Using Term Frequency to Identify Trends in the Media’s Coverage of Health

Delano J. McFarlane, M.Eng., PhD Candidate, Rita Kukafka, DrPH, MA
Department of Biomedical Informatics Columbia University, New York, NY

Abstract
This poster describes a method of analyzing news reports to detect emerging trends in the media’s coverage of health. The method examines term frequency and term usage in overall and health-specific news coverage. Term frequency calculation and analysis algorithms have been implemented in SalientNews, a news aggregation and analysis system. By using term frequency, SalientNews is now able to assist in the identification and analysis of emerging trends in the media’s coverage of health.

Problem
Media coverage of health topics and health policy debates in the US and world-wide can have a significant impact on how the public perceives health issues. Identifying trends in the media’s coverage of health can lead to the creation of more effective health communications and interventions which can in turn lead to better health outcomes. Unfortunately the methods used to identify and measure trends in news coverage are often limited, labor-intensive, and time consuming. Better, automated methods, which can identify and measure potential trends in health news coverage are needed.

Solution
Term frequency can be used as the basis for identifying emerging trends in health news reporting. Term frequency has been used extensively in information retrieval and natural language processing methods to index documents and to characterize corpora [1, 2]. In the method presented here, term frequency is used as a proxy for the frequency with which concepts are used in news coverage.

To use term frequency in this way, news reports must be collected and term frequencies calculated on a regular basis. Terms can then be ranked based on how often they appear in news reports over time. Trend analysis can then be performed to identify trends in term frequency. As trends in term frequency are observed, researchers and others interested in health communications can look more closely at the terms associated with these trends.

Current Work
We have built an online news aggregation and analysis system to study health news coverage. The system, SalientNews, has been collecting and analyzing general and health-specific news coverage since October 2006. As of March 2007 SalientNews is retrieving news stories from over 18 online news sources included CNN, BBC, Reuters and NPR. To date over 100,000 news articles have been collected.

Figure 1. Term usage over time from late 2006
To test the use of term frequency to identify trends in the media’s coverage of health, term frequency calculation functions were built into the SalienceNews system. Figure 1 shows the output of an analysis of term use in health news during the later fall months of 2006. In this graph terms related to abortion, HIV/AIDS and E.Coli peaked in their frequency of use. The dates associated with these increases in term use are approximately Election Day in the U.S., world HIV/AIDS day and an outbreak of E. Coli in a fast food chain in the northeastern U.S. By using term frequency SalienceNews can provide users with timely information on emerging trends in health news coverage.

References