to analyze and visualise news from a wide variety of perspectives. Although the R&D project's prototype only covers English language news articles, our technology is specifically intended to encompass other languages as well.

PROFILING NEWS ARTICLES

The technological foundation for SnapShot is the Grapeshot state-of-the-art probabilistic information retrieval system.

A result of over six years of commercial research and development, Grapeshot is the creation of our co-founder and chief scientist Dr. Martin Porter, a world renowned information retrieval researcher, and the inventor of the Porter stemming algorithm that is a key component of most online search engines including Google and Microsoft Bing.

Essence of the Story

Relying on sophisticated statistical modelling, our technology applies an experimentally validated, evidence based approach for automatically determining the keywords in the text that meaningfully characterise the content of a news article.

To understand the makings of a news story, we select the keywords that best exemplify the unique contribution an individual news article makes to the evolution of the collective news coverage. In order to prevent stylistic variations from obscuring factual content, we employ language models to help us identify which words are most likely to carry meaning and sentiment.

Keyword Profile

Avoiding the rigid constraints of linguistic templates and semantic ontologies that have so often limited the scope of competing approaches, we concentrate on figuring the odds that each word in the article is essential to conveying its message.

While many common words may appear frequently in the article, we focus instead on the few significant keywords whose frequency of appearance in the article is unexpectedly high when compared to other articles. For our purposes, the significance of a keyword is mathematically defined by its capacity to optimally distinguish the news article from the rest of the news.

Knowing what sets a news article apart, allows us to reliably predict its relevance to any given topic of interest. By applying an objective scientific procedure to evaluate the significance of every word occurring in a news article, we are able to capture the essence of a news article by a small set of optimally chosen keywords. We can then use these keyword profiles to retrieve, summarise, compare, categorise, and relate news articles in the context of the news stories they convey. In this way, keyword profiles serve to represent what makes a news article actually worth reading.

TRACKING NEWS STORIES

SnapShot's approach to news analysis is based on the simple observation that most news is not new; it is just more of the same. Hence, a series of news articles gathered around a common theme can serve to chronicle the development of their news story.

Shared News

As news articles can be regarded as contributions to a public conversation about the state of the world, each news article is situated in the news coverage according how it reflects what was said beforehand and how it affects what is said afterwards.

News articles conveying a shared message cast votes of confidence in the relative importance of their news story. Therefore, we evaluate the newsworthiness of a news article by assessing its similarity to other related articles published at around the same time.

Similarity Analysis

In much the same manner as biologists trace the evolutionary history of living organisms by measuring the degree of overlap between the DNA profiles that encode what makes them unique, we measure the similarity between news articles by the degree of overlap between their keyword profiles.

We simply count the number of keywords they have in common relative to the number of keywords each profile has in total. When the proportion of common keywords reaches a target threshold, we declare that the news articles are similar and therefore related to a shared news story. In this way, we can perform a content dependent, hierarchical categorization of the news, capturing how the different strands of news coverage weave together the components of the public message communicated by the media.

Duplicate Removal

In cases where the keyword profiles are virtually identical, we determine the news articles to be duplicates, even though they may originate from different news sources, which is not at all unusual given how widespread syndication is on the internet.

