Analyzing Online Discussion for Marketing Intelligence

Natalie Glance Matthew Hurst Kamal Nigam
Matthew Siegler Robert Stockton Takashi Tomokiyo

<firstInitial><lastName>@intelliseek.com
Intelliseek Applied Research Center
Pittsburgh, PA 15217

ABSTRACT

We present a system that gathers and analyzes online discussion as it relates to consumer products. Weblogs and online message boards provide forums that record the voice of the public. Woven into this discussion is a wide range of opinion and commentary about consumer products. Given its volume, format and content, the appropriate approach to understanding this data is large-scale web and text data mining. By using a wide variety of state-of-the-art techniques including crawling, wrapping, text classification and computational linguistics, online discussion is gathered and annotated within a framework that provides for interactive analysis that yields marketing intelligence for our customers.

Categories and Subject Descriptors: H.3.3: Information Search and Retrieval

General Terms: Algorithms, Experimentation

Keywords: text mining, content systems, computational linguistics, machine learning, information retrieval

1. SYSTEM OVERVIEW

The system is comprised of four main components. The content component crawls the web for weblog, message board and Usenet content and populates internal search indices. The search and relevance component uses a set of queries and a relevance filter to retrieve messages from the indices. The production component applies analyses to the relevant messages, producing a set of extracted facts. These form the project data over which the last component, interactive analysis, is used to discover marketing intelligence.

Discovery and harvesting of message data is the first component of our system. Data collection from Usenet newsgroups is straightforward. For weblogs, we collect lists of recently updated URLs from centralized services that are typically pinged by weblog authoring software. After crawling, we use a model-based approach to segment weblogs into multiple posts by discovering common xpaths to the date and title of the weblog, along with other heuristics. We complement our approach to model-based segmentation using weblog feeds when available. These contain the updated content of the weblog in standardized XML format (RSS, Atom), and are provided by a number of weblog hosting services. For online message board content, harvesting uses web-site wrapping and intelligent directed crawling to extract messages from web pages while minimizing the crawling impact on web servers.

Our content system indexes hundreds of millions of weblog, message board, and Usenet messages. For any given project, only a small fraction of these messages are relevant. The second component of our system, search and relevance, uses a two-stage approach to identify this relevant subset, combining complex boolean queries to our indices and a machine learning classifier trained by active learning. The analyst first configures the search queries and then labels training and testing sets using an active learning process that creates both a text classifier and performance estimates. This process incorporates a heterogeneous blend of traditional active learning strategies as well as strategies that leverage domain knowledge.

The third component of the system analyzes relevant messages to extract facts that represent at a fine level the expressions made within each message. A number of topic classifiers are created during configuration, by hand-writing rules (using words, phrases, stemming, synonymy, windowing and context-sensitivity) and training machine learning classifiers where appropriate. These topics can include things such as brands being tracked (e.g. *Dell Axim*) and features or concepts in the domain (e.g. *Screen*, *Technical Support*, *Price*). In addition, a sentiment (or polarity) analysis is performed to identify positive and negative language in messages. This analysis uses a shallow NLP approach based around a lexicon and semantic interpretation rules. The topics and sentiment analysis are combined to identify positive and negative expressions about brands, and their related features and concepts. Each identified fact is tied to a specific segment of text in a message. This collection of extracted facts forms the basis for the fourth component, interactive analysis, described as part of the case study.

Our work is similar to Takumi [2], an analysis system over extracted information from call center text data, and the CiteSeer project [1], providing a search interface over extracted information for research papers. Our work emphasizes challenges created by focusing on the web, and the appropriate analysis technologies to understand the data.

2. CASE STUDY

This section presents an example of how our system discovers marketing intelligence from internet discussion. A typical project will analyze anywhere from tens of thousands of messages to tens of millions of messages. The following case study presents such a project in the domain of handheld computers, including PDAs, Pocket PCs, and Smart-

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

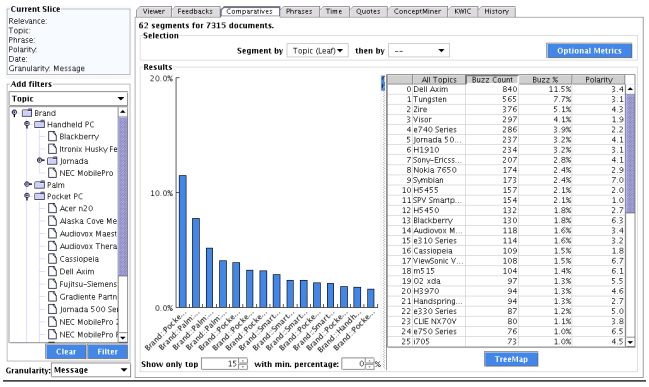

Figure 1: A breakdown of selected messages by brand, along with Buzz Count and Polarity metrics.

