

White Paper



Semantics Driven Consumer Insights

Leveraging Unstructured Data to Hear the Voice of the Consumer

Carey Chou, Kishor Gummaraju, Muralidhar Prabhakaran & Vaidyanatha Siva

Overview

Intense competition and ease of access to information have made it imperative for Consumer Packaged Goods (CPG) companies to listen to and understand the specific needs and feedback of shoppers. Traditionally, CPG companies have leveraged structured data sources to calibrate brand performance. However, with technology enabling new media vehicles, upwards of 80% of the consumer insights reside in unstructured data sources like consumer complaints, emails, presentations and, more recently, Internet blogs¹. Leveraging insights trapped in these sources is critical to strengthening brand loyalty and improving market share.

With the advent of Web 2.0 and related technologies, companies can effectively use Natural Language Processing (NLP) with semantics analytics and traditional structured data analysis. This 360-degree view of the shopper can provide insights into the shopper's mind that can help drive product innovation, marketing mix decisions and retail execution.

This thought paper addresses the current business challenges in the CPG industry along with the drivers that are driving the evolution of current solutions to glean consumer insights from amalgamated unstructured data. Following the assessment of existing technologies, the paper provides a reference architecture adopting the elements of Web 2.0 for NLP and semantics analytics. Finally, it discusses some of the promising vendors in this space and potential business and technology extensions.

The Need to Mine Unstructured Data

The central objective of every brand is to strive for stronger loyalty and image with greater market share. With competition for shelf space intensifying, there is a pressing need to provide shoppers with a highly differentiated value proposition through the right product and marketing mix.

In an attempt to reach shoppers, CPG companies spend billions of dollars on creating and launching new products. In addition, they also spend upwards of 20% of revenues on advertising and promotions². Despite this:

The cost of new product launch failures in the CPG industry is estimated to run into billions of dollars. Approximately 95% of all new products are considered to be failures, according to a recent Nielsen Bases and Ernst & Young study³. The number of new Universal Product Codes (UPCs) introduced by CPG companies in 2003 was 96,600 up from 72,400 in 2000⁴ (Source IRI), leading to clutter on the shelf and intense competition for the shopper's attention.

In the case of product promotions, with over \$100 billion at stake every year only 1 in 5 promotions are considered successful⁵. This can often be traced to promotion design and execution or competitor activity at retail.

Determining the return on investment (ROI) of advertising expenditure remains a challenge due to complexities in the marketplace such as simultaneous occurrence of multiple events and the fragmentation of media.

Typically, CPG companies religiously monitor brand performance, ad effectiveness and consumer need states through a variety of mechanisms that include consumer research, brand health monitors, syndicated and sales data analysis, and retail execution audits.

However, most of these data analyses leverage structured data - data that sits in rows and columns, files, or databases. Even in the case of consumer research, where some amount of unstructured data is encountered in the form of free text feedback, analysis is typically done by manually converting this information into a structured format.

Surprisingly, while maximum time is devoted to analyzing structured data, most business data (estimated at upwards of 80%) is unstructured. It sits in documents, presentations, e-mail, consumer complaint databases, consumer feedback, websites and Internet blogs, and is mainly composed of text rather than numbers.

Consider blogging: The millions of bloggers and blog readers growing exponentially represent educated consumer segments with higher disposable incomes. Clearly, they are important. More importantly they are ready to voice their needs and opinions. Hearing what they say about a brand can provide the key to understanding the target market and spotting early trends that may impact brand performance.

Customer insights posted at consumer blogs

A blogger posted that he could open a Kryptonite lock easily with a Bic pen. Other bloggers confirmed this and one made a video of the process leading to costly litigation that resulted in huge damages to the company. (<http://www.engadget.com/2004/09/14/kryptonite-evolution-2000-u-lock-hacked-by-a-bic-pen/>)

“Much to our surprise, we were able to hack our **Kryptonite Evolution 2000 U- Lock** with a ballpoint pen. This \$50 lock is supposed to be one of the best for “toughest bicycle security in moderate to high crime areas”—unless the thief happens to have a Bic pen. We used to use these to lock up our bicycles, but we’re switching to something else ASAP. (Oh, and just to be trite, the pen is mightier than the lock.). Click here to watch the video ...”

