Copyright

by

Russell Glasser

2007

Data Mining Analysis of Subject Priorities Among Prominent News Corporations

by

Russell Wain Glasser, B.S. Computer Science

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin
December 2007

Data Mining Analysis of Subject Priorities Among Prominent News Corporations

Approved by
Supervising Committee:

Dedication

This thesis is dedicated to my father Alan Glasser, whose guidance and encouragement
has always inspired me to seek new heights of learning;
To my son Ben, who may not have always understood what I was doing, but will
understand the extra time that I spend with him from now on;
Most of all to my wife Virginia, who cared for me and put up with me during these
difficult years.

Acknowledgements

I would like to acknowledge the contributions of Dr. Joydeep Ghosh, who served as supervisor on this report. My first class at UT was Data Mining under Dr. Ghosh, and I enjoyed his obvious enthusiasm for the subject.

I also wish to acknowledge the input of Iris Chyi, UT Journalism professor, who discussed issues surrounding the agenda-setting theory with me and pointed out several useful texts on the subject.

Finally, I wish to applaud the efforts of *The Daily Show* and the staff of Air America Radio. Collectively, they helped me understand that although entertainment does not constitute news, the two are not mutually exclusive.

November 30, 2007

Abstract

Data Mining Analysis of Subject Priorities Among Prominent News Corporations

Russell Wain Glasser, M.S.E.

Supervisor: Joydeep Ghosh

In recent years, major news corporations seem to dedicate an increasing amount of time and space to "fluff," reporting on celebrities, entertainment and crime stories, rather than more essential national and international news. As such news content is increasingly gathered online, it has become feasible to aggregate large amounts of data from a wide range of sites. This report proposes a model for collecting information from news agencies, then applying the techniques of Data Mining to organize this reporting in a way that identifies the priorities of individual organizations.

In addition, the rise of user-based taxonomies has made it possible broadly to evaluate the interests of people who actively read and recommend news. In the final analysis, data collected from users of Digg.com are compared with data collected from media sites. This provides a benchmark for determining whether the delivery of "fluff" news is delivered is a fair response to popular demand, or whether typical news readers are dissatisfied with the level of serious event coverage found in the media.

Table of Contents

List of Tables

List of Figures

List of Illustrations

INTRODUCTION

Chapter 1: The Nature of Modern Media

It is Monday evening, and Jon Stewart is on a roll. On the satirical news program, *The Daily Show*, Stewart briefly notes the important news of the day on June 11, 2007: Peter Pace has been ousted from his job as Chairman of the Joint Chiefs of Staff, America's highest-ranking military officer. This information is immediately followed by a clip of a CNN anchor saying: “We’ll talk about that more at the top of the hour; we do have some live pictures out of Paris Hilton that we want to talk about.”

Stewart watches in mock disbelief as clip after clip shows major news networks giving lengthy coverage to the supposedly important story of Paris Hilton being put in a car following her arrest. He notes: “All networks covered it gleefully… but know this: CNN didn’t want to!”

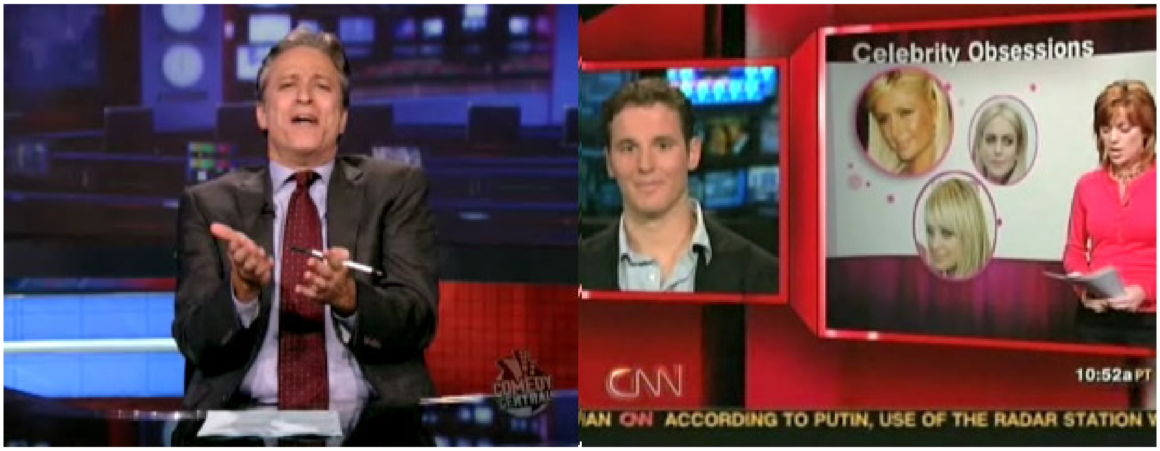


Illustration 1: Jon Stewart reacts to coverage of Paris Hilton

There follows a rapid montage of prominent CNN anchors scoffing at the insignificance of the story, even as they continue to talk about it around the clock.

