Leveraging Social Media Analytics for Competitive Advantage

May 2012
Overview - Social Media and Vertica

From the Internet’s earliest days computer scientists and programmers have worked to make it easier for machines to understand human beings. Tim Berners Lee, the inventor of the World Wide Web, was a strong advocate of developing “The Semantic Web”, where machines can understand the meaning of written text. The W3C foundation took the leadership role in creating open standards to find, share and create machine understandable data formats in support of this vision. Adoption of Semantic standards was slow until the social media revolution hit and companies became very interested in what people were commenting / tweeting on sites like Facebook and Twitter.

Companies discovered that the popularity of social media creates a viable forum for the voice of the customers, Facebook and Twitter’s combined membership is over 1 billion people worldwide. Internet users also blog and add comments to restaurant reviews on sites such as yelp.com and travel reviews at traveladvisor.com. This creates a major opportunity for companies to harness customer perceptions and use them to better understand what people are saying on a topic or about a company.

Finding meaning in people’s comments is a Big Data challenge. Many social media sites provide access to customer data via public API’s; however, combing through data in a variety of formats and extracting comments relevant to the subject of interest is a significant challenge. After extracting the data and determining the relevance of each comment, assessing the sentiment (positive, negative, or neutral) creates a further challenge. The volume of data, rapid pace of social opinions, and the time value of information all place scale and latency demands on the process — the value of social data is directly proportional to how quickly a company acts on it.

This paper will outline the advantages of using the Vertica Analytics Platform to collect, parse and process social media data using Natural Language Processing (NLP) techniques. Vertica is a “New SQL” Relational Database Management System (RDBMS) currently used by media and entertainment companies who understand and act on social media data as well as those organizations powering the social media world. Vertica is a linearly scalable, high performance Big Data platform that supports standard SQL, User Defined Functions (UDF’s) and connectors to other data sources including Hadoop.

The Challenge of Natural Language Processing and Big Data

The field of natural language processing (NLP) is in its infancy with both commercial and open source tools emerging to handle the complexities of finding meaning in text. A generalized process for NLP consists of connecting to a data source, collecting text, parsing each phrase or sentence, dividing each word into tokens, assigning parts of speech to each token, assembling tokens into logical chunks and then performing semantic analysis.

Although the process of parsing and categorizing text is straight forward, performing semantic analysis is complex. For example a person can read the phrase “blow it up” and easily determine
if the context is a hand gesture or a threat. Programming computers to perform this feat is much harder to do. Companies that ensure machines can consistently handle these distinctions at Internet scale will reap rewards in the marketplace.

Semantic analysis is performed using advanced statistical techniques and algorithms. Common methods of evaluating text include dividing text and words into n chunks [known as n-grams], reducing inflected words to their stem [e.g. “catlike” to “cat”], and measuring the “edit distance” between words or phrases.

“The cat in the hat.”

Unigram: “the” “cat” “in” “the” “hat”

Bigram: “the cat” “cat in” “in the” “the hat”

Trigram: “the cat in” “cat in the” “in the hat”

Figure 1.0 — The word “cat” is a collection of concepts rather than a straight lexical definition

Using Microformats for Natural Language Processing

Creating web content that is structured in a consistent format greatly helps make pages searchable and readable by people and machines. Online resumes and recipes are examples of information that is added, updated and read with high frequency and where semantic conventions known as Microformats make natural language processing easier.

Microformats designed to solve easy problems build upon existing and widely adopted open standards. They are gaining in popularity as companies such as LinkedIn, Facebook and
Google have adopted them. Google adopted a Microdata format called Rich Snippets to make search results more readable.

LinkedIn uses hResume as a standard for online resumes. Blogging sites have adopted XFN to better represent relationships. The table below summarizes popular and emerging formats:

### Examples of Popular Microformats

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>hCard</td>
<td>Contact info, an open standard 1:1 equivalent to Outlook vCard</td>
<td>Identify People</td>
</tr>
<tr>
<td>hCalendar</td>
<td>Open source iCalendar compatible events</td>
<td>Increasing in popularity for event information</td>
</tr>
<tr>
<td>hResume</td>
<td>CV information</td>
<td>Online resume information gaining traction due to sites like linkedin.com</td>
</tr>
<tr>
<td>hRecipe</td>
<td>Open standard for recipes</td>
<td>Popularized by recipe and blog sites including Wordpress</td>
</tr>
<tr>
<td>Microdata</td>
<td>W3C method of labeling content to describe a specific type of information</td>
<td>Google recommends format for Rich Snippets</td>
</tr>
<tr>
<td>XFN</td>
<td>Represent human relationships through URL’s</td>
<td>Popularized by blogging platforms - e.g. Blogger.com</td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
<td>Usage</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>Open Graph Protocol</td>
<td>Facebook Standard API</td>
<td>Access Facebook data through a standard set of API’s</td>
</tr>
</tbody>
</table>

The Vertica Analytics Platform can play a key role in processing social media data. At the core of Vertica’s offering is a fully parallelized True Column Store™ relational database that can be loaded and queried continuously eliminating latency for time sensitive analytics operations. Additionally, Vertica uses sophisticated encoding and compression algorithms to efficiently store and process the data throughout the entire processing pipeline. Columnar orientation and execution is an important feature for natural language processing, as row based databases will not be able to perform the queries required to parse and analyze semantic data.

A developer community has formed around the Vertica Analytics Platform SDK - with over 2,000 active participants. Several of the functions developed thus far are being used by companies around the world to analyze social data and better understand how customers perceive movies, television shows, products, brands, etc. Example functions (with documentation, samples, etc.) for web service integration, text analytics, and sentiment analysis are available on Github.com.

**Social Media Data Management and Vertica**

A classic data processing approach of capturing and transforming data is applicable to social media data, but much more complicated due to the large volumes and variety of sources. The complexity of processing, analyzing and storing the unstructured and structured social media data demands a new approach to the problem. Legacy database methodologies and implementations will not scale or provide adequate performance for social media analysis - organizations must rely upon an MPP scale-out architecture with advanced and extensible in-database analytic capabilities.

The Vertica Analytics Platform is designed to process, analyze and store large volumes of both structured and unstructured data while providing real-time insights to complex questions. Vertica’s native in-database analytic capabilities and user defined functions (UDF’s) were purpose-built to take full advantage of MPP parallelism, columnar storage and execution. Vertica executes complex analytics faster, and at higher scale than any other platform. Vertica offers full ANSI standard SQL support, SQL analytic functions, text regular expressions, native SQL extensions, and a User Defined Function (UDF) framework that is rapidly expanding beyond the SQL Macro UDFs currently available.

The flexibility, performance, and TCO of the Vertica Analytics Platform makes it an ideal choice for managing social media data. In addition to a growing community of developers, it also fully integrates and supports the existing ecosystem of data warehousing and analytics solutions. Developers can take advantage of industry-standard connectivity through ODBC, JDBC,
ADO.Net and a rich API in addition to the native integration and certification with a variety of tools like Cognos, MicroStrategy, Tableau, Informatica and others. Media and entertainment customers are currently using Vertica to process and analyze social media data and then take advantage of existing investments in reporting and visualization tools to create impactful reports and visuals that convey meaning and engage decision-makers.

Conclusion

The Vertica Analytics Platform is an ideal choice for companies interested in social media and natural language processing. Customers can take advantage of the growing developer community of open source algorithms and analytical functions as well as a mature ecosystem of data warehouse tools and reports to create a competitive advantage in understanding people’s comments, blogs, and tweets.