Chapter 21

Characterizing Buzz and Sentiment in Internet Sources: Linguistic Summaries and Predictive Behaviors

Richard M. Tong
Tarragon Consulting Corporation
1563 Solano Avenue, #350
Berkeley, CA 94707, USA.
Email: rtong@tgncorp.com

Ronald R. Yager
Machine Intelligence Institute
Iona College
New Rochelle, NY 10801, USA
Email: yager@panix.com

Abstract

Internet sources, such as newsgroups, message boards, and blogs, are an under-exploited resource for developing analyses of community and market responses to everything from consumer products and services, to current events and politics. In this paper, we present an overview of our exploration of effective ways of characterizing this large volume of information. In our approach, we first create time-series that represent the subjects, opinions, and attitudes expressed in the Internet sources, and then generate "Linguistic Summaries" that provide natural and easily understood descriptions of the behaviors exhibited by these time-series.

Keywords: Internet buzz, sentiment, linguistic summaries, marketing research, intelligence analysis, data mining, text mining, fuzzy sets, time-series analysis.

1. Introduction and Motivation

An often-unrecognized feature of the Internet is that it provides an unlimited number of forums for individuals, groups of individuals, and organizations to express their opinions about anything that concerns them. These forums include those that are inherently Internet-based, such as message boards, listservs and blogs, as well as those that are online extensions of traditional media, such as newspapers, magazines and newsletters. This vast array of "conversations" is increasingly seen, by

both Government and Industry, as rich, but mostly untapped, source of understanding of how communities and markets are responding to everything from current events, to political issues, to the latest consumer products, to cast changes on popular TV programs.

Several research groups, as well as a number of companies, have begun to explore the issues of mining these on-line sources for uses such as brand monitoring, new product feedback, assessing the impact of advertising, and image management. The recent AAAI Symposium on "Exploring Attitude and Affect in Text" (Qu et al., 2004) contains several papers on these problems and illustrates the broad range of challenges that work in this area entails. In this paper, we present the results of our own activities, with an emphasis on techniques for summarizing the aggregate behaviors and trends that emerge from the large-scale analysis of multiple on-line sources.
In the remainder of this paper, we first introduce the basic concept of a linguistic summary, in both its static and temporal forms. Then we illustrate the use of linguistic summaries in a variety of contexts using data taken from the Internet. We complete the paper with a brief overview of our data collection and analysis system, and with a short discussion of related work and open R&D issues.

2. Linguistic Summaries

A basic premise of our work is that much of the useful information in on-line sources is only apparent when we aggregate over time and across forums. This means that we are fundamentally interested in methods for combining information from disparate sources, and then with techniques for characterizing the dynamic behaviors of these sources.

In this paper, we focus on the latter, and specifically on ways to create summaries that are natural and easily understood by human intelligence analysts and decision-makers. We call such summaries "linguistic summaries" because we create them by using the mathematics of fuzzy set theory to map quantitative features of the underlying data into controlled language descriptions. We distinguish between static summaries that describe data independently of any time referent, and temporal summaries that focus on the change in data characteristics. We briefly describe each of these in the following sections.

2.1 Static Summaries

In Yager (1991) we introduced the idea of a (static) linguistic summary and described its role in summarizing information contained in a database. In this section we briefly summarize the basic ideas.

Assume V is some attribute in a database having as its domain the set X. Examples of V could be age, city of residence, years of education or amount of sales. Associated with V is a collection of elements drawn from X consisting of the values for V assumed by the objects in the database, we denote this as $D = [a_1, a_2, \ldots, a_n]$. A linguistic summary associated with V is a proposition containing meta-knowledge about the elements in D. If V is the attribute age then some examples of linguistic summaries are:

"*Most* people in the database are *about 25 years old*."
"*Nearly a quarter* of the people in the database are *middle aged*."

Formally, a linguistic summary is a statement of the form:

"*Q* objects in the database have *V* is *S*."

In the above, *S* is called the summarizer and *Q* is called the quantity in agreement. Associated with each linguistic summary is a value *T*, called the measure of validity of the summary. Given the dataset *D* the value *T* is used to indicate the truth of the statement that *Q* objects have the property that *V* is *S*.

A fundamental characteristic of this formulation is that the summarizer and quantity in agreement are expressed in linguistic terms. One advantage of using these linguistic summaries is that we can provide statements about the database in terms that are very natural for people to comprehend. A second advantage, which will be useful in database discovery, is that these types of propositions have large granularity.