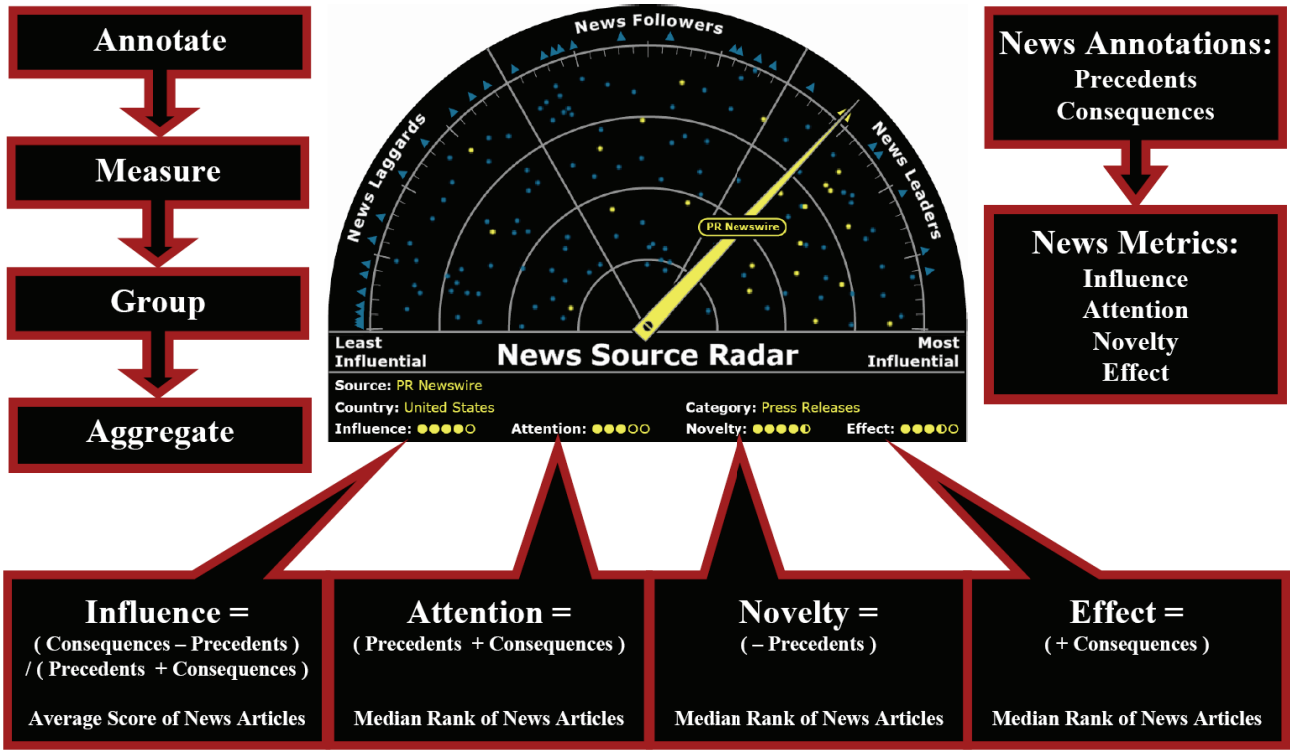
Depending on the type of analysis performed, duplicate articles can be removed from consideration, thus ensuring the user is not distracted by the needless repetition of identical information. Just as genetic analysis can serve to reunite ancestors and descendents separated by time, geography, or circumstances, we apply news analysis to connect each news article to its precedents and consequences, the related news articles that comprise its news story.

Note that precedents and consequences are tallied when a news article is first indexed by the SnapShot server in order to ensure that the news metrics derived from them are available without delay to the SnapShot's web client whenever the news article is retrieved in response to a user-specified news inquiry.

EVALUATING NEWS METRICS

For historical news analysis, typically performed over the course of a couple of weeks of news coverage, we compute the news metrics of novelty and effect for each news article in accordance to the number of its precedents and consequences.

Hence, we ascertain the novelty of a news article, a measure of the relative absence of precedents, by assessing its similarity to earlier relevant articles, typically taken from the week preceding its publication. Likewise, we ascertain the effect of a news article, a measure of the relative prevalence of consequences, by assessing its similarity to later relevant articles, typically taken from the week subsequent to its publication.



Classification

The evaluation of novelty and effect makes possible a four-way scientific classification of news articles:

- **Leaders:** A novel and effective article, with few precedents but numerous consequences, is considered influential
- **Followers:** An unoriginal but effective article, with numerous precedents and numerous consequences, is considered conventional.
- **Laggards:** An unoriginal and ineffective article, with numerous precedents but few consequences, is considered obsolete.
- **Eccentrics:** A novel but ineffective article, with few precedents and few consequences, is considered anomalous.

