A Text Mining Approach to Analyze Public Media Science Coverage and Public Interest in Science

Ying Sun

Abstract—**The presentation of science in the mass media is one of the most important questions facing social scientists who analyze science. In this paper, we use topic modeling technique to identify the scientific topic areas or themes most prevalent in mass media over a given period of time to inform the discussion about civic scientific literacy (CSL). Google Trends is used to analyze public interests in science. The two sets of data are compared and correlated to identify any relationship between traditional media and the new media in impacting public perceptions of new scientific developments and public's general understanding of science.**

Index Terms—**Data mining, civic science literacy, public interest in science, mass media.**

I. Introduction

In this paper, we explore the use of a statistical text mining technique called topic modeling, to identify the most important and potentially interesting scientific topics or themes in mass media over a given period of time to inform the discussion about civic scientific literacy (CSL).

Science coverage in the mass media was and still remains the major channel that bridges the gap between science and the general public. Most people acquire their information about science mainly from the mass media. So the presentation of science in the mass media is one of the central questions facing social scientists who analyze science. However, much of what is known about the public's interest in science originates from very general surveys [1], [2]. Science media coverage is also used in operationalization and measurement of CSL. Miller operationalizes CSL as a level of understanding of scientific terms sufficient to read a daily newspaper or magazine and to understand the essence of competing scientific arguments on a given dispute or controversy [3]-[5]. However, the core set of 12 basic science terms used in Miller method does not based on any analysis of media coverage. They may not be representative of a vast and ever expanding domain of science. Understanding mass media science coverage, public science interest and their relationships will help to "enable the citizen to become more aware of science and science-related issues in order to participate in the democratic processes" [6].

The present study intends to answer the following questions:

1) What are the science and technology topics that appear most frequently in the US mass media
2) Is there a correlation between the mass media report of a scientific topic with public interests on the Web about the topic?

The rest of this article is organized as follows. The relevant literature is first reviewed and described in literature review. Topic modeling technique is then introduced next. In experiments and evaluation we discuss the experiment results and evaluations. We then discuss findings and limitations of this study in the Discussion section.

Finally, conclusions and future research suggestions are presented in conclusion and future work.

II. Literature Review

According to National Science Board 2008 survey (2008), the percentage of Americans who say they follow science and technology (S&T) news closely has declined over the past 10 years, more than in other topics covered by the news media [1].

Mass media have traditionally been the main source of information about science for the majority of the American public. Social science searchers who analyze science have been studying the science coverage in mass media for a long time. In Schafer's meta analysis of media science coverage, 215 publications on the topic were identified and analyzed [7]. The study shows that "social scientific attention to media coverage of science has risen continuously over the past decades". About half of these studies are case studies which focus on science coverage of one issue in one region and within a given period of time. Cross-sectional study comparing different countries/disciplines/media counts for 22%. Longitudinal study analyzing the temporal development of coverage over time counts for 18%. The rest of the studies are combination of cross-sectional and longitudinal study.

Content analysis is the major method using by social scientists in identifying scientific topics in mass media. The scientific topic categories are predefined and assigned to articles during content analysis. The categories are usually broad as biology, physics, chemistry, geology, engineering, etc.

For the purpose of developing CSL measurement, the analysis of science coverage in mass media is not to characterize how different areas of science and technology are covered, but to identify more specific scientific topics that are popular in mass media. Given the large size of collection to be analyzed and the large number of potential topics, an automatic method is desirable.

The large amount of information stored in unstructured texts is not easy to analyze and requires specific processing methods to extract useful patterns. Text mining is an analysis

Manuscript received May 19, 2014; revised July 1, 2014.

Ying Sun is with University at Buffalo, USA (email: sun3@buffalo.edu).

DOI: 10.7763/IJMLC.2014.V4.461

process that discovers hidden features and extracts pivotal information for further processing. It uses techniques from information retrieval, information extraction, as well as natural language processing (NLP) and combines them with data mining, machine learning, and statistics. It has been widely used in various applications. In news media analysis, it is primarily used to determine the news category [8], [9].