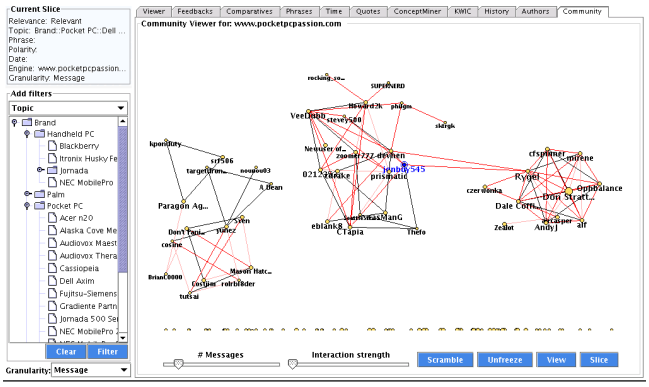

Figure 2: A display of the social network analysis for discussion about the Dell Axim on a message board.

phones. Some of the basic questions asked by our customers are "What are people saying about my brand?" and "What do people like and dislike about my brand?". This paper argues that these questions are best answered through interactive analysis of text-mined data. Manual review of the data or simple search are not comprehensive, deep or fast enough to provide robust answers to these basic questions.

Figure 1 shows our interactive analysis tool. One way to analyze messages is through a top-down methodology, that starts with broad aggregate findings about a brand, and then digs deeper to understand the drivers of those findings. The Comparatives analysis breaks down the messages and generates a variety of metrics over each segment. Figure 1 shows all handhelds discussion broken down by brand. The Dell Axim is the most "popular" brand, as measured by buzz volume, capturing 12% of all discussion. However, by a measure of overall sentiment, the Dell Axim does not do so well. The Polarity column shows a 1-10 score representing the aggregate measure of sentiment about this brand. This metric is based on the posterior estimate of the ratio of the frequency of positive to negative comments. The Axim's score of 3.4 is relatively low.

To understand drivers of this high-volume, low-sentiment discussion, an analyst selects the messages saying negative things about the Axim with just a few clicks. The Phrases tab identifies distinguishing words and phrases for negative Axim discussion through a combination of statistical and NLP techniques. Table 1 shows the top eight words and phrases, as calculated by our phrase-finding technology. Further drilling down on these words and phrases to the messages containing them reveals, for example, that a number of "SD cards" are "incompatible" with the Axim, and

Keywords	Keyphrases
Axim	Dell Axim
X5	Pocket PC
Dell	my Dell Axim
par	Dell Axim X5
today	battery life
ROM	SD card
problem	Toshiba e740
incompatible	CF slot

Table 1: The top eight words and phrases for negative comments about the Dell Axim.

that "ROM" updates are needed to make Personal Internet Explorer work correctly.

A second way of analyzing data is through a bottom-up methodology, starting with all discussion and identifying clusters of information that can be understood through interactive analysis. One of several such techniques in our application is a social network analysis. Figure 2 displays this network for discussion regarding the Axim on a popular Pocket PC discussion board. Each node in the graph is an author, and links between authors are created when authors interact by posting in the same thread. The length of each link connecting two nodes is inversely proportional to the strength of their interaction. Figure 2 shows three clusters of discussion in the message board. By selecting the rightmost cluster of messages, the analyst can quickly proceed to the Quotes analysis, which displays extracted facts with backing sentences having high sentiment about the selected brand. Table 2 shows results of this analysis for negative sentiment about the Axim within that authorial cluster. The quotes clearly indicate that a distinct group of people are unhappy about the audio and IR components of the Axim.

This case study illustrates two key points. First, an interactive analysis system can be used to quickly derive marketing intelligence from large amounts of online discussion. Second, the integration of many different state-of-the-art technologies are necessary to enable such a system.

3. REFERENCES

[1] K. D. Bollacker, S. Lawrence, and C. L. Giles. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In *Agents '98*, pages 116–123, 1998.

[2] T. Nasukawa, M. Morohashi, and T. Nagano. Customer claim mining: Discovering knowledge in vast amounts of textual data. Technical report, IBM Research, Japan, 1999.

- It is very sad that the Axim's audio AND Irda output are so sub-par, because it is otherwise a great PPC
- The Axim has a considerably inferior audio output than any other Pocket PC we have ever tested.
- When we tested it we found that there was a problem with the audio output of the Axim.

Table 2: Three representative automatically extracted negative sentences about the Dell Axim.