With the aid of fuzzy subsets we are able to provide a formal semantics for the terms used in the linguistic summary. In a procedure to be subsequently described, we shall use this ability to formalize the summarizers and quantity in agreement to evaluate the validity of the linguistic summary. This validation process will be based upon a determination of the compatibility of the linguistic summary with the data set *D*. It should be pointed out that for a given attribute we can conjecture numerous different summaries, then with the aid of the data set *D* we can evaluate *T* to determine which are the valid summaries. In Yager (1991; 1996) we discuss methods for quantifying the amount of information contained in a linguistic summary.

In our approach use is made of the ability to represent a linguistic summarizer by a fuzzy subset over the domain of the attribute. If *V* is some attribute taking its value from the domain *X* and if *S* is some concept associated with this attribute we can represent *S* by a fuzzy subset *S* on *X* such that for each $x \in X$, $S(x) \in [0, 1]$ indicates the degree of compatibility of the value *x* with the concept *S*. If we are considering the attribute age and if *S* is the concept middle age then *S*(40) would indicate the degree to which 40 years old is compatible with the idea of middle age. It should be noted that even in environments in which the underlying domain is non-numeric using this approach allows us to obtain numeric values for the membership grade in the fuzzy subset. For example if *V* is the attribute city of residence that takes as its domain the cities in the U.S. we can express the concept "lives near New York" as a fuzzy subset. The second component in our linguistic summary is the quantity in agreement *Q*. These objects belong to a class of concepts called linguistic quantifiers. Examples of these objects are terms such as *most*, *few*, *about half*, *all*. Essentially linguistic quantifiers are fuzzy proportions. We can represent these linguistic quantifiers as fuzzy subsets of the unit interval. In this representation the membership grade of any proportion $r \in [0, 1]$, $Q(r)$, is a measure of the compatibility of the proportion r with the linguistic quantifier we are representing by the fuzzy subset *Q*. For example if *Q* is the quantifier *most* then $Q(0.9)$ represents the degree to which 0.9 satisfies the concept most.

Having discussed the concepts of summarizer and quantity in agreement we are now in a position to describe the methodology used to calculate the validity *T* of a linguistic summary. Assume $D = [a_1, a_2, \ldots, a_n]$ is the collection of values that appear in the database for the attribute *V*. Consider the linguistic summary "*Q* items in the database have values for *V* that are *S*."

The basic or default procedure for obtaining the validity *T* of this summary in the face of the data is as follows:

(1) For each $a_i \in D$, calculate $S(a_i)$, the degree to which a_i satisfies the summarizer *S*

(2) Let $r = \frac{1}{n}\sum_{i=1}^{n} S(a_i)$ be the proportion of D that satisfy S

(3) $T = Q(r)$, the grade of membership of r in the proposed quantity in agreement.

As an example, assume we have a database consisting of 10 entries. Let D be the collection of ages associated with these entries: D = [30, 25, 47, 33, 29, 50, 28, 52, 19, 21]. Then:

(1) Consider the linguistic summary "*most* people are *at least 25*". In this case the summarizer *at least 25* can be expressed simply as $S(x) = 0$ if $x < 25$ and $S(x) = 1$ and $x \geq 25$. We define *most* by the fuzzy subset $Q(r) = 0$ if $r < 0.5$ and $Q(r) = (2r - 1)^{1/2}$ if $r \geq 0.5$. The first eight items in the data collection $S(x) = 1$ while for the remaining two items $S(x) = 0$. Thus in this case $r = 8/10$ and hence $T = Q(8/10) = \left(\frac{16}{10} - \frac{10}{10}\right)^{1/2} = \left(\frac{6}{10}\right)^{1/2} = 0.77$.

(2) Consider the proposition "*about half* the ages are *near 30*". We define *near 30* by:

$$S(x) = \exp^{-\left(\frac{x-30}{25}\right)^2}$$

and we define *about half* as:

$$Q(r) = \exp^{-\left(\frac{r-0.5}{0.25}\right)^2}$$

In this case $r = 3.94/10 = 0.394$ and $Q(0.394) = 0.95$.

(3) Consider the proposition "*most* of the people are *young*". We define *young* as:

$$S(x) = 1 \qquad \text{if } x < 20$$
$$S(x) = -\frac{1}{10}x + 3 \qquad \text{if } 20 \leq x \leq 30$$
$$S(x) = 0 \qquad \text{if } x > 30$$

In this case $r = 0.27$ and using our previous definition of most we get $T = Q(r) = 0$.

Thus far we have considered linguistic summaries involving only one attribute. The approach described above can be extended to the case of multiple attributes from a database. We first consider summaries of this form "*Most* people in the database are *tall* and *young*." Assume U and V are two attributes appearing in the database. Let R and S be concepts associated with each of these attributes respectively the generic form of the above linguistic summary is:

"Q people in the database have U is R and V is S."