Brossard and Shanahan used occurrence of scientific terms in the media to build a measuring instrument for CSL [10]. They took a sample from the Oxford Dictionary of Science by randomly selecting a term from every page. They used these terms to search newspapers to find the number of articles each term appeared at least once. The terms with the most articles were included in measurement instrument.. There are two major drawbacks of this approach. First, the initial sets of words were randomly sampled from a dictionary; many important words might be excluded at this initial selection. Second, the relationships among words were totally ignored. Concepts or topics are a better unit to define CSL than an individual word because a concept or topic is more meaningful.

Tseng *et al.* automatically extracted terms and term associations from 901,446 newspaper articles in Taiwan, matched the terms against index terms from Taiwaneses science textbooks, and constructed a concept map of 95 scientific terms to support the development of a measurement instrument for Taiwan CSL [11]. The result shows that the automatically identified concepts are reasonably representative of CSL in Taiwan. Tseng *et al.* used a rule-based extraction algorithm to extract a list of "key" terms from each text. Term Frequency (TF), the number of times the term occurs in a text, was used as the basic extraction criterion. Multi-word phrases are supported. The term boundaries are identified by simply identifying the longest repeated strings in a document. The first *k* terms were treated as topics for the next step of analysis. The mining is limited to pure word and co-word frequency analysis. In other words, the identified terms are isolated from the contexts in which they are used. Thus, this method fails to provide the important contexts in which the identified scientific topics appear.

III. Topic Modeling

Topic modeling is a probabilistic, statistical technique to analyze large text collection to identify patterns in the uses of words to identify (or extract) topics. A topic is constructed as "a cluster of words that frequently occur together" in the same document [12]. Topic modeling techniques assume that any piece of text is composed by selecting words from possible buckets of words where each bucket corresponds to a topic. If that is true, then it becomes possible to mathematically decompose a text into the probable buckets from whence the words first came. The techniques go through this process over and over again until it settles on the most likely distribution of words into buckets, which we call topics. In other words, in topic modeling, a "topic" is a probability distribution over words. So the topics are not predetermined by the researcher but instead emerge from the patterns uncovered by the statistical algorithm. Generally these words are semantically related and interpretable. A topic can often be identified simply by examining the most common words in a topic. Following two examples show topics extracted from our collection:

Topic X: cell human gene DNA scientists stem genome researchers university mice protein cloning body

Topic Y: climate water global carbon warming change ice environmental energy sea emissions oil dioxide

Beyond these groups of words, a topic model also provides proportions for each document a list of topics with importance weights, providing quantitative data that can be used to locate documents on a particular topic.

Topic models have been used on newspaper corpora to discover topics and trends over time. Griffiths and Steyvers identify research topic trends by topic modeling abstracts of scientific papers [13]. Newman and Block analyzed a collection of eighteenth-century American newspaper using topic modeling method [14]. Nelson uses the method to mine topics from the Civil War era newspaper Dispath [15].

There are different topic modeling algorithms. We used Latent Dirichlet Allocation (LDA) implemented in MALLET [16]. LDA is a probabilistic model using a mixture of multinomials. It is not the first topic modeling technique, but by far the most popular one.

IV. Experiments and Evaluation

A. Mass Media Science Coverage

To answer our first research question, we selected *the New York Times, ABC news, CBS news, Fox news* and *NBC news* as our data source. The print version of *the New York Time* is the largest local metropolitan newspaper in the United States and third-largest newspaper overall. It also has a strong presence on the Web. Its website is America's most popular newspaper site, receiving more than 30 million unique visitors per month. Taking into consideration of influence of both its print version and digital version, we believe that it has the largest reader group in America compared with other newspapers and is one of the most influential public media. For science coverage in TV news broadcast, we include news scripts from all four major TV broadcast companies in US. A comprehensive understanding of the scientific content in the collection will give us a reliable, yet valid view of science coverage on public media.