> "We've always avoided even mentioning the name of the hotel heiress, because we can never figure out what she was famous for."
>
> "What is the obsession with the woman – and I'm talking about Paris Hilton – who does absolutely nothing?"
>
> "We're not sure what upsets you [viewers who sent feedback to CNN by mail] more: Paris going to jail, or the fact that we're even **covering** this story…"
>
> "Of course it is the case of Paris Hilton and – I know! But hear me out…"
>
> "Are we just so pathetic and so lonely that we have to live through people like Paris Hilton?"

Naturally, Jon Stewart is ready with the punch line. "If by 'we' you mean 'CNN', and if by 'lonely' you mean 'nobody's watching you,' then, uh… yes. Poor CNN. 'Why are they making us do this?'"

So reluctant were they to cover the story, he adds, that they covered it all day. (The Daily Show, 2007.)

DO MEDIA OUTLETS COVER ISSUES THAT DON'T INTEREST PEOPLE?

Though *The Daily Show* is a comedy program, it raises a serious point. If CNN anchors really do not wish to cover stories like the incarceration of Paris Hilton, then who forces them to? In theory, the for-profit corporation is answerable to its viewers, whose eyeballs translate directly into advertising dollars. But do cable subscribers really want wall-to-wall discussion of Paris Hilton? Or is it really the case, as one correspondent seemed to say, that the feedback was overwhelmingly negative? And if the latter is true, then why does the story receive so much attention?

These are questions which are presumably well explored by the marketing experts and business leaders who manage CNN. For ordinary consumers of news, however, their reasoning is a bit more opaque. Presumably a news channel has two goals, which may sometimes come into conflict. The first is to present serious news which is intended to inform viewers of important current events. The second is to entertain people and make a profit.

Suppose that it were possible to objectively divide the news into two categories: "relevant" news and "sensationalist" (or "fluff") news. Suppose also that it were possible to track these stories along two axes: first, the amount of attention given to each type of story by media outlets; and second, the level of interest in each type of story among readers. If this level of categorization could be accomplished, then we might discover one of three possible outcomes:

1. The coverage of sensationalist news is less than the public demand for sensationalist news. In this case, we might explain the data by claiming that media managers view their mission as one of presenting serious news, and to a certain extent manage to disregard their function as entertainment and avoid catering to the lowest common denominator.
2. The coverage of sensationalist news is about equal to the public demand for it. In this case, one could explain the data by saying that managers sees their primary mission as one of catering to public demand, providing entertainment to the widest possible consumer base, in order to maximize profit regardless of journalism concerns.
3. The coverage of sensationalist news is significantly greater than the public demand for it. This may be the most interesting outcome, although perhaps the hardest to explain. If the media are intentionally delivering fluff that the public does not want to know about, then there is some other motive besides either journalistic integrity or broad appeal to an audience.

> Although the true motives cannot be determined by this study, some speculations will be discussed at the end of the paper.

Of course, “the media” is a plural term, which does not describe a single monolithic entity. Many different corporations exist, with highly variable motives and missions. For this reason, it is necessary to analyze media outlets on a case by case basis, rather than as a uniform group.

With this in mind, the purpose of this project is to collect reporting patterns from a wide range of media sources and compare their priorities with that of the public interest. Sources covered will include respected mainstream newspapers, cable news, and some (perhaps) deliberately biased organizations.

ANALYSIS OF THE PROBLEM

Chapter 2: The Journalism Angle

Maxwell McCombs is a professor of journalism at the University of Texas. Along with Donald Shaw, McCombs is widely known for conducting the first systematic study of the agenda-setting hypothesis of media in 1972. This developed into a popular theory of journalism which states that "the news media, by their display of news, come to determine the issues the public thinks about and talks about." (Severin and Tankard, 2000, p 207.)

The agenda-setting theory stands in contrast to the notion that popular media merely reacts to public sentiment. It is also in contrast to the "magic bullet theory" of journalism, which holds that the media directly plants opinions into people minds. Agenda-setting theory states that the media does not necessarily have the power to directly change people's minds, but it does have the power to change the issues on which people focus their attention.