In this case our data set D consists of the collection of pairs, $D = [(a_1, b_1), (a_2, b_2), ..., (a_n, b_n)]$ where a_i is the value of the V attribute and b_i is the value of the U attribute for the i^{th} object in the

database. Our procedure for evaluating the validity of the linguistic summary in this case is defined as follows:

(1) For each i calculate $R(a_i)$ and $S(b_i)$.

(2) Let $r = \frac{1}{n}\sum_{i=1}^{n}\left(R(a_i)S(a_i)\right)$

(3) $T = Q(r)$

We now consider another class of linguistic summaries manifested by statements like:

"*Most tall* people in the database are *young*."

In this case we have as our generic form:

"*Q U* is *R* objects in the database have *V* is *S*."

In the above we call *R* the qualifier of the summary. The procedure for calculating the validity of this type of linguistic summary has the same three-step process except that the calculation for *r* is as follows:

$$r = \frac{\sum_{i=1}^{n} R(a_i)S(a_i)}{\sum_{i=1}^{n} R(a_i)}$$

We note that we can naturally extend this procedure to handle summaries of the form "*Few well paid* and *young* people in the data base live in the *suburbs*."

2.2 Temporal Summaries

In our current work we are concerned with the extension of the preceding ideas to cover situations where the behavior of data over time is of primary interest. Such an extension will allow us to characterize time-series data, such as Internet buzz, in ways that lead to an understanding of the significance of the patterns the data exhibits.

In this context, a linguistic summary has the form, "In the last few weeks knowledgeable insiders have become strongly optimistic about the new operating system to be introduced by Apple." or, "Since his announcement in August the media coverage of Arnold Schwarzenegger's candidacy has been friendly." Each of these has a similar structure that involves a time interval over which the summary holds, a source (or set of sources) from which the data is derived, a variable that is the focus of the summary, and a characterization of the behavior.

We capture this idea by defining a temporal linguistic summary (TLS) as a 4-tuple:

TLS := $\{T, S, V, B\}$

where:

Time Extent (*T*): is the time interval over which the summary holds
Source (*S*): is the set of sources from which the data is derived
Variable (*V*): is the subject or concept that the data represents
Behavior (*B*): is the characterization of behavior of the data over *T*

and where each component is an independent facet that we can compute from our knowledge of the underlying time-series data.

So the first example above can be mapped (somewhat loosely) onto:

T: "In the last few weeks"
S: "knowledgeable insiders"
V: "opinion of the new operating system to be introduced by apple"
B: "have become strongly [optimistic]"

and the second example can be mapped onto:

T: "Since his announcement in August"
S: "the media"
V: "coverage of Arnold Schwarzenegger's candidacy"
B: "has been [friendly]"

In both cases note that the behavioral component is defined by a combination of a behavioral pattern and directionality (or polarity) that reflects the underlying sentiment - "optimism" in the first case, and "friendliness" in the second.

Given this framework, we view the summarization task as one of generating the most useful TLS over an underlying database of time-stamped data objects. That is, we treat the task as a data-mining exercise in which we pre-specify the set of elements from which we can construct the extracted information descriptions.

In the remainder of this section we briefly review the underlying concepts for each facet and discuss the issues we face in computing a summary from the data.

2.2.1 Time Extents

In our model, time extents are treated as fuzzy time intervals and often anchored at one end with respect to a date or an event. That is, we think of these intervals as being relative to a reference point that corresponds to a date around which a user is interested in exploring the behavior of the source.

So for example, an extent such as "Several days prior to the policy speech … " uses a specific event (here a particular speech), and hence a specific date, to anchor the interval and then extends it backwards in time to some fuzzy end-point. Similarly, an extent such as "In the early part of the year … " refers to an interval that begins on a specific date (here January 1st) and extend forwards in time.

Fuzzy time intervals are a well-understood mathematical concept (Dubois and Prade, 1989; Yager, 1997) so the challenge for our approach is to define the formal equivalents of descriptions like "several days" and "early part of the year." For now we are relying on a pre-defined set of parameterized intervals that correspond to prototypical extents.

2.2.2 Data Sources

In our approach, a data source is a general designator for any collection of documents that, when taken together, provide the basis for the kinds of behavioral analysis we want to perform. Note that in this definition, we use "document" as the generic term for any information-bearing object that uses written human language. This very general perspective allows us to create arbitrary "sources," although normally we expect that a user would be interested in a specific forum, or author, or publication.

Sources are then characterized using a set of meta-data that encodes information such as the type, the language, geo-location, as well as information that is specific to a type (e.g., whether a on-line news source is state controlled or independent), and also information about the reliability and authoritativeness of the source content.

The meta-data are then used to create sets of taxonomic organizations that we think of as hierarchical dimensions and that form a framework within which we can do the kinds of data analysis we envision. So, for example, we might organize on-line news sources according to their geo-location as shown in Figure 1 below.