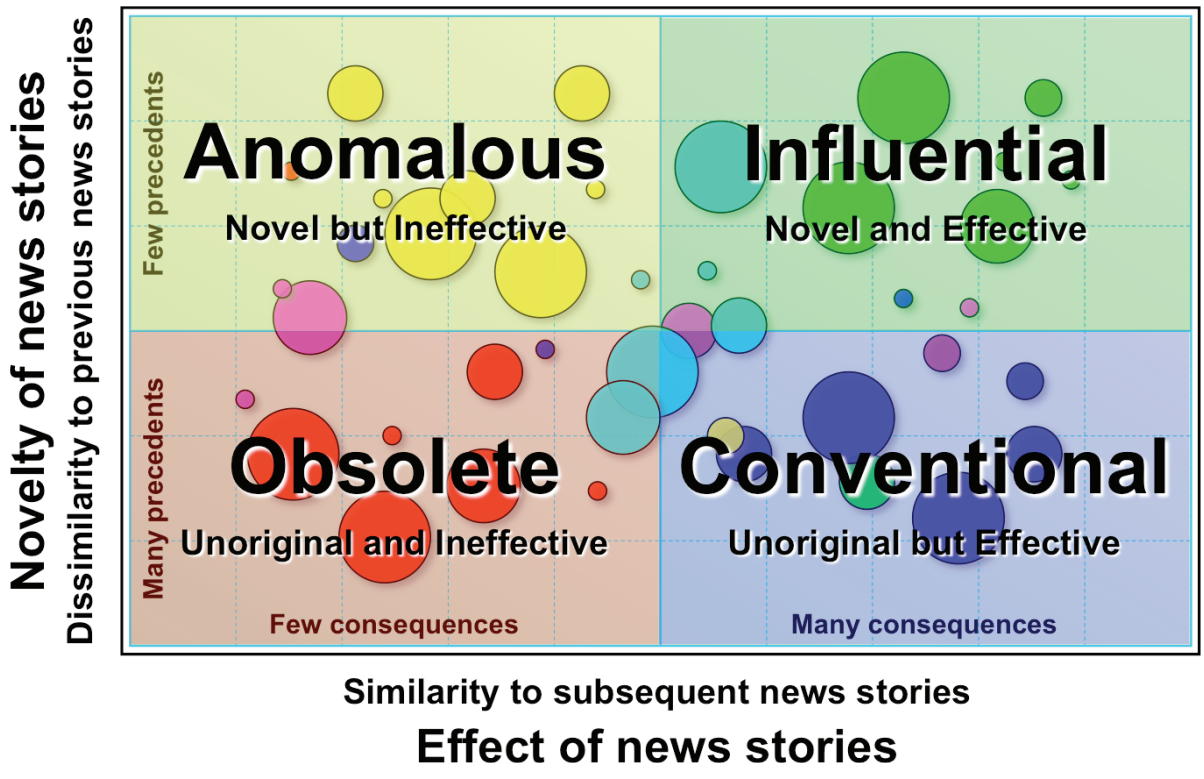
Accordingly, the media attention of a news article can be calculated from the sum of its consequences and its precedents while the media influence of a news article can be calculated from the difference between its consequences and its precedents.

Furthermore, the ratio of media influence to media attention determines its outlook, which is positive when the news is primarily proactive and negative when the news is primarily reactive. In this way, relevant news articles can be scored, ranked, and filtered according to their novelty, effect, attention, influence, outlook and classification.

Signal and Noise

For any given topic, we can take the pulse of the media by measuring its signal-to-noise ratio, which is the ratio of the number of influential articles to the number of obsolete articles published in the same time frame. In an analogous fashion, for recent news analysis, usually performed over the course of the last 24 or 48 hours of news coverage, although we cannot foretell future consequences, we can continuously monitor the rate at which contemporaneous similar news articles are being published.