In the case of the company manufacturing “high security bicycle locks”, monitoring blogs could have helped ensure timely action. Similarly, consumer complaint databases and call center feedback notes have a wealth of insights. However, given the overload of information and number of sources (some credible and others not), the challenge is to effectively capture accurate and authentic feedback.

While analyzing structured data is important, this unstructured data can provide brand managers lead indicators on consumer needs, product efficacy, competitor activity, ad effectiveness, etc., for timely action.

Technology Implications and Emerging Solutions

The challenge presented in the previous section is unique from the technology perspective. Despite the abundance of data immersed in various sources and the value thereof, the ability of current technology solutions to seamlessly extract data from these sources and glean insights is limited.

The typical solution employs disparate sets of technology elements like Natural Language Processing (NLP), text mining and stochastic-based algorithms to uncover and collate insights. Infosys believes this approach presents several challenges:

- **Actionable results predicated on human involvement** – While text mining is able to uncover patterns among different data clusters, it is difficult to consistently trace back the causality of these patterns. Lack of understanding of root causes can lead to misinterpreting consumer insights while predicting consumer trends.
- **Lack of leverage of domain context** – The ability of systems to interpret domain context is of utmost importance in the quest for patterns within text and subsequent insights. The challenge is in the struggle to integrate domain knowledge with the text mining process enabling it to weave the domain context during clustering and pattern recognition. Since the text mining technique does not understand the meaning of the data, manual intervention is required to monitor and fine-tune the execution. Obviously this process is not scalable
- **Lack of multi-layer deduction** – Another challenge for CPG companies is that text mining cannot differentiate between an entity and the corresponding attributes that are part of the entity. The inductive reasoning that today's text mining solutions are capable of is insufficient for a CPG company to act on. With a multi-layer deduction approach, it is possible to decipher entities, attributes and their relationships enabling CPG companies to make accurate decisions. Recent technology developments have served to narrow the gap in the “text mining” space. Notable is the evolution of “semantic analytics”, which extends the capability of text syntactic pattern analysis to the semantic meaning in unstructured data. Some new technologies enabling this concept are:
- **NLP** – Natural Language Processing has been a prominent research topic for years. NLP is capable of deep text parsing in collaboration with language grammar analysis. NLP identifies terms (entities) and their corresponding characteristics (attributes), which exist as descriptive forms compliant with the grammars. Although NLP focuses on text syntax, it is able to generate an output form that relates entities and attributes.
- **Semantics Processing** – While NLP can identify entities and attributes, it does not understand the associated semantics. In the past, NLP results were generated in disparate proprietary formats, making it difficult to employ a common semantic processing technology. With the creation of the Semantic Web, there is opportunity for text analytic solutions to use a common standard framework in describing semantics that can be shared and reused. A prominent example from the Semantic Web is Resource Description Framework (RDF) and Web Ontology Language (OWL), which is based on Description Logics (DL), a language of knowledge representation. Given a standard framework like RDF/OWL, text analytic engines transform NLP results into the standard format, making them reusable and expandable across different domains.
- **Ontology Repository** -- It is critical to understand the relevance of ontology in the context of semantics processing. Ontology is the hierarchical structuring of knowledge about things by sub-categorizing them according to their essential qualities. At the grammar level, the meaning of identified entities and attributes is neutral to the domain context. During semantic processing, such entities and attributes are transformed so they can be translated in the context of target domains. Such transformation requires an existing ontology repository, which may be created manually to begin with, and enriched semi-automatically over time through semantic processing. There is a wide range of ontology technologies and tools to help users create ontology repositories. With the adoption of the Semantic Web, many of these support RDFS/OWL as an ontology representation language.
- **Semantic Reasoning and Inference** – There is a key difference between text mining and semantic analytics solutions. During semantics processing, the semantic analytics engine can detect semantic conflicts and try to reconcile or resolve them with semantic reasoning. Further, with the help of inference rules, the engine can infer additional meanings or extend existing relationships to other entities. The results of semantic reasoning and inference can be used to augment the semantic ontology. There are products available in the market place that provide sophisticated stochastic algorithms that, used against semantic data, make analysis results more accurate and actionable in contrast with text mining.

- **Semantic Search** – A key benefit of text analytics is that it enables search with more relevant results. Several enterprise search product vendors are adopting the Semantic Web technology, which elevates search capabilities from syntactic pattern recognition to semantic relationship discovery. As a consequence, search is processed not only from keywords, but also by navigating entities and their relationships.