This hierarchy provides a natural set of aggregates that we can use to perform "roll-up" and "drill-down" operations as we mine the underlying data for interesting behaviors. That is, we use source organizations of this kind to control the focus of our analysis. This in turn allows us to detect that, say, the behavior that has the highest informativeness emerges at the level of "Mid-East Sources" rather than at the level of a specific publication such as the Tehran Times.

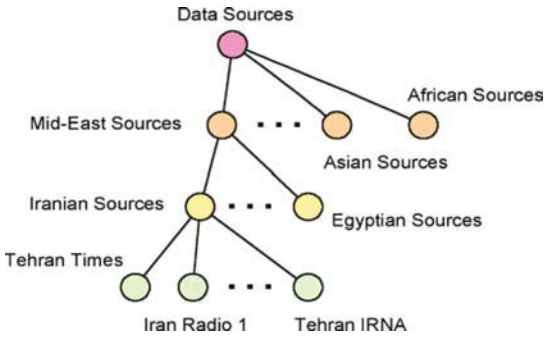

Figure 1. Taxonomy of Media Sources by Geo-Location.

2.2.3 Variables

The third element in our model is the concept of variable. By variable we mean any topic or subject that is mentioned in the documents we are analyzing. What constitutes an interesting set of

topics is highly domain and application specific, but can range from such things as the particular characteristics of a consumer product to high-level political and economic concerns.

Whatever the domain, we organize the topics hierarchically to facilitate our data analysis. So, for example, Figure 2 shows a fragment of a current affairs taxonomy.

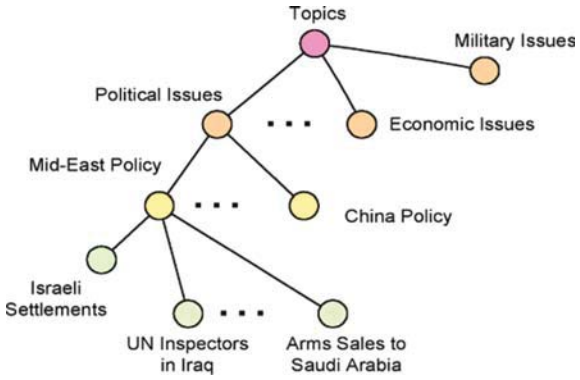

Figure 2. Taxonomy of Current Affairs Topics.

Other typical subject organizations would be by person or organization or event types, and multiple such hierarchies are likely to be in use within a single application context.

Determining that a particular document is "about" a particular subject is, of course, a significant challenge that researchers in the fields of Information Retrieval and Natural Language Processing have been investigating for many years. Our own techniques are evidentially based, and allow us to assert that a document is about a particular topic with some degree of confidence. Since in this paper our focus is on summaries of behaviors, we will not elaborate further, except to note that any technique that can make assertions of this form could be used to provide inputs to our summarization algorithms.

Once we have determined the topicality of a text, then in the general case we are also interested in the attitudes and opinions that are being expressed about the topic. Detecting and classifying opinions is also a hard problem, and we use a mix pattern-based and syntactic analysis tools within our evidential framework to extend our basic assertion model to include qualified statements about polarity. We make no special theoretical claims for these techniques since they are similar in spirit to those used by other groups, but in our application efforts we have focused on the questions of efficiency and cost-performance tradeoffs.

2.2.4 Behaviors

The final element of our TLS model is the idea of a behavior. This is the key element in the summary since it carries the information that is critical to the user in both understanding the historical characteristics of a source, and in alerting them to significant current changes.

Since each document is represented as a time-stamped data object with information about the source, the topical content, and polarity, our behavior detection algorithms work by attempting to

match time-series segments to one of a number of canonical behavior classes. This is an active area of investigation for us, but currently we have classes that describe trends (e.g., up, down), level changes (e.g., higher than, same, lower than), and simple dynamic patterns (e.g., spike, oscillation). We use modified landmark detection techniques to identify inflection points in the time-series (Dunham, 2003) and fuzzy pattern matching techniques (Dubois et al., 1988) to map time-series sub-segments to the behavior class descriptors.

The output of this step is a candidate behavior descriptor that we then combine with the other elements of the TLS to produce the final formal summary. Creating the actual English language gloss is done using straightforward template-based techniques.

3. Example Applications

In this section we present two illustrative excerpts from a variety of projects in which we have explored techniques that analyze and then characterize buzz and sentiment in online sources. The first example is taken from a pilot project to track attitudes and opinions towards US foreign policy as expressed in state-controlled media. The second example is taken from a longer-term project in which the objective was to mine online discussions of movies for leading indicators of a movie's financial performance.