For example, early in the 20th century, a reporter named Lincoln Steffens wrote about how he got in a competition with a reporter another newspaper, where each tried to find crime stories in order to outsell each other. Soon, many other New York papers worked to find crimes and keep up with the others. As a result, there was widespread public perception of a crime wave, which even drew the attention of Teddy Roosevelt. Yet there was no crime wave; merely eager reporters shifting the public's attention so that they came to see crime as the most important issue of the day. (Severin and Tankard, 2000, p. 207.)

The 1972 story by McCombs and Shaw focused on polling undecided voters in Chapel Hill, North Carolina. A similar study was performed five years later in which they demonstrated that over time, public perception of the relative importance of an issue is more likely to be influenced by the number of stories appearing on that topic, rather than the other way around. (Severin and Tankard, 2000, pp. 209-211.)

Chapter 3: Modeling the Problem with Web 2.0

One problem with conducting studies via survey is that it surveys are expensive, requiring significant time and manpower to collect a representative sample of the general public. Another problem is that people who respond to opinion polling may be a self-selecting and non-representative set of the general public.

As the last decade of the 20th century saw the rise of the World Wide Web as a serious media phenomenon, the early 21st century has witnessed the rise of a suite of software technology collectively described by the buzzword "Web 2.0." Web 2.0 sites focus on distributed content creation. No longer is web content generated exclusively by technically adept gurus; instead, a growing arsenal of server tools allows people with essentially no programming experience to effortlessly contribute to a vast information network.

Examples of Web 2.0 sites include Wikipedia[1], a user-created encyclopedia that can be edited by anyone with an account; Blogspot[2] and other "web logging" sites (or "blogging" for short), where individuals can publish textual posts in which they present news content or personal stories to anyone who is interested in keeping track of their blog; and Digg[3], a social content sharing website at which some users submit the URL of existing web pages and then other users rate the contents of those pages.

With all this information being spontaneously generated by a user base numbering in the millions, the time seems right to ask whether there is a cheaper and more efficient method for identifying the focus of public opinion. Rather than hiring people to laboriously pore through physical newspaper clippings and watch hours of television, we

[1] http://www.wikipedia.org
[2] http://www.blogspot.com
[3] http://www.digg.com

can take advantage of convenient sites like the Google News archive[4], where dedicated software engineers have already done the hard work of collecting searchable text from many thousands of online media outlets. Rather than calling people on the phone to learn their opinions, we can use content-sharing sites such as Digg to learn about broad patterns of people's reading habits. That information has already been made freely available; we need only figure out how to assemble it in meaningful ways.

[4] http://news.google.com/archivesearch

SOFTWARE DEVELOPMENT

Chapter 4: Planning the Analysis

The first problem that must be addressed in searching the Internet for content is that, while existing software is very fast at retrieving archived news stories, it is not always easy to identify what a particular story is about. A human reader who looked at a news article might scan the first paragraph and immediately conclude that the article is about, for instance, presidential candidate Barack Obama. Individuals who are well informed about the news might even learn most of what they need from an article merely by reading through a few paragraphs. However, despite efforts that researchers have made towards artificial intelligence since the invention of the computer, a simple software package that reads the content of a story and identifies its significance is currently beyond easy access.

It might have been possible to create an “intelligent” news reader for this project, which could take a given news article and parse it to find key topics. However, developing an automated news categorization program is a tricky proposition, and this was deemed to be a tangential issue to the data mining issues that are the focus of this project. Also, reading and organizing news is something that Google News has already done, and I wished to duplicate their efforts as little as possible.

THE SEARCH FOR NEWS TAGS

Web tagging is a subject that has gained considerable attention lately. On many sites that are geared toward user content creation, contributors are allowed to assign “tags” to each resource that they generate. On Wikipedia, for example, a given article may be placed in multiple categories, which are arranged by users in a hierarchy to enable later visitors to easily browse items with similar tags. Tags are searchable, offering keywords that can be used by a database. Unlike reading the full text of an article, the use of tags allows an article to be briefly identified by subject.

Therefore, my first hope was that Google News would itself assign tags to articles and make the job of collecting data easy. Unfortunately, it turned out that this was not the case. Failing to find this easy solution, I then searched the web for a site that, like Digg.com, might post contemporary articles and encourage others to assign manual tags. However, I could not find any such site.

