

## Leverage the voice of your customers.

An Anderson Analytics and SPSS predictive text analytics case study. In the new information age, any customer can become a brand enemy or evangelist and reach millions of other customers on the web. These opinions, expressed freely on the internet, encompass attitudes, thoughts and behaviors of past, current and potential customers.



Companies are rushing to tap into the explosion of customer opinions expressed online. The most innovative companies know they could be even more successful in meeting customer needs, if they just understood them better. Text analytics is proving to be an invaluable tool in doing this.

Anderson Analytics, a full-service market research consultancy, tackles this issue using cutting-edge text analytics and data mining software from SPSS that allows the application of linguistic, statistical and pattern recognition techniques to extremely large text data sets.

A text analytics project is usually part of a much larger data mining project that would typically involve the identification of some core strategic questions, the allocation of resources and the eventual implementation of findings.

However, the focus of this case study is to describe the tactical aspects of a text analytics project and to delineate the three basic steps involved in text analytics:

- **Data Collection and Preparation**
- **Text Coding and Categorization**
- **Text Mining and Visualization**

In this case study, Anderson Analytics “content-mined” data available on Flyertalk.com’s discussion boards. Flyertalk.com is one of the most highly trafficked travel domains. It features chat boards and discussions that cover the most up-to-date traveler information, as well as loyalty programs for both airlines and hotels.

Note that the text analytics techniques applied in this case are not limited to discussion boards or blogs but can be applied to any text data source, including survey open ends, call center logs, customer complaint/suggestion databases, emails, etc.

# Step 1: Data Collection & Preparation

Having quality data in the proper format is usually more than half of the battle for most researchers. For those who can gain direct access to a well-maintained customer database, the data collection and preparation process is relatively painless. However, for researchers who want to study text information that exists in a public forum such as FlyerTalk.com, data collection can be more complex and usually involves web-scraping.

**“It is crucial for us to get an understanding of how our most loyal customers think and what they value most in their travel experience... any data that helps us meet our loyal customers and attend to needs of future customers is worth its weight in gold.”**

**Robin Korman,  
Vice President, Starwood  
Loyalty Marketing Program**

Even with the availability of powerful web-scraping tools and techniques, text mining a popular blog or a message board like the one at FlyerTalk.com presents unique data collection and processing challenges. The amount of free text available on such sites usually prohibits an indiscriminate approach to data scraping. A strategy with clear objectives and a well-defined data extraction method are needed in order to increase the reliability of data analysis in the latter stages of the research.

In this particular case, researchers at Anderson Analytics narrowed the scope to just discussion topics within a 12-month period (from August 2005 to August 2006) on the five major forums intended for discussing the hotel loyalty programs of Starwood, Hilton, Marriott, Inter Continental, and Hyatt hotels.

Specific web-scraping parameters differ depending on the structure of the target sites. In a discussion board format, the text data tend to follow a simple hierarchy. Typically, each forum contains a list of topics, and each topic consists of numerous posts. Therefore, the web-scraping process of FlyerTalk.com initially retrieves data such as the

discussion topics, topic ID, topic starter, and topic start date. Then, by using the topic ID, the web-scraping application constructs and submits query strings to the FlyerTalk.com site to retrieve messages associated with each specific topic.

A good web-scraping tool should allow the capture of information that exists in the source data of an html page, not just the displayed text. Therefore, hidden information such as the topic ID, date stamp, etc. also becomes available to the researcher.

Besides making sure the fields in the final dataset are in the correct format, another problem unique to discussion board text needs to be addressed. It is very common for posters to quote others' text within their own posts. These quotes should typically be extracted from the message field and placed in a separate field so as to prevent double counting and inadvertently weighting certain posts.

In addition to the text messages posted on the forum, the web-scraping process should also capture the poster's ID and 'handle' as well as any other available poster information such as forum join date and forum registration information (in this case: location, frequent traveler program affiliation, etc.)

Web scraping (or screen scraping) is a technique used to extract data from websites that display output generated from another program. There are many commercially available applications that can scrape a website and turn the blogs or forum messages into a data table.

Even with the availability of powerful web-scraping tools and techniques, text mining a popular blog or a message board like the one at FlyerTalk.com presents unique data collection and processing challenges. The amount of free text available on such sites usually prohibits an indiscriminate approach to data scraping. A strategy with clear objectives and a well-defined data extraction method are needed in order to increase the reliability of data analysis in the latter stages of the research.

## WEB SCRAPING PROCESS

### CRAWL

Crawl the website and scrape for topic, ID and thread initiator.

### DOWNLOAD

Use topic ID from the first step as part of the URL query string to download messages.