Text Analytic Solution: Reference Architecture

To address the needs of CPG enterprises, Infosys proposes a six-layer reference architecture for a solution to seamlessly extract, process and leverage insight from various unstructured sources (Figure 1).

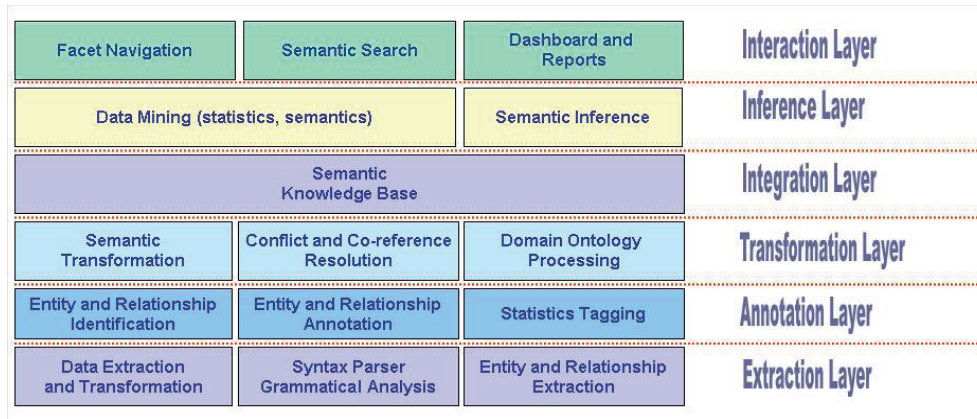


Figure 1 – Reference Architecture for Semantic Analytics Solution

Extraction Layer – Using NLP in collaboration with data access technologies that fetch and aggregate data from disparate sources, this layer extracts all detected entities and relationships and feeds them to the “Annotation layer”.

Annotation Layer – Taking the input from NLP, this layer runs through a series of recognition steps to tag entities and their relationships using pre-defined ontology dictionaries. As a result, the annotation layer creates a list of entities of interest and their corresponding relationships in a form of a semantic graph.

Transformation Layer – While the annotation layer creates a semantic graph, some relationships may be contradictory or ambiguous. With the help of semantic reasoning and existing domain ontology, the transformation layer validates each entity and relationship to fix potential conflicts; some relationships may be dropped and some established due to the resolution. Sometimes, this step also requires manual intervention. As a result of reasoning, the consistent semantic graph it creates can be used to augment existing domain ontology or export to other external ontology repositories.

Integration Layer – This is a central repository layer that stores results from the transformation layer. As a general practice, a semantic knowledge base also equips a semantic-based ontology to apply ontology rules for validation of instance data before it is stored. The integration layer establishes semantic relationships between structured and unstructured data, with which it maintains consistent meanings across different data communities.

Inference Layer – This layer combines the power of data mining and semantic inference. With a semantic knowledge base in place, inferences can be drawn either by ontology rules or using stochastic algorithms. Statistics based approaches are often used to uncover affinities among entities or to discover unknown relationships. The inference layer is where domain experts work to define and validate hypotheses and eventually derive conclusions. In this layer, new identified relationships may become part of the ontology base.

Interaction Layer – This is the layer most end users interact with. A de facto application at this layer is semantic search, which searches the semantic knowledge base with keywords. Some sophisticated applications adopt a semantic query function which gives users finer control of search execution. Such search applications often incorporate a compelling navigation capability forming hierarchical navigation trees by clustering search results based on entity characteristics. With such capability, users can drill down or up search results by navigating the tree. Other possible applications include reporting and dashboard applications that leverage the combined power of both structured and unstructured data and create conjoined views of consumer insights.

Key Players in Semantic Analytics

Leveraging the reference architecture, Infosys has evaluated several vendors, aligning their product features to each layer.

Vendors like ClearForest and Siderean offer robust extractions and annotation solutions with deep domain tagging. Siderean and Attensity also provide sophisticated analytics solutions at the interaction layer with text search and drill-down navigation capabilities. Though most vendors do not provide out-of-the-box capabilities covering all layers, some are moving aggressively to bridge these gaps by extending partnerships with independent software vendors (ISV) or system integrators (SIs).

