Social Media Analytics

Background

John Naisbitt, author of bestselling book Megatrends, stated in 1995 that "we are drowning in information but starved for knowledge." Around the same time a study at Berkeley University has estimated that by the end of 1999, the sum of human-produced information (including all audio, video recordings and text) to be about 12 Exabyte of data, where one Exabyte is 1 million terra bytes. By 2009, only a decade later, 494 Exabyte of information were transferred seamlessly across the globe every day, according to The Digital Britain Final Report.

The voice of John Naisbitt on data overload was resounded in early 2009, in an article titled "Too Much Information, Too Little Understanding", which was published by The Washington Post. Kathleen Parker, an American syndicated columnist and author of the article feared that toxic assets can soon exhausts our cognitive resources while making the nonsensical [seem] significant. She requested to "turn off."

Few months before this article was published, a storm was already brewing in the venue of Web 2.0 Expo on the issues of information overload. Clay Shirky, a New York University new-media professor, writer, and consultant on the social and economic effects of internet technologies, argued that there have always been complaints about information overload, whenever a new media entered our life - it happened with radio, TV, and now it is time for internet and social media. He argued not to turn off but to design new filters for information processing.

On December 9th 2011, a 25-year-old animation movie "Laputa: Castle in the Sky" was beamed on Japanese television. During one point in the TV broadcast, viewers joined forces, sending tweets at the same time to symbolically help the movie's characters cast a spell. Viewers together had managed to send 25,088 tweets in just one second. It is estimated otherwise, that 600 tweets are sent per second globally by now, and a higher number of messages are exchanged in Facebook - of about 700 of them per second.

As social media usage turned commonplace in astounding speed, the impact of this can clearly be felt in the corporate world as tipping point has just inched near. According to a survey by the Worldcom Public Relations Group, more than half (54%) of companies surveyed plan to increase spending on social media in 2011. This bodes well for Twitter, the most popular channel, used by 85 percent of global respondents, followed by Facebook (74%), LinkedIn (72%), YouTube (69%) and corporate blogs (60%).

What prompts companies to channelize funds in social media is that it has shifted much of the almost always labor-intensive market research work of intelligence-gathering and marketing efforts of brand building to the customer. Surely the bandwidth of social media is very wide and noise level is little too high, but a right tuning and intelligent filtering can let you hear true voice of your customer.

Challenge

Creating intelligible information by analyzing social media data is hard. What makes it hard is the sheer volume of information and high degree of noise. Tweets and status updates are small, often written using abbreviated words, discarding linguistic rules. High rate of data demands more online processing of the information so that only intelligence is captured and summarized rather than building local copies of data-ocean.

At Nuvento, our primary understanding is text treatment and preparation phase has a direct influence on the analysis quality and reliability. Cleaner the input data the better is the analysis quality. However, more cleaning is computationally time consuming and tends to make the approach more offline. It is challenging to design data cleansing and preparation phase in a very efficient manner.

Prepared data is subjected to statistical and machine learning algorithmic techniques to reduce problem dimensionality, to identify correlations, create clusters, establish linkages, built relationships, and to create predictive rules among buzz elements. Right kind of analysis is not straight forward either. Consider the problem of understanding whether a tweet is simply informative or actionable e.g. request for information, requires innovative methods derived from deep linguistic and statistical ideas. While analysis builds intelligible information – they are still not in human readable format, and require visualization.

Visualization can be knowledge enriching. Graph-based visual analysis is a highly effective method for capturing and understanding relationships between data that are not quantitative in nature. However, effectively presenting unstructured intelligence is a real challenge. When we discover a collection of time varying themes of discussion around a query over various social media streams in real-time – representing them meaningfully also requires innovation.

Nuvento

Our approach

Social networking sites provide interface to obtain data in various flavors. However, they necessitate that the connection to be authorized first. This authentication followed by authorization procedure is primarily handled by building an application following Open Authorization standard, namely OAuth 1.0. Once authorized by a social networking site, gathering various non-private data of users are possible via API calls. However, the challenge is in maintaining a long duration connection that can stream real-time conversations. Social networking sites currently do not provide such guarantee on reliability of long duration connection. Nuventos' watchdog software mode helps us keep an eye on long running queries to ensure that no conversation is missed.

Using your query and interests expressed as words we gather real-time conversations from various social networking sites and traditional electronic media like blogs and product review sites. Queries could be simple list of words, or simple logical combination or even phrases. As we start receiving data from various sources, we run our language identification technology to identify language and encoding of these text messages or data. The collected data also includes the geo-psycho-demographic profiles of the people in conversation.