3.1 State Controlled Media

The State Controlled Media Project (SCMP) looked at the technical and operational challenges of using state-controlled media as a source of attitudes and opinions towards the US in general and towards US foreign policy initiatives specifically. The source data were taken from the Foreign Broadcast Information Service (FBIS) and organized based on geographic location. The topics of interest were taken from a predefined set of foreign policy themes and issues similar to the one shown in Figure 2 above.

To illustrate the processing concept, the following is a text taken from the Islamic Republic News Agency (IRNA), the official news agency of Iran:

```
<FBIS_DOC>
<ID>HZQ23000110</ID
<DTG>23 May 02 1810 GMT</DTG>
<SOURCE> Tehran IRNA</SOURCE>
<TEXT>
Tehran, May 23, IRNA -- Iranian Foreign Minister Kamal Kharrazi here Thursday
described the ongoing situation in the Middle East &quot;convulsive and
critical&quot; and blamed &quot;repeated US mistakes&quot; for that.

Talking to IRNA, he said that the US policy line in the aftermath of the
September 11 attack on American landmarks was centered on meeting the
Zionists&apos; interests.

&quot;The US decision-making and its foreign policy line is based on  meeting
illegitimate goals of the Zionists and this should not be allowed to lead to
instability and collapse of the international community,&quot; Kharrazi said.

&quot;The new American policy has adopted use of pressure as a means to carry
other countries along its side, which has led to the spread of spite, hatred
and war,&quot; Kharrazi further said.
</TEXT>
</FBIS_DOC>
```

Using a set of topic classifiers and sentiment analysis tools, each message text is represented as an XML data object that records the assessment of the sentiment of the text with respect to the issues of interest. In this case we get something like:

```
<MSG_METADATA docid="HZQ23000110">
<DTG>23 May 02 1810 GMT</DTG>
<SOURCE>Tehran IRNA</SOURCE>
<ISSUE>
  <ISSUE_NAME>US Mid-East Policy</ISSUE_NAME>
  <ISSUE_SENTIMENT>-0.80</ISSUE_SENTIMENT>
</ISSUE>
</MSG_METADATA>
```

where the value of -0.80 in the <ISSUE_SENTIMENT> slot indicates that there is a significant degree of negative sentiment in this text. Sentiment values range over the interval [-1, +1], and, typically, a text would get multiple <ISSUE> tags.

Over a period of time, the set of such data objects allows us to create a time series that represents the changing state of attitudes in a specific source or an aggregation of sources with respect to a topic of interest.

Figure 3 shows an example time-series created by analyzing a set of Mid-East sources and aggregating all of the negative <ISSUE_SENTIMENT> scores.

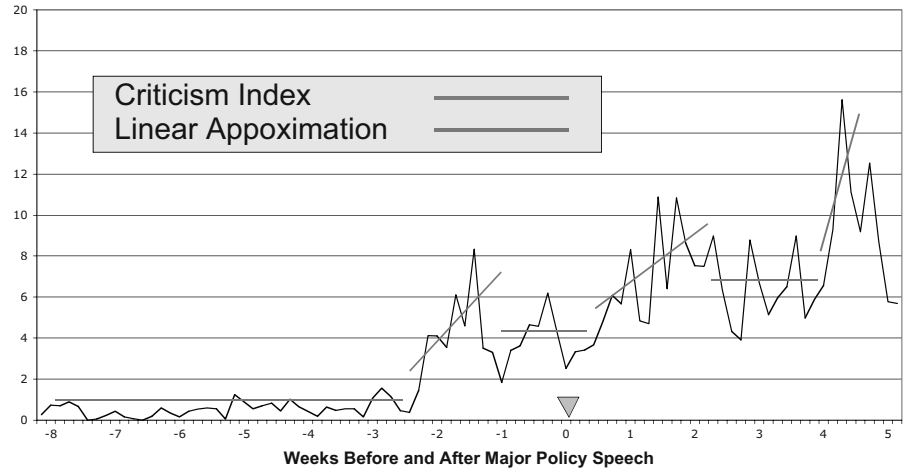

Figure 3. Mid-East Sources via FBIS (April 2002 through July 2002).

This time-series segment spans April 2002 through July 2002 and is centered on a date in late June 2002 when President Bush gave a speech on the need for new leadership in Palestine. The x-axis is labeled with major time divisions at weekly intervals and minor time-divisions at daily intervals. The y-axis is the aggregate amount of negative sentiment (i.e., the sum of the negative <ISSUE_SENTIMENT> scores) on a daily basis across all sources.