### STORE

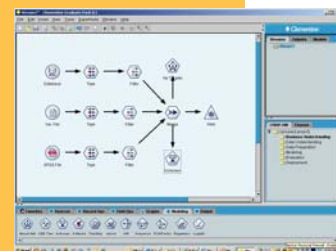
Web crawl and store message display pages.

### SCREEN SCRAPE

Screen scrape stored webpages and extract data into a structured format.

### LINK

Link extracted posts with topics from the first step, along with other extracted fields to create the final dataset.



# Step 2: Text Coding & Categorization

**T**ext coding and categorization is the process of assigning each text data record a numeric value that can be used later for statistical analysis.

Text coding can apply either dichotomous codes (flags & many variables) or categorical codes (one variable for an entire dataset). Short answers to an open-ended survey question typically use categorical codes. However, the amount of text included in most discussion board posts typically requires dichotomous codes.

Text coding is usually an iterative process. This is particularly true for coding messages on a site such as FlyerTalk.com. This is because compared to survey answers, the text information from discussion boards tends to be less focused. The text data on most discussion boards tend to be “user-driven” rather than “provider-driven.”

Before creating categories, researchers at Anderson Analytics first randomly examine a sample of text messages to gain a basic understanding of the data. This step is required to understand the type of acronyms, shorthand and terminologies commonly used on the forum of interest.

SPSS Text Analysis for Surveys and Text Mining for Clementine are powerful tools. However, the text coding results can be greatly improved if the programs can be “trained” to better understand text information particular to the industry and topics of interest.

With a list of industry specific themes, concepts and words, the researchers at Anderson use tools such as SPSS Text Analysis for Surveys to create a custom dictionary. Then the SPSS text analytics applications can be used, in conjunction with an SPSS developed dictionary, to extract highly relevant concepts from the text data.

**“I strongly believe that the travel industry can not only learn by listening to their customers in real-time, but that by being active where the customers are on the Internet, they can create a unique generation of loyal, repeat customers... Customer service these days is where the customer is, not where you are”**

**Randy Petersen,  
CEO WebFlyer.com**

## CODING & CATEGORIZATION

### PRELIMINARY CODING

Use both computer and human coder to obtain the preliminary understanding of the data.

### INITIAL CLASSIFICATION

Use SPSS Text Mining tool to perform initial categorization on a sample data set (1/100 of the entire dataset).

### COMPUTER CLASSIFICATION

Information and knowledge gained from the initial concept extraction is used by human coder to assist in computer categorization.

### CODING & CLASSIFICATION REFINEMENT

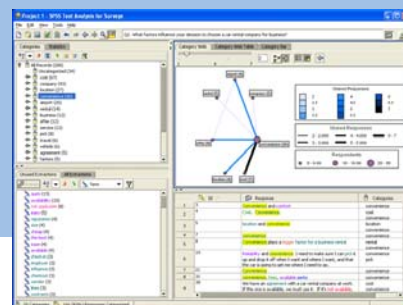
Categorization and coding are an iterative process. Custom libraries are created to refine the process. Text extraction is performed multiple times until the number of and the details of categories are satisfactory.

### CODING & EXTRACTION RULES

Once the coding result becomes satisfactory, the same coding and extraction rules are used on the entire dataset.

### CATEGORIZATION RESULTS

Categorization results are exported for further analysis with tools such as SPSS Text Analysis.



In this case, examples of some of the basic concepts in the messages that can be detected by the software include: ‘rates’, ‘stay’, ‘breakfast’, ‘points’, ‘free offers’. The text extraction and categorization processes are repeated with minor modification each time to fine tune results.

# Step 3: Text Mining & Visualization

The coded text data can be interpreted in many different ways depending on the needs of any given research project. In this case, the data is examined via the following methods:

## Positive/Negative comments and overlapping terms

The Flytalk.com data indicate that negative discussions among the posters are centered on the payment process, condition/quality of the bathroom, furniture, and the check in/out process. The praises seem to be centered around topics such as spa facility, complimentary breakfasts, points and promotions.

## Data patterns within different hotel brands

By comparing the coded text data of Starwood and Hilton forums, the researchers find that the posters seem to be relatively more pleased with beds on Starwood's board, but more pleased with food and health club facilities on Hilton's board.

## Longitudinal data patterns

As this study contains data from a one year period, data can be analyzed to understand how topics are being discussed on a month-to-month basis. The data in this particular case revealed that the discussion about "promotions" on the Starwood board was particularly frequent in February 2006. Cross-checking with Starwood management confirmed that special promotions were launched during that time period. This demonstrates one way to measure the impact of various communication strategies, promotions and even non-planned external events.

## Analysis of Poster Groups

Web mining may be helpful in understanding the aggregate motivation of some of the most active users of the products. Though it may be difficult to segment posters with only one post, frequent posters can provide a relatively rich set of segmentation variables. In this case, some general motivational themes found were the need for being "in the know", "finding deals", and the desire to "give back".

