



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



D. McFarlane MEng, R. Kukafka, DrPH, MA, Department of Biomedical Informatics, Columbia University, New York, NY

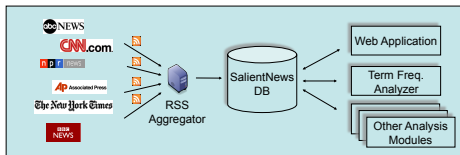


Objective

Determine if term frequency can be used to characterize the media's coverage of health news.

Methods

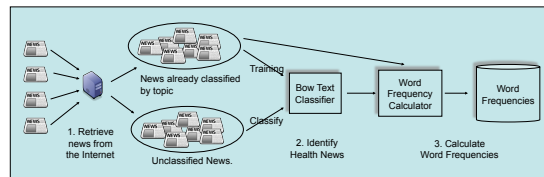
SalientNews, a news analysis system that we built, was used to collect and analyze over 10000 news articles published online between October 2006 and March 2007. News sources included Associated Press, NPR, ABCNews, BBC, New York Times and CNN. Bow¹, a statistical text classification program, was used to identify health related news articles. Frequencies were calculated for terms appearing in article titles and descriptions. A stoplist was used to filter out very common words (e.g. *a, an, it, the*). An implementation of the Porter Stemming Algorithm² was then used to combine word variations. Trend and usage analysis was conducted for terms appearing in the top 20 most frequent for any week of analysis.



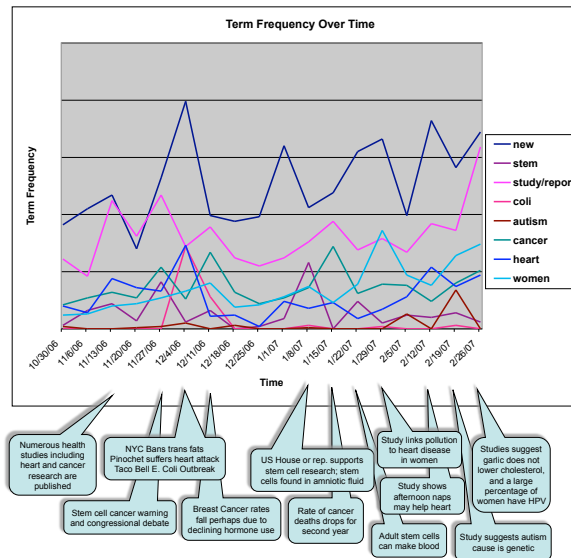
SalientNews Architecture

SalientNews

SalientNews is a news aggregation and analysis system that we are developing to evaluate news analysis methods and to analyze the media's coverage of health topics. Its major components are an RSS (Really Simple Syndication) news aggregator, a Web based user interface, analysis modules and a database containing news articles, configuration information and analysis results. At present, news articles are retrieved from over 25 regional, national and international news sources. Standard Internet technologies were used in its development (e.g. RSS, MySQL, PHP, Java and Ruby on Rails).



Methods Schematic



Results

132 terms met the criteria for analysis. Some words (e.g. *new, child, report*) had consistently high frequencies (over 10 weeks in top 20) and appeared in various different health news stories. Other terms (e.g. *coli, autism*) exhibited more dramatic changes in frequency (under 3 weeks in top 20) and were typically associated with specific health related events and debates.

	new	stem	study/report	coli	autism	cancer	heart	women
29-Oct-02	91	3	61	0	2	21	20	12
5-Nov-02	105	16	48	0	0	27	14	15
12-Nov-02	117	22	112	0	0	32	44	23
19-Nov-02	70	7	81	0	1	27	36	22
26-Nov-02	132	41	117	0	2	54	33	27
3-Dec-02	199	0	72	72	5	29	73	33
10-Dec-02	99	16	66	26	0	67	11	40
17-Dec-02	94	0	62	0	3	32	12	18
24-Dec-02	98	2	55	0	0	22	2	21
31-Dec-02	160	3	62	0	0	27	24	28
7-Jan-03	106	56	76	3	1	36	18	37
14-Jan-03	119	0	94	0	0	72	23	23
21-Jan-03	155	24	69	0	0	31	9	35
28-Jan-03	166	5	79	2	0	39	17	36
4-Feb-03	99	12	67	0	13	38	28	47
11-Feb-03	182	10	92	0	0	24	54	38
18-Feb-03	141	14	89	3	34	49	37	64
25-Feb-03	172	6	159	0	0	51	47	74

Selected High Frequency Terms

Conclusions

Term frequency can be used during automated news analysis to characterize trends in the media's coverage of health. Based on preliminary analysis, some terms exhibited consistently high frequencies and were associated with a number of different health news stories over the entire period of analysis (e.g. *new, study, report*). Specific health related events resulted in dramatic fluctuations in the frequency of use of terms related directly or indirectly with those events (e.g. *autism, stem, women*).

Future Work

- Explore how automated news analysis methods can impact research into the media's coverage of health.
- Associate health related words in news stories with concepts in health terminologies (e.g. MeSH, UMLS).
- Explore how health news coverage impacts and can be used in the development of consumer health terminologies.
- Compare the health topics highlighted in health news coverage to expert health knowledge, public health priorities and trends in health journal publications.

Impact

This research will improve the way that the media's coverage of health is analyzed. By accurately characterizing the media's coverage of health, deficiencies in health reporting can be more readily identified. Individuals and organizations involved in health communications can in turn create communications and interventions that address shortcomings in the media's coverage before there is any negative impact on individual or public health.

1. McCallum, Andrew Kachites. "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." <http://www.cs.cmu.edu/~mccallum/bow>. 1996.
2. <http://tartarus.org/~martin/PorterStemmer/>