On further reflection, it is obvious why such a search must be futile. To give an idea of the scope of the categorization problem, consider these figures. When you visit news.google.com, you might typically be greeted with a number of stories resembling the following:

> **State Dept. steps up watch on Blackwater**
> Chicago Tribune - 23 hours ago
> By Aamer Madhani and Bay Fang | Washington Bureau October 6, 2007
> WASHINGTON - With Congress poised to expand laws overseeing private security contractors in Iraq, the State Department announced a new set of procedures Friday that will allow for closer ...
> Iraqis Claim Jurisdiction, But US Also May Oversee Incident ... ABC News
> Iraqi-US commission holds first meet on Blackwater AFP
> New York Times - The Courier News - Worcester Telegram - JURIST
> all 1,352 news articles »

In turn, clicking on the link labeled “all 1,352 news articles »” takes you to a “cluster” of similar articles, all of which are dated within the last 5 days or so, and many being roughly within a single day.

Furthermore, the front page of Google News alone contains links to some 30-40 clusters, each cluster representing its own individual group of related news articles. The above cluster contained 1,352 articles, which makes it unusually large. A more typical front page cluster may contain something closer to 100-500 stories. A quick calculation reveals that if the front page of Google News displays 35 clusters which each contain perhaps 200 stories that were posted within a single day, this means that we would have to rely on users to accurately tag 7,000 new stories per day on the front page alone. This does not even include all other stories, which are not selected by the search algorithm to appear on the front page.

By way of comparison, Wikipedia (arguably the most successful web 2.0 site) has a statistics page[5] showing that a total of just over two million English language articles have been produced within its six year history. Dividing this up, we can verify that somewhere around 913 new articles have been generated per day. Users cannot be expected to tag all the stories on the front page of Google News by hand, when this would require more than seven times more daily activity than is applied on Wikipedia to accomplish.

To take a brief tangent from the topic of the report, it would be very valuable if major news corporations could be persuaded to apply a standard tagging format to their own articles as they are generated. This would greatly improve the browsing experience for readers of media web sites. However, this is not available at the present time, and therefore it is an empty wish for the purposes of this project.

INFERRING TAGGED NEWS FROM SEARCHES

The Google Corporation is known first and foremost for their innovative approach to web searching, and so it seemed logical to attack this problem from the opposite direction. Rather that looking at a large selection of stories at random and trying to

[5] http://en.wikipedia.org/wiki/Special:Statistics

determine their topic, it made sense to choose a number of specified topics as fields of interest, and then do a Google News search on that text string. That way, Google could interpret the topics and deliver stories or clusters which were relevant to previously identified chosen issues.

Therefore, my first programming task was to write a web page reader which could identify common topics. Luckily, Google News had a way to help out: each day, the front page shows a list of eight common terms, under the heading "In The News." I developed a simple page reader that I could run each day, which would transfer those terms into a "Topic" table in my database. After this had run for a couple of weeks, I was given a list of some 100 unique topics to choose from. I also added topics of my own interest, and then labeled the topics as "Interesting" (a binary field under the Topic table). I eventually wound up with the following list:

Topic	**Category**	**Description**
Abu Ghraib	US News	Prison in Iraq, currently under control of United States military. Abu Ghraib has been a prominent news subject since May 2004, when U.S. soldiers were discovered torturing and abusing prisoners.
Anna Nicole Smith	Celebrity	Former Playboy model who married elderly billionaire J. Howard Marshall. Briefly starred in a reality TV show. Died of a drug overdose on February 8, 2007, a few months after her 20 year old son also died
Barack Obama	Presidential candidate	Currently the second highest polling candidate in the field of Democratic candidates.
Britney Spears	Celebrity	Young female pop singer and sex symbol.
Giuliani	Presidential candidate	Current Republican front runner. Since first name is alternately reported as “Rudolph” or “Rudy,” only his distinctive last name is used here, in order to catch the most possible stories.
Harry Potter	Entertainment	Fictional protagonist of a popular series of books and movies.
Hillary Clinton	Presidential candidate	Current Democratic front runner.
John Edwards	Presidential candidate	Currently the third highest polling candidate in the field of Democratic candidates.
Mitt Romney	Presidential candidate	Currently the second highest polling candidate in the field of Republican candidates.
New Orleans	US News	Major United States city, mostly decimated by Hurricane Katrina in August 2005. Reconstruction of the city has continued to be slow and many destroyed homes have not been rebuilt.
Paris Hilton	Celebrity	Daughter of Richard Hilton, and heiress to the Hilton Hotel fortune. Hilton is regarded as a somewhat vacuous individual, yet receives frequent media attention. Former star of a reality television show called “The Simple Life.” Her arrest in 9/06 generated brief, intense coverage.
Tiger Woods	Entertainment (Sports)	Successful African-American professional golfer; this topic is intended to represent sports interests.