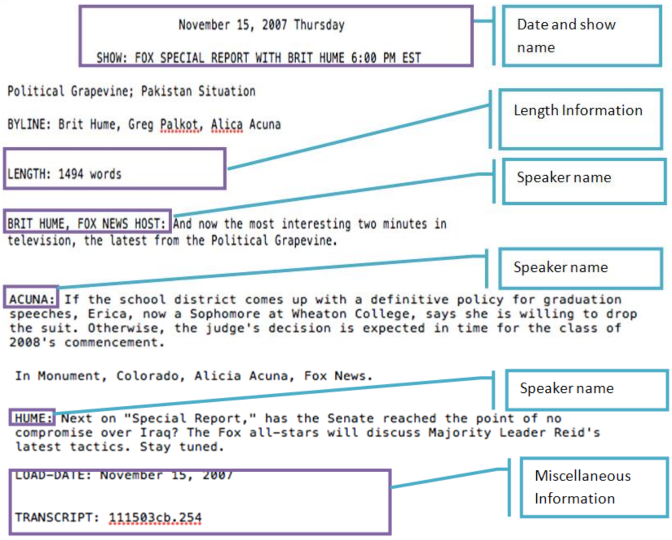
We assembled our corpus by extracting from LexisNexis database about 19K articles from *The New York Times,* and 8K TV news scripts from *ABC news, CBS news, Fox news and NBC news* published form the years 1993 through 2012 that are indexed with the LexisNexis subject index term "*science and technology*".

We removed some duplicates and wedding announcements (which were indexed with s*cience and technology* based on education information for the couple).

To prepare the collection for topic modeling, we removed metadata from each document, including date, byline, article length, host names, transcript information, copyright statement, etc. We also removed punctuation and common words (and, the, or, …) words using the default MALLET stop word list augmented with some corpus-specific common words like “New York Times”. To reduce computing cost, all single-character and two-letter words, and all words that occurred no more than five times in the entire collection are also removed. Given the size of the collection, these terms are rare terms that are not likely to represent common contents across documents in the collection. Fig. 1 illustrates our metadata preprocessing.

After preprocessing, each article is an individual text file that includes the text and some metadata.

We applied topic modeling to the entire 20-year corpus to gain some insight into popular scientific topics in mass media. We set the topic modeling parameter *K* to 15. *K* is the number of topics to extract. It is to some degree affect the granularity of the extracted topic. Generally, the larger the number, the more fine-grained the results are. Given the purpose of our study is to identify relatively general topic areas to develop civic science literacy instruments, we experimented with various *K* values and picked 15 based on the granularity of the topics.



```
November 15, 2007 Thursday

SHOW: FOX SPECIAL REPORT WITH BRIT HUME 6:00 PM EST

Political Grapevine; Pakistan Situation

BYLINE: Brit Hume, Greg Palkot, Alica Acuna

LENGTH: 1494 words

BRIT HUME, FOX NEWS HOST: And now the most interesting two minutes in
television, the latest from the Political Grapevine.

ACUNA: If the school district comes up with a definitive policy for graduation
speeches, Erica, now a Sophomore at Wheaton College, says she is willing to drop
the suit. Otherwise, the judge's decision is expected in time for the class of
2008's commencement.

 In Monument, Colorado, Alicia Acuna, Fox News.

HUME: Next on "Special Report," has the Senate reached the point of no
compromise over Iraq? The Fox all-stars will discuss Majority Leader Reid's
latest tactics. Stay tuned.

LOAD-DATE: November 15, 2007

TRANSCRIPT: 111503cb.254
```

Fig. 1. TV data-formatting example (FOX news 11/15/2007).

