

## **Computer Coding of 1992 ANES Like/Dislike Responses and ANES Responses on Knowledge about Rehnquist**

**David Fan**

I'm going to talk about the InfoTrend software for the computer content analysis of text that I have been developing over a number of years (Fan, 1994, U.S. patent 5,371,673).

As background (Slide 2), I was doing laboratory work in biology on hormones and noticed that these are chemical messages sent in identical copies by one group of cells to influence other cells in the body. I translated that to opinion and behavior in which the news media consist of messages in identical copies sent by one group of people, the publishers, to influence people in the general population.

The same mathematical modeling could be used to predict both hormone action and changes in public opinion in response to persuasive information. The hypothesis was that the mass media contained the major information used by the public for opinion change. Therefore, press coverage was scored for its expression of ideas likely to alter opinions. Then the scores were used to predict opinion time trends.

In an example (Slide 3), the scoring was performed on 45,145 Washington Post and Associated Press stories on the economy from 1977-2000. The scoring was performed using the InfoTrend software to identify paragraphs favorable and unfavorable for the economy. The media scores were entered into the differential equation model of ideodynamics to predict monthly values for the Index of Consumer Sentiment from the University Michigan constructed from 5 survey questions about the economy (Fan and Cook, 2003, *Journal of Mathematical Sociology*, 27,1-23).

The Index data (Slide 4) are shown with the height of each bar giving the confidence assigned by the University of Michigan. The two lines overlaying the Index values give ideodynamic predictions starting with different initial values. The lines rapidly converge and show the robustness of the predictions.

Scores from InfoTrend content analyses have been consistently successful in the ideodynamic model for predicting opinion time trends with examples in the political domain including presidential approval (e.g. Watts, et al., 1999, *Journal of Politics*, 61, 914-952; Shah, et al., 2002, *Public Opinion Quarterly*, 66, 339-370; Fan, 2008, in Scheuren and Alvey, "Elections and Exit Polling, Wiley-Blackwell), and pre-election polls (e.g. Fan, 1988, *Predictions of public opinion from the mass media: Computer content analysis and mathematical modeling*, Greenwood Press, Westport, CT; Fan, 1996, *Political Analysis*, 6, 67-105; Domke, et al., 1997, *Journalism and Mass Communication Quarterly*, 74, 718-737; Watts, et al., 1999, *Communication Research*, 26, 144-175).

A major theme of this conference has been the desirability of consistency and transparency in the scoring of open-ended responses. Clearly, machines can evaluate consistently.

However, there can be two meanings for transparency. One is that nothing about the scoring procedure is hidden from view. In that sense, any machine method is automatically transparent because its decisions can be deduced from the software.

However, transparency also has another meaning, namely comprehensibility of the scoring algorithm by humans. In this sense, an algorithm is opaque to the extent that it is difficult for a human to grasp the scoring criteria.

For the use of scores from open-ended survey questions, comprehensibility is essential because human researchers must understand the meanings of the scores.

A notable example here is an ANES algorithm for scoring open-ended responses for knowledge about William Rehnquist (see Slide 17 below). This algorithm was written for implementation by human coders and was thus both explicit and comprehensible. Human reading could quickly show how the algorithm was so problematic as to be an impetus for organizing this conference.

Without this clear documentation, the scoring criteria would likely have been so unclear that the reasons for puzzling scores might not have been understood. Thus the Rehnquist example illustrates why transparency of the algorithm in the sense of human comprehensibility is so important.

The InfoTrend method has the advantage of implementing algorithms that are easily comprehended by both people and machines. In fact, InfoTrend instructions for an expanded version of the Rehnquist algorithm are given in Slide 18.

Slide 5 provides a general overview of the InfoTrend system for analyzing text.

Slide 6 covers the steps in the InfoTrend method starting with the entry of the input text. For an InfoTrend study, the unit of analysis is usually a small block of text, a paragraph or a sentence, because text within such a block is usually on the same topic thereby improving scoring precision.

Slide 7 outlines an analysis of some ANES open-ended responses provided for this conference. The responses were to a question asking about a respondent's "like" and "dislike" of Bill Clinton in the 1992 presidential election. The responses were quite varied and could be scored for their expression of a wide range of ideas.

Every paragraph in the responses was scored for the presence of each of the chosen ideas. A new paragraph began whenever the interviewer asked a prompting question. Since more than one paragraph could be scored to express a given idea, a sum of the paragraph counts gave an indication of the respondent's interest in the idea.

The InfoTrend method scores by machine so there is no penalty for rescoring the text. Therefore, it is both feasible and desirable to develop scoring instructions for one limited group of related ideas at a time. The domain restriction is useful for minimizing distractions and hence avoiding mistakes during the development of scoring categories.

To illustrate the InfoTrend method, instructions were developed to score the ANES like/dislike responses for references to the ideologies of being liberal or conservative. This topic was arbitrary, and any other could have been chosen.

Once ideology was chosen, a sample of text enriched for this issue was read for assigning the detailed ideas or categories to be scored. The reading led to the surprising finding that the responses referred not only to the ideology of Bill Clinton but also the political leanings of the respondent. Therefore, the text was scored for the ideologies both Clinton and the respondent. The decision to expand the scored ideas to include the respondents' ideologies is typical in text analysis where it is often desirable to modify categories formulated in advance.

For designing the scoring algorithm, it is advantageous to begin by selecting text with a high density of discussion of the chosen topic. That way, uncommon but important ideas can be surfaced. This gives a more sensitive analysis than merely basing scoring instructions on a representative sample of the entire corpus.