Table 1: List of collected topics

This list is meant to represent a range of interesting subjects, some of which qualify as serious news, and many of which definitely imply “fluff.” Several major presidential candidates are included, since it will be interesting to compare the relative level of coverage that the candidates are given.

Once the topics were chosen, it was apparent what would be the correct way to proceed in collecting data. First, do a web search on one of the above news topics. Then, find all links on the page labeled “all N news articles,” note the number, and save the link. This link would represent a single news “cluster”, an event that is covered by one or more media sources. (Stories that appear by themselves, and not in a cluster with other stories, are ignored.) Finally, read the contents of the cluster by following the appropriate link and reading off-site links.

Each link can be parsed fairly simply to identify what site it comes from. For instance, a link to

< http://www.washingtonpost.com/wp-dyn/content/article/
2006/04/13/AR2006041302159.html >

can be identified as coming from the Washington Post by extracting the first string after “http://” and before the next “slash” character: “www.washingtonpost.com”. To make later searching easier, a number is immediately associated with all web pages as soon as they are added to the database, uniquely identifying the source site, which then does not need to be extracted from the full URL again.

Naturally, the list had to be altered a bit throughout the course of the project. The list originally included the topics “Rupert Murdoch,” “Gulf Coast,” and “Blackwater,” which came up as popular subjects during my initial scans. The “Rupert Murdoch” topic was eliminated because it consistently resulted in a very low story count, which made the sample difficult to work with. “Gulf Coast” was eliminated because the subject was somewhat redundant with “New Orleans.” Although interesting and topical, “Blackwater” (a privately owned security organization which receives military funding for operations in Iraq) generated very little news until early 2007, so it lost value as a story when the program began scanning older stories. The "Abu Ghraib" topic was not sampled initially, but it was added near the end in order to include more subjects in the "News" category.

SELECTION OF MEDIA SOURCES

The first attempt to collect data simply involved doing topic searches several times over a few days, then recording the news that appeared in every cluster retrieved from these searches. After doing this data collection, I analyzed the database to determine which sites were responsible for most of the content. Based on this list, supplemented with a certain amount of personal preference, I eventually tagged the following web sites as "interesting" and therefore worth tracking.

I have listed my online sources below. The descriptions of these sites are subjective to an extent; they are based on opinions I had going in and cross-checked by reading their mission pages and other external descriptions, such as Wikipedia references.

Source	**Web Address**	**Perceived characteristics**
The New York Times	http://www.nytimes.com	Online location of large, established newspaper
The Washington Post	http://www.washingtonpost.com	Online location of large, established newspaper
Cable Network News (CNN)	http://www.cnn.com	Founded in 1980 by Ted Turner, CNN was the first major news channel to feature 24-hour news coverage.
ABC News	http://www.abcnews.com	ABC News is a division of the ABC network. They distribute news to television, radio, and the internet.
USA Today	http://www.usatoday.com	Daily national newspaper; widest circulation of any in the US. Known for simplifying stories for a broad audience and using colorful charts and tables.
Fox News	http://www.foxnews.com	Cable news network owned by Rupert Murdoch. Despite the slogan "Fair and Balanced," Fox News has a reputation as a highly partisan right wing network.
New York Post	http://www.nypost.com	Tabloid-style newspaper, also owned by Rupert Murdoch.
Wall Street Journal	http://online.wsj.com	Financial newspaper that covers business news. Conservative op-ed, but neutral reporting. The paper was also recently bought by Rupert Murdoch.
Washington Times	http://www.washingtontimes.com	Politically conservative daily newspaper, founded in 1992 by Sun Myung Moon, leader of the Unification Church.
BBC	http://www.bbc.co.uk	British Broadcasting Corporation, primary television news coverage in Great Britain.
The Guardian	http://www.guardian.co.uk	Somewhat left-leaning British newspaper
Salon	http://www.salon.com	Online-only magazine, focusing on American politics with a distinctly liberal point of view.

Table 2: List of media sites used

The principle behind this selection of web sites is to represent a wide variety of media sources. Some are print and some are television. Some are respected mainstream sources, while others have a particular bias associated with them. Some are considered serious news sites, while others, such as USA today, lean towards "fluff." Whether deserving or not, several of the sites have a specific reputation for deliberate political bias. More than other sources, television news may tend toward sensationalism by nature of the format.