TABLE I: MOST COMMON 15 TOPICS IN THE 20-YEAR COLLECTION

Rank	Most likely words in topic	Topic label
3	cell human gene DNA scientists stem genome researchers university mice protein cloning body	**LIFE-HEREDITY**
4	people brain children women study time researchers men behavior social percent found male	**BRAIN**
5	company technology computer research industry business market products information internet	**IT**
7	nuclear weapons government China security Lee war international program military energy	**NUCLEAR**
9	space earth NASA mars mission planet life spacecraft sun miles moon system solar light stars	**SPACE**
10	light physics energy theory particles laboratory experiment research nature matter molecules	**PHYS-MATTER**
11	Cancer disease health people study drug patients virus medical blood risk heart treatment doctors	**MEDICAL**
12	climate water global carbon warming change ice environmental energy sea emissions oil dioxide	**GLOBAL WARMING**
13	food fish species plants forest trees wildlife eat animals water wild conservation agriculture	**SUSTAINABILITY-SPECIES**
14	city people island year long day area state water center park air project feet building back	**ENVIRONMENT-LOCAL**
15	years found scientists ago human million animals evolution evidence modern long birds Africa	**EVOLUTION**
1	case court evidence federal investigation government law judge officials department police cases	**LITIGATION**
2	research president state money administration federal bush house congress government million	**GOV**
6	science world life book people time art man century history things show made good day museum	**SCIENCE vs. RELIGION**
8	science university school research students professor work high college year institute education	**EDU**

The 15 topics automatically extracted from the 20-year collection are listed in Table I. Each row shows the rank of a topic, the most likely words in the topic, and our label to describe the topic.

The 15 topics provide a broad overview of the science technology related contents of the New York Times. The rank of a topic indicates the relative size of the topic in the collection. We reviewed the top 100 documents on each topic and created a label based on the general contents of those topic documents. Journal Science’s subject collections classify articles published by Science since 1996 into three large groups: life science, physical science and other subjects (Science Subject Collections, 2014). We divide our topics into two groups of topics based on it: traditional science topics (including life science and physical science subjects) and other subjects (including Economics, Education, Science and Policy, etc). Three topics: LITIGATION, GOV, SCIENCE vs. RELIGION and EDU are in the second group. They are grouped together at the bottom of the table. The rest of the topics are focused on a certain scientific topics in either life sciences or physical sciences.

B. Public Science Interests on the Web

Queries entered into search engines can be useful resource for detecting people’s information needs. *Google*, as the most widely used search engine today, is selected as our data source to answer our second research question. An estimated 12 billion search queries were conducted at *Google Search* by Americans in November 2013 alone, representing 66.7% of all search queries run during that time period in the USA [17].

In particular, we use Google Trends (http://www.google.com/trends/). Google Trends analyzes and displays the proportion of searches for terms compared to the total number of searches made on Google over a defined period of time (between 2004 and the present). The tool also allows limiting analysis in certain categories, like, science, news, health, etc.

In order to use Google Trends, we need to select a list of specific queries. We picked our queries from the topics we automatically extracted from mass media. We only used the 11 scientific topics and exclude the 4 context topics in our analysis. For each topic, several search queries were formulated using the topic words obtained in the topic modeling process. We also limit Google Trends to certain categories to reflect public's interests on the scientific topics, rather than other issues of the topic. The rank of average number of search over time is in Table II. Fig. 2 shows the trends of 11 topics from 2004 to 2012.

TABLE II: SEARCH POPULARITY OF SCIENTIFIC TOPICS

Rank	Topic label	Average
8	LIFE-HEREDITY	5.91
1	BRAIN	18.61
11	IT	1.94
5	SPACE	7.45
10	NUCLEAR	3.19
6	PHYS-MATTER	7.41
9	MEDICAL	5.69
2	GLOBAL WARMING	15.21
4	SUSTAINABILITY-SPECIES	9.73
7	ENVIRONMENT-LOCAL	6.36
3	EVOLUTION	14.02

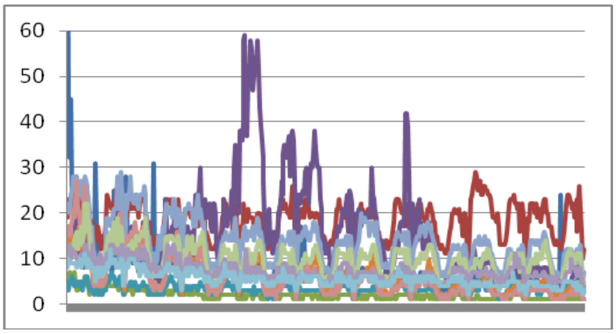


Fig. 2. Trends in 11 popular mass media science topics.