For each news article, this enables us to compute news metrics such as size, velocity, and momentum. We can then use temporal news analysis to attempt to alert our users to the emerging news topics that are poised to take off. By adopting these objective news metrics, we effectively take the guesswork out of figuring out "what's hot" and "what's not".



CHARTING NEWS COVERAGE

SnapShot takes an ongoing census of the news and presents the big picture, using charts and league tables to track the relative performance of the different segments of the news coverage.

Just as market research seeks to understand what motivates customers to buy, relying on demographic analysis to assign people to demographic groups according to their shared characteristics, SnapShot seeks to reveal how the media influences public opinion to react, applying news analysis to classify news articles into news groups according to their shared attributes.

Media Map

Going beyond the mere headlines of the news articles and the themes of their news stories, Snapshot is designed to provide a map to the media, from the point of view of our clients' interests.

Segmenting news articles into new groups not only renders our analysis more reliable, less dependent on the vagaries of individual reporters, but more importantly it allows us to present our users with a conclusive, up-to-date overview of how news coverage relates to their specific concerns, which is much more useful than just the stream of endless, out-of-date examples one can readily obtain from any news portal.

Trajectory of News

To reveal the trends that shape public perception, visualisation can allow us to depict the trajectories of how news stories and their coverage in the media evolves over time.

For example, it becomes possible to compare how an issue is covered by mainstream versus alternative press, American versus European sources, industry trade journals versus general interest newspapers, blog opinions versus journalist reports. SnapShot technology can enable the coverage of news topics to be tracked in real-time with the precision of stock quotes: the level, the change, the low, the high, the running average, the sector index and sector rank.

User-Centric Query

The SnapShot web client works by asking the user to specify their topic of interest using just a few simple search terms. For virtually any type of query, whether a buzzword or a brand, a celebrity or a product, a social issue or a political event, we can not only gauge who is gaining and who is losing public attention by objectively measuring the rate of change in media coverage but we can also attempt to explain why by examining the composition of media coverage.

The number of news articles retrieved can be adjusted according to the amount of media coverage expected. Furthermore, the news analysis can be optionally constrained to only consider news articles from highly reputable news sources. To direct the news analysis, the user can select the desired grouping criterion, such as source, genre, featured industry sectors or companies.

The system responds by grouping the retrieved news articles accordingly, skipping over any groups deemed too small to be accurately represented. It then calculates the aggregated news metrics for each news group, such as average novelty, average effect, average media attention, and average media influence. The news groups shown can be filtered as necessary, with the most relevant ones highlighted in different colours.

Four Quadrants

For the purposes of news visualisation, we provide a news radar to display the relative positioning of the various news groups. For instance, a typical news visualisation features novelty as the vertical axis and effect as the horizontal axis, so that, in clockwise order, the top right quadrant corresponds to the leaders, the bottom right quadrant corresponds to the followers, the bottom left quadrant corresponds to the laggards, and the top left quadrant corresponds to the eccentrics. Hence, at a glance it is possible to see who leads where others follow, what distinguishes the front-page news from the back-page curiosity.

Further Visualisations

We can readily prepare other types of news visualisations depending on the requirements of the application. For instance, we can provide dashboard dials or traffic lights, lightweight widgets that can be embedded in a corporate portal to indicate the current level of media influence and attention its brand is attracting in the news as compared to its closest competitors. Our news visualisations can be animated across a time line and saved in a gallery for later viewing.

It is easy to assess how adding search terms to refine the scope of the analysis presents different views of the news landscape. In effect, our clients can gain direct insight into the key factors that promote or suppress the media coverage of their business interests.

Our information visualisations are designed to be understood at a quick glance, without recourse to any complicated explanations. They are meant to be suitable for copying and pasting into a PowerPoint presentation, a management report, or a corporate blog. Moreover, to articulate the findings that the picture suggests, it can be possible to express the trends in everyday language, in the form of an automatically generated quote that can be inserted into any text document.

At Grapeshot, we measure the attention span of the media.