In order to determine the priorities of the various news sites, I chose to look at all the clusters identified by Googe which ran over a minimum size (initially clusters with 10 stories or more) and see which news sites were featured stories within that cluster. Then I planned to identify each combination of web site and topic based on the percentage in which they were present in the relevant clusters.

IDENTIFYING POPULAR OPINION WITH DIGG.COM

Searching Digg proved to be much easier than searching Google News. In contrast to the 7,000 daily front page stories on Google News, Digg has a much smaller base of user-submitted stories to read. In fact, it turned out that I could collect every story relating to each of my topics, back through the beginning of Digg's existence.

Unlike Google News, Digg provides a fairly obvious way to "score" stories based on overall interest or disinterest in a topic. The process of visiting Digg as an ordinary user works like this:

1. Find an interesting web link.
2. Log in to an account on Digg, and click "submit new."
3. Enter a headline and a brief description for the link.

Other users who encounter a previous submitted article can click on a direct link to view the original story. They can also click a second link to arrive at a page within the Digg site, where they can comment on the significance of the story or engage in discussion with other users. Finally, and most importantly to this study, a link on one side says “Digg it.” Clicking this link will assign one point value to the story, and the overall score will displayed next to each story. At most, one point may be assigned by each reader.

In this way, all stories can be viewed and their scores can be tallied in just one page load. Once this is done, we can determine the total number of stories submitted for each topic, and the average score assigned to stories within that topic.

Chapter 5: Database Design

The database consists of six tables. Each table name is prefixed with "NM_", indicating that it is part of the News Miner group of tables.

The table schema is as follows:

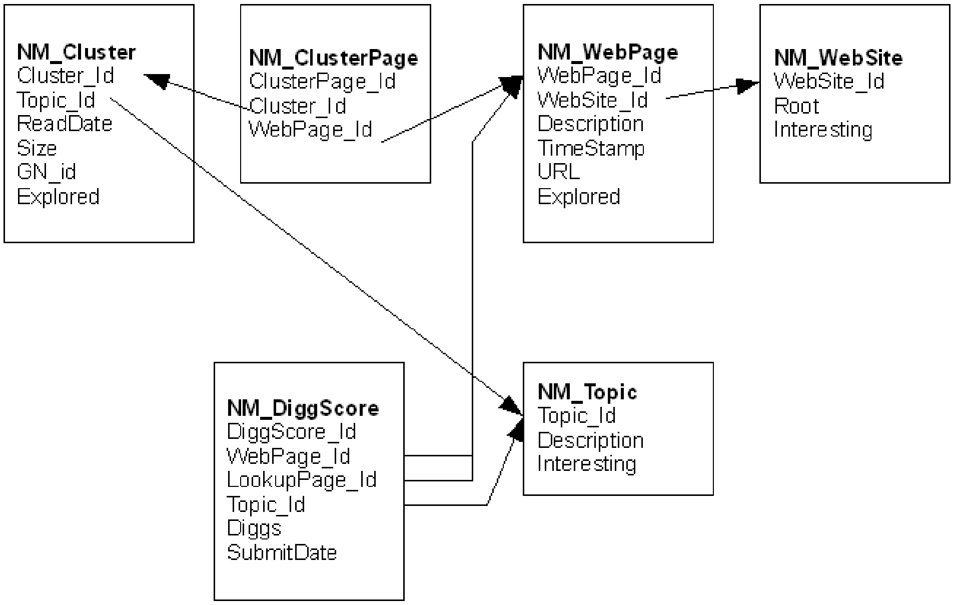


Figure 1: Entity-Relation Diagram

The description of the tables is listed below.

- **NM_Topic** is a descriptor of one topic in the news. Hundreds of topics were entered, but only the topics listed in Table 1 were scanned.
- **NM_WebSite** represents an individual site, such as www.nytimes.com or www.foxnews.com. The "Interesting" field is a binary value indicating whether it is being treated as an important news source in the analysis.