Space has the highest weekly search in early 2004 which are not fully shown in the figure. Global warming is another one with periods of surge of search. This growth in searches might be accompanied by an increase in media attention.

C. Mass media and Web Search

To prove the cause relationship between media coverage and information needs on the topic represented as web search, we analyzed the relationship between the amount of news coverage on a topic and the number of Google search at the same period of time.

We first need to decide how to calculate the amount of mass media coverage on a topic. In Topic modeling, each document is treated as a composition of a set of topics. Some topics may be very important; some may be barely mentioned in a document. We use the topic modeling assigned composition weight of a topic in a document to measure the relevance of a document to a topic. The simplest method will be the sum of total pieces of news report. However, we believe that different types of news reports have different levels of influence on the public. We decide to assign different weights to different media types based on their influence. TV news reports are divided into two categories, World news and others. Articles in World news category are weighted by 20 while others are 5. For newspaper reports, we categorize them on their positions into three groups: articles in Section A and at Page 1 weights 20, articles at Page 1 but not of Section A weights 5, all others 1. So for a topic, its media coverage in a certain period of time is calculated as:

$$C = \sum_{i}^{n} w_{t(d_i)} \times p(d_i, t)$$

d_i is document *i* in the set of documents on the given topic. There is a total of *n* documents in the set. *w* is the weight of d_i 's category. $p(d_i, t)$ is the relevance of document d_i on topic *t*.

We also need to pick a time unit for correlation analysis. Given that the topics are general we chose month as our analysis unit, rather than the smallest weekly unit available in Google Trends. Google Trends downloaded data are by weeks. So we grouped Google Trends data into monthly data.

We conducted a Pearson correlation analysis between monthly weighted media coverage and Web search about a topic reported through Google Trends. Pearson correlation coefficient is calculated as the measure of the linear relationship between the two variables. Table III summarizes the results.

TABLE III: PEARSON CORRELATION BETWEEN MEDIA COVERAGE AND GOOGLE SEARCH

Topic	Google Trend Category	Correlation	Sig
LIFE-HEREDITY	science	0.26	0.01*
BRAIN	Mental Health	0.171	0.77
IT	Computer Tech	0.039	0.69
SPACE	science	0.78	0.00*
NUCLEAR	science	0.424	0.00*
PHYS-MATTER	science	0.136	0.16
MEDICAL	health	0.264	0.01*
GLOBAL WARMING	Ecology environment	0.437	0.00*
SUSTAINABILITY-SPECIES	science	0.227	0.02*
ENVIRONMENT-LOCAL	Ecology environment	0.407	0.00*
EVOLUTION	science	0.461	0.00*

The results show 8 topics with significant correlations between media coverage and Google search. By examining the three topics, IT, PHYS-MATTER, and BRAIN we don't see significant results, we found they are more general compared with other topics. Because we only pick a few queries to represent Web search on the topic, it might be the queries we picked could not cover the whole scope of each topic. MEDICAL is also a broad topic, however, the media report and general public interests on it are relatively focused, say, certain types of diseases.

V. Discussion

The above presented results provide answers to the research questions:

What are the science and technology topics that appear most frequently in a representative US mass media and what are the contexts in which these topics appear? We have found that the 10 most common and persistent science and technology topics that have appeared in the *New York Times* and four TV channels are in rank order: Life-heredity, Nuclear power, Space exploration, Medical research, Conservation of species, Information Technology, Education, Physics-matter, global warming, and evolution. These topics involve all contexts of Personal, Social and Global as well as Health, Natural Resources, Environment, Hazards, and Frontiers of Science and Technology. We have also identified specific terms that define the above topics. For example, the following terms define the topic of Global Warming: climate, water, global, carbon, warming, change, ice, environmental, energy, sea, emissions, oil, and dioxide. We can also drill in to specific news reports to find out how science and technology is involved. The topics are clearly interdisciplinary and multidisciplinary, a trend in current science and engineering research (NRC, 2005).