The piecewise linear approximation segments show (simplified for presentation purposes) the main sub-segments detected by our algorithms. In this example, the sub-segments get mapped onto either a trend or level-change behavior, and support generation of TLSs like, “In the weeks immediately preceding the speech, criticism reached new levels” and, “Since the speech, criticism has continued to rise.”

3.2 Entertainment Forums

The second example is taken from a large-scale effort to analyze on-line discussion forums about movies to determine if patterns of conversations, and especially the sentiments expressed, can be used as indicators of such things as the response to advertising, or the likely box office receipts.

An edited example posting taken from the Usenet group rec.art.movies.current-films is shown below:

```
<ARTICLE>
<DOCID>tgn-2972</DOCID>
<BOARD>rec.arts.movies.current-films</BOARD>
<SUBJECT>Crouching Tiger, Hidden Dragon</SUBJECT>
<BODY> <TEXT>
I must say that the film blew me away. It&apos;s the best martial arts film
I&apos;ve ever seen, and probably the best action film I&apos;ve seen. The
direction by Lee makes all the difference. His camera moves with each blow.
There is an amazing fludity in the film that, comibined with the scenery,
makes the film easily the most ravishing of the year.
</TEXT> </BODY>
</ARTICLE>
```

As with the state-controlled media example, we process texts of this type to generate sets of structured representations, which include the sentiment about specific aspects of the movie (e.g., the camerawork, the acting, the production, etc.), and then perform our analysis at various levels of granularity.

Figure 4 shows the aggregated positive and negative sentiment for the movie “Crouching Tiger, Hidden Dragon” which was released on December 22, 2000. Positive sentiment is shown as a continuous curve with positive value. Negative sentiment is shown as a continuous curve with negative values. Overlaid on this are the figures for the weekly box office receipts in millions of US dollars (shown as columns).

As with Figure 3, the x-axis is labeled at weekly and daily intervals, but in this case the left y-axis measures aggregated daily sentiment in both directions.

We do not have the space, or the permissions, to discuss the details of our analysis of these kinds of time-series, but in general we have not found any significant correlation between sentiment and absolute box office receipts. We have found, however, that certain characteristics of the basic buzz time series allow us to make good predictions of what the movie industry calls the “longevity” of a movie (usually defined as the ratio of the total theatre receipts to the opening weekend receipts). We also found many cases in which national advertising campaigns had detectable effects on both the volume and content of the on-line conversations.

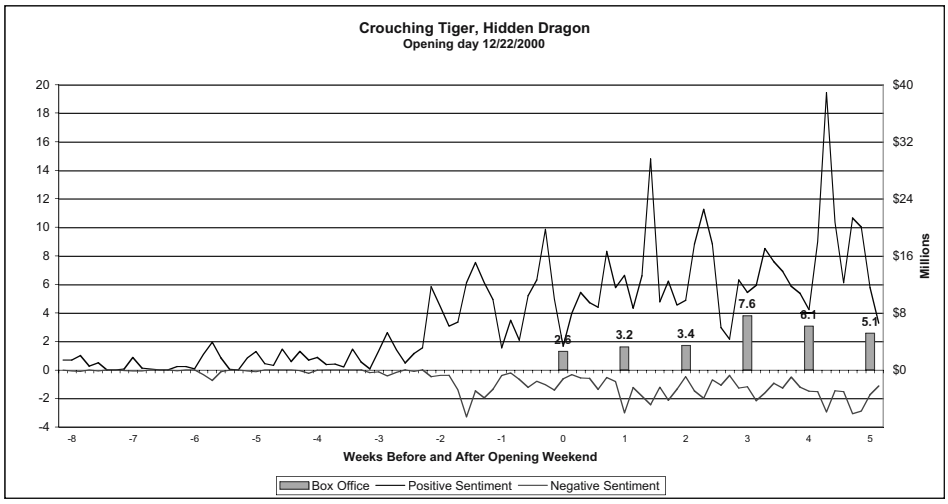

Figure 4. Usenet movie groups (October 2000 through February 2001).

4. TRENDS-2™ Infrastructure

In the work reported in this paper, we made use of Tarragon's TRENDS-2™ content acquisition system shown in Figure 5 below.