- **NM_WebPage** represents a single web page. It is associated with one **NM_WebSite** entry. The URL is the Uniform Resource Locator that identifies the site on the web. The TimeStamp represents the date and time at which the page was added to the database. The Description is a string that is usually pulled directly from the link text from which the story was obtained. However, the program is also designed in such a way that the link can be retrieved from the title field after directly visiting the page.
- **NM_Cluster** represents a cluster of Google News stories. Each cluster is associated with exactly one topic. The ReadDate is a date an time which is provided by Google News to indicate on what date the stories within the cluster occurred. GN_id is a unique identifying number that can be used to retrieve the search page from Google News. The Explored field is a boolean value which indicates whether or not stories have been collected from within the cluster yet. Stories are assigned to Clusters within the database via the **NM_ClusterPage** table.
- **NM_ClusterPage** is a table that associated WebPages with Clusters. Each ClusterPage entry links exactly one cluster id with exactly one web page id. The relation between WebPages and Clusters is intended to be one-to-many, so each cluster will have multiple stories but a web page can be associated with at most one cluster.
- **NM_DiggScore** represents one story collected from a Digg search. A Digg score has two web page associated with it: the original story, and a URL for the discussion page. Like Cluster, the Digg story has a topic based on the search term. IT also has a score (the “Diggs” field).

Chapter 6: Program Architecture

The News Miner program is written in Java. It uses the Java Database Connectivity (JDBC) library to access an online MySQL database at my website, http://mysql.apollowebworks.com. It also uses the JTidy library, a package which interprets HTML text which may not be well-formed, and converts them into Document Object Model (DOM) trees. JTidy can be found at http://jtidy.sourceforge.net/.

CLASSES

The program consists of 23 classes and approximately 3500 lines of code. Many of the Java classes directly correspond to tables in the database. Instantiations of classes such as **Cluster**, **WebPage**, and **DiggScore**, represent single table entries. Methods in these classes communicate with a common **MiningDB** class to read entries straight out of the database, and save new entries to corresponding rows in the table.

The **WebPage** class also has methods for scanning a page and collecting links, or viewing lists of other selected HTML tags in order to facilitate collecting new information.

There is also a **Digg** Class and a **GoogleNews** class, both inherited from a common **AggregatorSite** class. The classes for these aggregator sites contain static methods for retrieving the front page and individual search pages. There is also a **GoogleNewsPage** class and a **DiggPage** class, both of which extend the **WebPage** class. These classes are responsible for reading HTML patterns specific to those sites.

For instance, the **GoogleNews** class has a “getMonthlyClusters” method for exploring stories within clusters that have not yet been associated with an entries in the

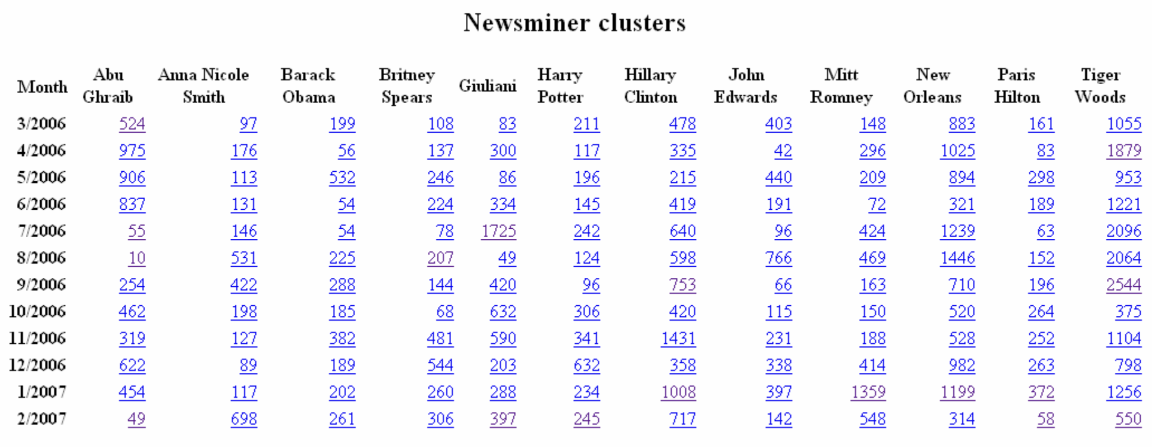
ClusterPage table. Calling "getMonthlyClusters" will generate a series of **GoogleNewsPage** instances. These pages are then read using the method **GoogleNewsPage**.getNewsStories, which in turn populates the database with new **WebPage** and **ClusterPage** entries. While a cluster is being explored, the autocommit property of the database is turned off. New web pages are committed only after all pages in a cluster have been retrieved. The **NM_Cluster** row is marked with the "Explored" field set to true in the same operation. Thus, a cluster in the database is always either completely explored, or completely unexplored.