Is there a correlation between the mass media report of a scientific topic with public interests on the Web about the topic? We found that the popularity of 11 topics extracted. In general, it was found that science related searches decreased over the past 20 years, both for general and specific topics. “Global Warming” is the only exception. We observe significant correlation between news media coverage and Google web search on 8 extract topics. They are more specific topics. For a general topic, our method might not pick enough queries to cover general publics’ interests. In general, the results show the influence of media coverage on scientific topics on general public’s web interest on those topics. This provides new insights into public understanding of science. Even it is reported in the survey that the importance of traditional media as a channel for general public to obtain scientific information, the news coverage does have influence in initiating public interests on the topics and to explore on the Web.

References

[1] National Science Board, “Science and Technology: Public Attitudes and Understanding,” *Science and Engineering Indicators 2008*, Washington, DC: US Government Printing Office, 2008

[2] National Science Board, *Science and engineering indicators*. Arlington, VA, 2012

[3] J. D. Miller, “Scientific literacy: A conceptual and empirical review,” *Daedalus*, vol. 112, no. 2, pp. 29-48. 1983.

[4] J. D. Miller, “The measurement of civic scientific literacy,” *Public Understanding of Science*, vol. 7, pp. 203-223, 1998.

[5] J. D. Miller, “The conceptualization and measurement of civic scientific literacy for the twenty-first century,” in *Science and the educated American: A core component of liberal education*, J. Meinwald and J. G. Hildebrand, Eds. American Academy of Arts and Sciences, 2010.

[6] B. S. P. Shen, “Science literacy and the public understanding of science,” in *Communication of Scientific Information*, S. B. Day Ed., pp. 44-52, Basel, Switzerland: S. Karger AG, 1975.

[7] M. S. Schafer, “Taking stock: a meta-analysis of studies on the media’s coverage of science,” *Public Understanding of Science*, vol. 21, pp. 650-663, 2012.

[8] R. A. Calvo, “Classifying financial news with neural networks,” in *Proc. the 6th Australasian Document Computing Symposium*, 2001.

[9] W. Zheng, E. Milios, and C. Watters, “Filtering for medical news items using a machine learning approach,” *AMIA Annual Symposium Proceedings*, pp. 949–953, 2002.

[10] D. Brossard and J. Shanahan, “Do you know what they read? Building a scientific literacy measurement instrument based on science media coverage,” *Science Communication*, vol. 28, no. 1, pp. 47-63, 2006.

[11] Y.-H. Tseng, C.-Y. Chang, S.-N. C. Rundgren, and C.-J. Rundgren, “Mining concept maps from news stories for measuring civic scientific literacy in media,” *Computers & Education*, vol. 55, pp. 165-177, 2010.

[12] A. K. McCallum. (2002). *MALLET: A Machine Learning for Language Toolkit*. [Online]. Available: http://mallet.cs.umass.edu

[13] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proceedings of the National Academy of Sciences of the United States of America*, vol.101, no. 1, 2004.

[14] D. Newman and S. Block, “Probabilistic Topic Decomposition of an Eighteenth-Century Newspaper,” *Journal of the American Society for Information Science and Technology*, vol. 57, no. 5, 2006.

[15] R. K. Nelson. (2010). *Mining the Dispatch*. [Online]. Available: http://dsl.richmond.edu/dispatch/

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[17] ComScore. (December 2013). U.S. Search Engine Rankings. [Online]. Available: http://www.comscore.com/Insights/Press_Releases/2014/1/comScore_Releases_December_2013_US_Search_Engine_Rankings

Ying Sun received her B.S. degree in information science from Peking University in 1996. In 1999, she received her MLS degree from the Information Management Department at Peking University. She entered the Ph.D. program in School of Communication, Information and Library Studies, Rutgers University, in 1999. She received her Ph.D from Rutgers, the state university of New Jersey in October 2005. She also obtained a master degree in computer science (also at Rutgers) while completing her Ph.D studies in 2003. Dr. Ying Sun is an associate professor of information studies at the University at Buffalo, the State University of New York. Her research interests are in the areas of information retrieval (searching for information), text mining (detecting significant patterns or trends) and formal information system evaluation.