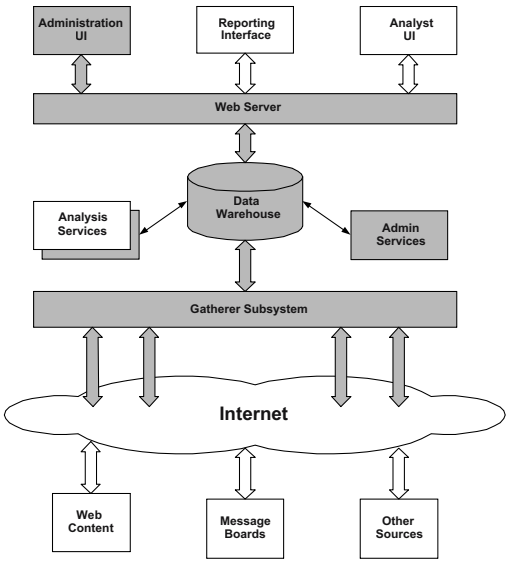

Figure 5. Basic TRENDS-2™ Architecture

TRENDS-2™ is a totally automated approach to Internet data gathering and analysis that combines smart web crawlers, data warehousing technology, fully configurable data processing and analysis workflows, and interactive report generation. In the standard configuration (shown shaded in Figure 5) the TRENDS-2™ infrastructure includes a set of standard "Analysis Services," namely: a generic query language and search capability; families of meta-data extractors and taggers; content analyzers that use fuzzy regular expression techniques; document deduplication algorithms; and, time series analysis tools. Specialized modules, such as advanced sentiment detection algorithms, can be inserted into the processing architecture as needed to meet specific application requirements.

5. Previous and Related Work

Our basic techniques for detecting and tracking opinions, as first reported by Tong (2001), rely on the use of custom lexicons that capture what we call speaker emotion, topic features, and language tone. The lexicons are developed using a mixture of handcrafted and automatic, corpus-derived patterns, and then used in various scoring algorithms to generate "sentiment scores" of the kinds we illustrate in Section 3. So in the case of the movie analysis, we model the way posters feel about the movie (e.g., "I really liked this movie", "I hated it"), the subjective comments they make on various aspects of the movie (e.g., "great acting", "wonderful visuals", "terrible score", "uneven editing"), and the overall use of vocabulary that indicates the general tone of their comments (e.g., "exhilarating", "go see it", "it sucks", "entertaining", "waste of money", "highly recommended").

Our original approach was influenced by the work of Wilson and Rayson (1993) on the analysis of transcripts of market research interviews, and by the work of Subasic and Huettner (2000) on the development and use of affect lexicons. This work also drew on earlier ideas by Hearst (1992) for looking at directionality in text, on the tools for generating lexicons developed by Spertus (1997) in her system for recognizing hostile messages, and on the methods used by Sack (2000) to characterize discussion themes in Usenet newsgroups. More recently, we have been encouraged by the work of Turney (Turney, 2002; Turney and Littman, 2003) and Wiebe (Wiebe, 2000; Wiebe et al., 2003) to explore more robust attitude and opinion detection techniques.

Our work on linguistic summaries is based on the original work by Yager (1991; 1996), together with the advances and applications of this work made by Kacprzyk, Yager and Zadrozny (2001). We drew from the extensive literature on time-series modeling to create our time-series segmentation and approximation algorithms.

6. Open R&D and Application Issues

There are a significant number of challenges that need to be addressed if we are to make effective, large-scale use of on-line, open-source material. In this final section, we review the key technical challenges and application concerns that we face in our own work.

The first technical issue is what we call "source characterization." In order to effectively aggregate data from multiple sources, we need a methodology for characterizing a source along a number of dimensions, including properties like reliability, coverage, bias and timeliness. These are all difficult concepts in their own right, but do need to be factored into the kinds of large-scale analysis we envision.

The second technical issue is sentiment detection and classification in text. The techniques we have used to date have worked well on sources in the consumer product spaces (e.g., movies, automobiles, and personal healthcare), but begin to lose their effectiveness when we look at sources that address current events and politics. Some of the reasons for this are the highly informal nature of language usage in politically oriented forums coupled with the often chaotic nature of the discourse, and the use of standard language constructs in state-controlled media that need to be calibrated to ascertain their real attitudinal value.

The third technical issue is how develop more effective techniques for generating linguistic summaries. These are technical concerns that are specific to our approach and include such questions as how to formally represent the information we discover in the data, how we test the validity of conjectured observations, and the kinds of measures we need to assess of value of the summaries we produce. We are also interested in trying to produce more complex linguistic summaries like, "Most Mid-East sources are critical of US policy on …" or even, "Sources that were united in their views … are now exhibiting a diversity of opinions."

As important as the technical challenges are, there is an equally important set of application issues that need to be addressed if this kind of technology is to be successfully deployed in end-use organizations.

The first application-related issue is the one of scalability. Given our premise that only when we look at large amounts of material do the important patterns emerge, then we need our processing algorithms to be able to work with large volumes of data that have a low "signal to noise" ratio. We also need to be able to handle very diverse sources that encompass multiple genre, languages and contexts, and we need to be able to do much of the processing in near real-time.