Customer Insights: Proof-of-Concept

Generating customer insight from unstructured data (online feedback form sites like Amazon), call center records, customer complaints, blogs, and newswire reports could be critical to the product innovation process for CPG companies and the key to their success. Infosys conducted a pilot with a start-up vendor covering the scope of the problem to mine customer insights from –

- One website – Amazon.com
- One product line – Digital cameras
- Two competing / leading manufacturers – Canon and Nikon

Methodology

- Define a “controlled vocabulary” for the “bounded domain” of digital cameras, leveraging manufacturer product description and customer feedback
- Create the controlled vocabulary by manually identifying entities and predicates from product description and customer feedback, across a small sample of products
- Based on the controlled vocabulary, create a problem statement definition, scope and deliverables that are realistically achievable and add value
- For each track, build an effective guided navigation capability - an emerging technique wherein the semantics and relationship between attributes or facets of data is better qualified. This approach increases the effectiveness of search on a website
- For each track, build queries and basic insights (even transitive relationships are allowed) and achieve deliverables

Approach

- Request digital camera data from Amazon web services as feed
- Receive camera data and load it into an Resource Description Framework (RDF) model
- Request RDF Site Summary (RSS) feed with reviews per product through RSS connector
- Apply sentiment analysis on the data and update RDF model with the results of sentiment analysis
- Web browser requests camera data with brand, average rating and sentiment filters
- Web browser receives navigation results with bar graph visualization

Deliverables

- Faceted navigation capability for each manufacturer (Canon and Nikon)
- Insight into competition capability – in areas where products compete
- Inference problems with products. Inference capabilities customers wished products had, after they used them



Figure 2 - Sorting by Sentiment Analysis

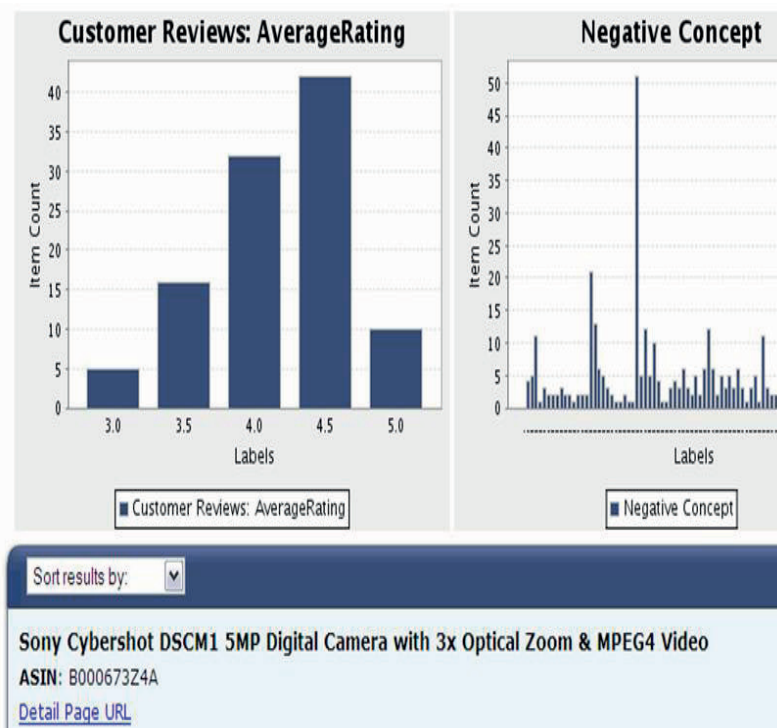


Figure 3 - Drill-down to analysis of one camera

great value for money and offers video and all kinds of stuff. If you're looking mostly looking to take landscape or portrait stills, this camera will do a good job!, I really like this camera! It is fairly simple to use, takes pictures and is packed with features. It fits nicely in the hand and is compact enough to slip in a coat pocket or purse.

We bought this camera for our daughter's birthday. She is using it with ease & especially loves the movie with sound option. In my opinion, the camera is a winner for any adult or kid wanting a good all around digital camera.