Slide 8 describes the InfAlign method for visually examining text enriched for the scoring topic (Fan and Fan, 2009, U.S. Patent 7,519,521). This procedure differs from the usual method of performing a key word search on a set of documents and receiving as output a few lines of text from each document. Only a small number of such text fragments can be presented on a page.

In contrast, the InfAlign map shows the concurrent presence of several ideas belonging to a document. Each document is represented by a series of stacked bar graphs that can be made narrow enough that information about hundreds of documents can be presented simultaneously on a single page.

Slide 9 presents an infAlign map for all 83 documents (3 percent) containing at least one the words “conservative” or “liberal” among the 2665 responses analyzed. This low percentage shows the advantage of filtering the documents to focus on a topic of interest. The development of ideological scoring categories could concentrate on just 83 relevant responses without the need to look at text on other issues. The other topics could be scored in separate analyses.

All responses responding to the “like” question for Clinton were given the arbitrary date of 01/03/1992 and all “dislike” responses were given the date of 01/04/1992. That way, the date could be used to sort and group responses by the “like” and “dislike” questions. In the map of Slide 9, the top row gives the date with all “like” responses on the left in the shaded band spanning the height of the map. All “dislike” responses were on the right and not shaded. In the unshaded region on the right hand side of the top row, each response was indicated by a full-height blue bar indicating the date of 01/04/1992, and in the shaded regions all responses had no blue bar corresponding to date 01/03/1992. Thus the blue bars were vertical bar graphs giving the date value extending over the full date range from 01/03/1992 to 01/04/1992. A bar was drawn for every response so fewer bars on the left meant that fewer “like” response used one of the ideological terms.

In general, bar graphs space their bars so that a legend can be written beneath each bar. The InfAlign map removes all labels so that the bars can be compressed in width. Instead, the legend information is provided in the horizontal region at the top of the map. That information corresponds to the response highlighted by red cursor. This cursor can be moved horizontally across the map using a computer mouse or arrow keys. In this way, a user can choose any response using the cursor and see the legend information for that response at the top of the map.

In addition to the top date row, the InfAlign map also can have additional rows for other variables associated with the responses. For each response, Rows B and E give the count of the numbers paragraphs containing the words “conservative” and “liberal” respectively, in the highlighted response. Other rows give the numbers of paragraphs scored using the InfoTrend method (see Slides 11-16 below) containing the ideas of Clinton being conservative (Row C), the respondent as being conservative (Row D), Clinton being liberal (Row F), and the respondent being liberal (Row G). Row H reports the total number of paragraphs per response.

The user can see the values for each of the attributes of a response by the heights of the blue bars under a red bar and by the numbers at the top of the map. For the highlighted response in the map, there was one conservative word (Row B) but Clinton was scored as being liberal (Row E).

To see why, the user can click on the red bar or hit enter and a window opens (Slide 10) with the text of the response. In the text, the word “liberal” appears in bold type and upper case. The user can see that Clinton (through use of the pronoun “he”) is described as not being liberal. The following slides show how the words “he”, “not” and “liberal” are combined to give the final meaning of Clinton being conservative.

In the InfoTrend content analysis instructions, the user defines variable names representing ideas on the left margin of lines in the InfoTrend instructions. Beneath each idea definition line are text strings in {curly brackets} associated with that idea (Slide 11). Some ideas like “ClinConserv” referring to Clinton being conservative have no words associated with them. These ideas are evaluated using user written rules (Slide 12).

The rule specifying the negation condition is shown by the red arrow to Rule 6 in Slide 13. Rule lines are numbered on the right. The arrow points from the table at the top of the slide. The table shows how the rule follows logic used by human coders to specify how a word with the FwdNeg idea changes the idea of Liberal to the idea of Conservative.

Then Rule 11 (Slide 14) shows how “he” is linked to the idea of conservative to give the idea that Clinton is conservative. A reading of the like/dislike responses showed that “he” rarely meant anything but Clinton.

Thus the idea that Clinton is conservative is obtained in two steps using Rule 6 followed by Rule 11. The sequential action of both rules is shown in Slide 16.

Slide 15 indicates that the InfoTrend instructions written by the user are applied to sample text chosen from an InfAlign map to be particularly relevant to the coding topic. Then, the results are shown in diagrams like that in Slide 16 to help guide the user in making changes to the instructions to improve the scoring. The changed instructions are then retested on samples of text in an iterative fashion.

Clearly, a paragraph could be scored for a broad spectrum of ideas. InfoTrend instructions can be written to extract more nuances from the responses. For example, a paragraph mentioning Clinton as liberal could also contain a reference to a particular policy. The joint occurrence of both a policy and an ideology usually meant that the policy was a reason the ideology. Instructions can be written for that linkage to require more information than just the joint occurrence in a paragraph.

Another example for the use of the InfoTrend method is to implement the ANES algorithm (Slide 17) for scoring knowledge about William Rehnquist.

Slide 18 shows that these instructions are easy to implement using the InfoTrend system. In fact, the instructions in Slide 18 add synonyms for the concepts of Chief (e.g. head) and Justice (e.g. honcho) so that the instructions could capture variant ways of expressing the idea of Chief Justice. The instructions also included the specification that Chief Justice could apply to not only the Supreme Court but also to the United States, a part of the official title, to give the desired specification of Rehnquist's job.

Slide 19 considers some additional uses for the InfoTrend scoring method in areas such as crisis management.

Obviously, the InfAlign map can display values for open as well as closed ended questions. Using such a mixed format, it is possible to scan quickly and simultaneously for consistency in a large number of responses. Thus an open ended statement that a respondent was an engineer can be mapped in one row with the education of the respondent in another. Rapid inspection of the map would show whether education was at least at the college level and hence how the respondent used the term engineer.