The second application issue is the need to develop appropriate metrics, both at the component level as well as at the system level. Although we have adopted an obvious numerical representation for sentiment, it is not clear exactly what it means to talk about the amount and directionality of sentiment, nor is it not obvious that the kinds of sentiment aggregation we do in our current implementation is justified. We also need metrics to help us understand the trade-off between the costs of achieving increased accuracy in text processing and summary generation, and the benefits that the ultimate end-user sees.

The final application issue is the fundamental need to demonstrate that the kinds of text data mining we are proposing do indeed provide significant added value for intelligence analysis and decision-making. In order for Government and Industry to invest substantial resources in this endeavor, it is important to have convincing evidence either that we can provide insights that were either not available before, or that we can provide them sooner and at a lower-cost than using existing techniques.

7. Bibliography

Dubois, D., Prade, H. and Testemale, C. (1988) Weighted fuzzy pattern matching. *Fuzzy Sets and Systems*, 28, 313-331.

Dubois, D. and Prade, H. (1989) Processing fuzzy temporal knowledge. *IEEE Transactions on Systems, Man and Cybernetics*, 19, 729-744.

Dunham, M. (2003) *Data Mining*. Prentice Hall, Upper Saddle River, NJ.

Hearst, M. (1992) Direction-Based Text Interpretation as an Information Access Refinement. In Jacobs, P. (Ed.) *Text Based Intelligent Systems*. Lawrence Erlbaum, Mahwah, NJ.

Kacprzyk, J. and Yager, R. (2001) Linguistic summaries of data using fuzzy logic. *International Journal of General Systems*, 30, 133-154.

Kacprzyk, J., Yager, R. and Zadrozny, S. (2001) Fuzzy linguistic summaries of databases for efficient business data analysis and decision support. In Abramowicz, W. and Zaruda, J. (eds.) *Knowledge Discovery for Business Information Systems.* Kluwer Academic Publishers, Hingham, MA.

Qu, Y., Shanahan, J. and Wiebe, J. (Co-chairs) (2004) *Exploring Attitude and Affect in Text: Theories and Applications.* AAAI Spring Symposium SS-04-07. AAAI Press, Menlo Park, CA.

Rasmussen, D. and Yager, R. (1997) A fuzzy SQL summary language for data discovery. In Dubois, D., Prade, H. and Yager, R. (Eds.) *Fuzzy Information Engineering: A Guided Tour of Applications*. 253-264. John Wiley & Sons, New York, NY.

Sack, W. (2000) Conversation Map: A Content-Based Usenet Newsgroup Browser. In *Proc. ACM International Conference on Intelligent User Interfaces*. New Orleans, LA.

Spertus, E. (1997) Smokey: Automatic Recognition of Hostile Messages. In *Proc. 9th Innovative Applications of Artificial Intelligence*. Providence, RI.

Subasic, P. and Huettner, A. (2000) Affect Analysis of Text Using Fuzzy Semantic Typing. In *Proc. 9th IEEE International Conference on Fuzzy Systems*. San Antonio, TX.

Tong, R. (2001) An Operational System for Detecting and Tracking Opinions in On-Line Discussions. In *ACM SIGIR 2001 Workshop on Operational Text Classification Systems*. New Orleans, LA.

Turney, P. (2002) Thumbs Up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA.

Turney, P. and Littman, M. (2003) Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21 (4), 315-346.

Wiebe, J. (2000) Learning Subjective Adjectives from Corpora. In *Proc. 17th National Conference on Artificial Intelligence*. Austin, TX.

Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E., Wilson, T., Day, D. and Maybury, M. (2003) Recognizing and Organizing Opinions Expressed in the World Press. In *New Directions in Question Answering*. AAAI Spring Symposium SS-03-07. AAAI Press, Menlo Park, CA.

Wilson, A. and Rayson, P. (1993) The Automatic Content Analysis of Spoken Discourse. In Souter, C. and Atwell, E. (Eds.) *Corpus-Based Computational Linguistics*. Rodopi, Amsterdam, The Netherlands.

Yager, R. (1991) On linguistic summaries of data. In Piatetsky-Shapiro, G. and Frawley, B. (Eds.) *Knowledge Discovery in Databases*. 347-363. MIT Press, Cambridge, MA.

Yager, R. (1996) Database discovery using fuzzy sets. *International Journal of Intelligent Systems*, 11, 691-712.

Yager, R. (1997) Fuzzy temporal methods for video multimedia information systems. J*ournal of Advanced Computational Intelligence*, 1, 37-45.

Zadeh, L. (1975) The concept of a linguistic variable and its application to approximate reasoning: Part 1. *Information Sciences*, 8, 199-249.

Zadeh, L. (1999) From computing with numbers to computing with words - From manipulation of measurements to manipulations of perceptions. *IEEE Transactions on Circuits and Systems*, 45, 105-119.