**"Understanding the needs of your customers as well as the strengths and weaknesses of your competitors is paramount for building brands that people love. Thus, I believe that the mining of customer comments on the web will become a cornerstone for future innovation and the detection of competitive threats."**

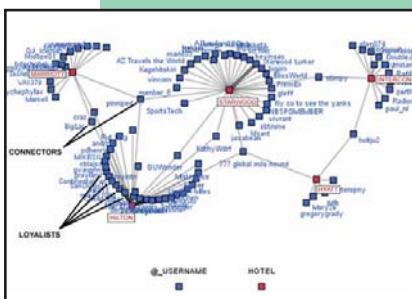
*Isaac Collazo,  
Vice President, Performance Strategy & Planning  
InterContinental Hotels Group*

## CONCLUSION

Companies have found that they can compete far more effectively if they gain a true, 360° view of their customers. The feedback that current and potential customers provide in blogs, forums and other online spaces provides a rich source of feedback. Using text analytics to monitor this information helps organizations gauge customer reaction to products and services and, when combined with analysis of "structured" transactional data, delivers predictive insight into customer behavior.

This paper described how text analytics was applied to information posted by users of travel and hospitality services; but the same techniques can be applied to other industries. A company might find, for example, that when it launches a special promotion, customers mention the offer frequently in their online posts.

Text analytics can help identify this increase, as well as the ratio of positive/negative posts relating to the promotion. It can be a powerful validation tool to complement other primary and secondary customer research and feedback management initiatives. Companies that improve their ability to navigate and text mine the boards and blogs relevant to their industry are likely to gain a considerable information advantage over their competitors.



## ANALYSIS

Super Posters—  
Connectors vs. Loyalists.  
Analysis using  
Clementine's Web  
Visualization Tool

# Find out how you too can harness the power of text analytics. Contact Anderson Analytics & SPSS.

## About Anderson Analytics

*More than Market Research*, Anderson Analytics is a next generation marketing consultancy that combines new technologies, such as data and text mining, with traditional market research. We focus on helping clients *Gain the Information Advantage* by combining the efficiencies and business experience found in large research firms with the rigorous methodological understanding from academia and the creativity found only in smaller firms. Our clients put their customers first and so do we, visit our website to learn about “The AA-Assurance.”

## About SPSS Inc:

SPSS Inc. (Nasdaq: SPSS) is a leading global provider of predictive analytics software and solutions. The company's predictive analytics technology improves business processes by giving organizations forward visibility for decisions made every day. By incorporating predictive analytics into their daily operations, organizations become Predictive Enterprises—able to direct and automate decisions to meet business goals and achieve a measurable competitive advantage. More than 250,000 public sector, academic, and commercial customers rely on SPSS technology to help increase revenue, reduce costs, and detect and prevent fraud. Founded in 1968, SPSS is headquartered in Chicago, Illinois.

## SPSS TEXT ANALYSIS FOR SURVEYS™

SPSS Text Analysis for Surveys uses natural language processing (NLP) software technologies and allows users to combine automated and manual techniques in analyzing open-ended responses to survey questions. SPSS Text Analysis for Surveys is uniquely able to distinguish between positive and negative comments and opinions, which is extremely valuable in understanding customer feedback.

## TEXT-MINING FOR CLEMENTINE®

Text Mining for Clementine enables users to extract key concepts, sentiments and relationships from textual or “unstructured” data and convert them to a structured format that can be used to create predictive models. This has been shown to improve the “lift” or accuracy of predictive data models and significantly improve results.

## About FlyerTalk.com

Founded in 1986, Frequent Flyer Services has created a unique niche for itself within the travel industry as a company that conceives, develops and markets products and services exclusively for the frequent traveler. Its focus and distinctive competency lie in the area of frequent traveler programs. Worldwide, these frequent traveler programs in the airline, hotel, car rental and credit card industries have more than 75 million members who earn an excess of 650 billion miles per year.

Headed by Randy Petersen, who *The Wall Street Journal* calls, “...the most influential frequent flier in America,” Frequent Flyer Services has had alliances with major companies such as AOL and provides, or has provided, content and information to many of the leading sites on the Internet. The company is probably most famous for its Inside Flyer magazine, which has grown from a simple newsletter into the leading publication in the world for members of frequent traveler programs.



Anderson Analytics, LLC  
154 Cold Spring Road, Suite 80  
Stamford, CT 06905  
Tel: +1.888.891.3115  
[www.andersonanalytics.com](http://www.andersonanalytics.com)



SPSS Inc.  
233 S. Wacker Drive, 11th Floor  
Chicago, IL 60606  
Tel: +1.312.651.3000  
[www.spss.com](http://www.spss.com)