My only negative is that it is a battery hog. As with most digital cameras, you will want to **make sure you get rechargeable batteries** and large memory card. This is my 3rd digital camera and I bought this one to supplement my larger 10X zoom camera. This one fits into a pocket or small camera bag to easily take it to all events or for travel. I've found that it takes nice photos indoors when the flash is used...Even when the cats or dogs I'm shooting are moving around. I've found the life of a battery set is pretty good, but I've been using nice rechargeable ones. Definitely get a set then you have 2 to use and 2 as a backup set. The settings are pretty easy to figure out and change and 99% of my photos were taken on the "auto" setting and came out pretty good. As others have indicated indoors - without flash - can produce blurry photos, but that's the nature of almost all cameras, not unique to this model. Glad I bought it and would highly recommend it.

Customer Reviews: Date: 2005-12-21, 2006-01-07, 2006-08-15, 2005-09-15, 2005-12-19

Customer Reviews: Summary: Good deal, Best camera for the price, Some Important CONS you must be aware of, Excellent Camera with Many Features, Great camera - especially at this price

Customer Reviews: Rating: 4, 5

Positive Concept: [camera](#), [camera](#), [features](#), [camera](#), [bag](#), [battery](#), [Auto](#)

Negative Concept: [camera](#), [feature](#), [close](#), [battery](#), [cam](#), [capture](#), [blurry](#)

Figure 4 - Further drill-down to each customer review

Where this Technology can be Extended

Semantics can be applied wherever knowledge management of unstructured data sources is needed. For example, significant advances have been made in the life sciences industry with the creation of medical informatics ontologies –

- Foundational Model of Anatomy (FMA) Gene Ontology
- Health Level 7 (HL7) data types and top-level RIM classes
- Guideline Interchange Format (GLIF)

Of course, this can be extended to mining information from both structured and unstructured data in a dashboard format. For instance, a brand health dashboard for a Consumer Goods Brand manager could tie syndicated data, demographic information, new product launch information (from structured data sources) and customers feedback knowledge (from blogs, customer complaints, panel data) and use this dashboard to provide a timely view to brand health and provide the ability to act on information. This could be a powerful tool that integrates both unstructured and structured data.

Conclusion

There is clearly a need for CPG companies to look at unstructured information to better manage their brands. Current and emerging technologies can effectively deliver the solution required. Leveraging Infosys' reference architecture and prescribed approach, CPG companies can accelerate their adoption and implementation of these technologies for competitive advantage.

References

1. IBM Unstructured Information Management Architecture (UIMA) <http://www.alphaworks.ibm.com/tech/uima>
2. IBM Services for UIMA based Knowledge Integration (SUKI) <http://www.research.ibm.com/UIMA/SUKI/index.html>
3. Web Ontology Language (OWL) <http://www.w3.org/TR/owl-ref/>
4. Resource Description Framework (RDF) <http://www.w3c.org/RDF>
5. The Description Logic Handbook: Theory, Implementation Applications: F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider, Cambridge University Press, Cambridge, UK, 2003:

Acronyms & Abbreviations

CPG – Consume Packaged Goods

NLP – Natural Language Processing

OWL- Web Ontology Language

RDF – Resource Description Framework

RSS - RDF Site summary

Sources For Quotes / Figures

1. <http://www.b-eye-network.com/view/2098>
2. Infosys Analysis.
3. http://www.copernicusmarketing.com/about/docs/six_sigma_accountability.htm
4. Information Resources Inc. IRI
5. Information Resources Inc. IRI
6. Figures 2, 3 & 4 courtesy of Siderean Software, Inc.

About the Authors

[Muralidhar Prabhakaran](#) is a Principal Architect with the Retail, CPG and Distribution Business Unit of Infosys. Muralidhar has several years of experience executing complex data management projects. He can be contacted at Muralidhar_P@infosys.com

[Vaidyanatha Siva](#) is a Principal Architect and heads Technology Architecture and Innovation for the CPG Industry vertical in Infosys. He can be contacted at siva_vaidyanatha@infosys.com.

[Carey Chou](#) is a Senior Technical Architect with CPG Solutions at Infosys. He can be contacted at carey_chou@infosys.com.

[Kishor Gummaraju](#) is Associate Vice President and Head of CPG Solutions at Infosys. He has several years of experience in working with leading CPG companies and retailers. He can be contacted at Kishor_gummaraju@infosys.com



For more information, contact askus@infosys.com

About Infosys

Many of the world's most successful organizations rely on Infosys to deliver measurable business value. Infosys provides business consulting, technology, engineering and outsourcing services to help clients in over 30 countries build tomorrow's enterprise.

For more information about Infosys (NASDAQ:INFY), visit www.infosys.com.