It is not difficult to observe that social media is inherently noisy, and filtering out what is meaningful requires throwing out undesirable and repetitive content. Effectively identifying spam when messages are short and written with many abbreviations can be challenging. Mathematical medaling helps us assign a usefulness-indicative probability independently to each word of a message. Another formula then selects the message for which summative probability indicates it as useful.
Once we have the input data in a clean format, we store them in scalable and highly efficient storage. We use No-SQL storage so that very efficient horizontal scaling is possible. The storage is inherently fault tolerant and every bit of data is replicated in multiple machines. However, the biggest benefit comes from this type of storage in servicing heavy read/write workloads, which is typical in analytics scenario.

Solution

Gathered and stored information allows us to build first level metrics and basic comparative trends of following kinds:

- Daily Volume: Day wise volume of texts categorized by geo-psycho-demographic measures with respect to the query of interest.

- Share of Voice: We allow you to create criteria for clustering the voices and then volume of each category is presented in suitable manner.

- Author Tags: We collect the tags that authors often assign to their media, e.g. in blogs as tags, or Hash-tags with tweets. Metrics and comparative trends of these tags in your area of interests are also made available.

- Trending of recurring key words and phrases: at heart of qualitative data analysis often lies the task of discovering themes. Such themes are more easily captured by recurring key-words or phrases, and they can provide you more insights into how brands are perceived by customers. We build the history of such trends over time to fine granularity so that you can understand how perception changes over time.

The real strength of Nuvento'ssolution is in the features that we provide with our state-of-the art text analytics algorithms. To start with, we analyze the sentiment in the text and identify positive, negative and neutral expressions in them. We treat this as a classification task of labeling a text as expressing either an overall positive, negative, or neutral opinion.

Nuvento

Automated classification of any kind has an operative margin of error and inherent in any solution offered in the market. If you integrate any sentiment analysis solution with your Social CRM application -a false positive in such algorithm is ignoring a disgruntled customer for you and a false negative is wasting valuable time to think a positive or neutral feedback as negative.

Recognizing this, our approach has three stage counter measures to reduce both kinds of errors. In our implementation, we execute multiple state-of-the-art algorithms on the same input and take majority vote to conclude the classification. This significantly reduces the misclassification rate. Secondly, we use Industry-specific taxonomy to train our algorithms to reduce mistakes, and thirdly we provide power in your hand to mark or denote errors as soon as you notice them, so that our algorithms can learn and improve over time even after deployment.

Along with sentiment, we provide comparative opinion mining with respect to selected feature space. This allows you to compare your product with that of your competitors by defining a set of features for comparison. We mine review sites, blogs and social buzz in order to build a model and compare product features, as expressed by reviewers, and as conversed by your customers in the social networking sites.

Another innovation from Nuvento research lab allows us to identify which voice is actionable and which buzz is just informative. While sentiment analysis provides you the power to understand voice of customer, our ability to understand actionable item allows you to focus on addressing problems, thus cutting down on time of response significantly. Actionable item identification algorithm relies on rules of linguistic patterns and in understanding and classifying them with precision. Indeed, to facilitate a smoother functioning of the customer interaction – our API can inject actionable items directly into your CRM software, or your mailbox. API can also be configured to create delegation rules based on responsibility, severity, or the tone of the message.   Furthermore, API can also be configured to automatically reply or post back (e.g. private message in Facebook and LinkedIn, or @userid message in Twitter) selectively with template messages.

Our API capability includes full text search on the collected corpus of data of your interest. While traditional search engine indexes text by terms or words appearing in the content, we employ state-of-the art semantic analysis algorithms side-by-side traditional Tf-Idf indexing. We built statistical models for discovering the abstract themes that occur in a collection, and cluster buzz based on such underlying themes. Search queries are not just limited by words appearing in the voices but extended to underlying themes that algorithms automatically discover. To ensure high performance storage, and retrieval of relevant informationwe use NoSQL database with options for single machine or cluster deployment, depending on the data volume and are always scalable by adding one more machine.

Identifying underlying themes of a conversation allows us to automatically tag them by the theme. Thus every individual buzz within our system is tagged by its geo-psycho-demographic features, sentiments, key-words, phrases, themes, and whether it is actionable. Along with this we also allow you to add customizable tags on any item and filter them at will using any of the generated or added tags.

Such filtered collection of buzz can then be subjected to closer scrutiny. At times a collection may be large in number and recognizing this we have developed multi-buzz summarization technique. This innovation from Nuventoresearch lab can create a synopsis from a collection of text messages. Our multi-buzz summarization technique is capable of important content selection, and generating a readable presentation. One such readable presentation mode also allows the presentation in the form of a timeline.

For more information on Nuvento solutions please contact us:

email: info@nuvento.com
call: +1 (913) 338 4400
www.nuvento.com

Nuvento