I designed a simple interface on the Java application, offering a menu displaying options for various data-gathering tasks. The tasks are: 1. Get monthly clusters (one topic); 2. Get monthly clusters (all topics); 3. Explore current clusters; 4. Get new Digg scores (one topic); 5. Get new Digg scores (all topics); 6. Generate results table; 7. Analyze results.

Throughout the months when I was collecting data, I would execute one or more instances of the program each day to fill in gaps in the current data.

Chapter 7: Web Interface

A simple web interface was written in Perl/CGI, using the DBI library to access the SQL database. The resulting web pages are primarily intended to provide a way to conveniently drill down into the data, which is organized by topic, then date, then cluster (or Digg Score); and finally links to the stories on the original source pages.



Newsminer clusters

Month	Abu Ghraib	Anna Nicole Smith	Barack Obama	Britney Spears	Giuliani	Harry Potter	Hillary Clinton	John Edwards	Mitt Romney	New Orleans	Paris Hilton	Tiger Woods
3/2006	524	97	199	108	83	211	478	403	148	883	161	1055
4/2006	975	176	56	137	300	117	335	42	296	1025	83	1879
5/2006	906	113	532	246	86	196	215	440	209	894	298	953
6/2006	837	131	54	224	334	145	419	191	72	321	189	1221
7/2006	55	146	54	78	1725	242	640	96	424	1239	63	2096
8/2006	10	531	225	207	49	124	598	766	469	1446	152	2064
9/2006	254	422	288	144	420	96	753	66	163	710	196	2544
10/2006	462	198	185	68	632	306	420	115	150	520	264	375
11/2006	319	127	382	481	590	341	1431	231	188	528	252	1104
12/2006	622	89	189	544	203	632	358	338	414	982	263	798
1/2007	454	117	202	260	288	234	1008	397	1359	1199	372	1256
2/2007	49	698	261	306	397	245	717	142	548	314	58	550

Figure 2: Screenshot of the web interface

The above figure is a screenshot of the top level Google News browsing interface. This page displays a grid with topics on the horizontal axis, and months scanned on the vertical axis. Each cell in the grid displays a number representing the total number of prominent (front page) stories collected on one topic in one month.

Clicking on any cell takes you to a page where a number of clusters (up to 20 per monthly topic) are displayed. The clusters range in size from two up to several hundred stories. Each cluster can be clicked in order to visit the original Google News page, and you can then browse the stories in that cluster.

A similar interface also exists for exploring the submitted stories in Digg at their original locations.

At the current time, this page can be visited at http://www.apollowebworks.com/newsminer

Chapter 8: Revising the Approach

As every programmer knows, even the best laid plans are always subject to change based on new information discovered in the process of developing software. This section lists some of the issues that I had to deal with in generating my data.

MOVING FROM DAILY NEWS TO THE HISTORICAL ARCHIVE

After a few days of collecting data from the main Google News site, it became clear that reading current news on a daily basis would not generate a sufficient amount of data to create an acceptable sample size. It was at this point that I began to explore the Google News archive. The archive is a recently developed tool, available from the main news page, which collects stories from many sources going as far back as the 19th century. Naturally, the farther back in time you search, the sparser are the stories you can locate. This wasn't a problem for me, however. For my needs, it was sufficient to search back through the last year of stories.

I soon discovered that there is a large gap in the search capabilities of the archive. The archive only contains stories up to mid-February of 2007, and the main news search page only contains stories going back through the last month before the date on which a search is conducted. Running the program in late revealed a large coverage gap: stories from March through August were not searchable by either method. Therefore, I set the range of my searches to cover one year, spanning from March 2006 to February 2007.

Some of the news sites had a much lower presence in the archive site than in the main news site. Switching from an ordinary Google News search to an archive search required adding and dropping a few web sites from the "Interesting" category, in order to make sure that all sites had a similar level of presence in the clusters.

GETTING AROUND GOOGLE'S MALWARE DETECTION FEATURE

Once the program began gathering data from the archive, a new problem unexpectedly materialized. Topic searches suddenly began to return no clusters at all. The results of one search seemed to indicate that there had been no news at all about Anna Nicole Smith during the entire month of March. A quick check with a browser made it clear that this was not true.

Going through the program in a debugger led to a visited URL which contained this message:

> *We're sorry...*
>
> *... but your query looks similar to automated requests from a computer virus or spyware application. To protect our users, we can't process your request right now.*
>
> *We'll restore your access as quickly as possible, so try again soon. In the meantime, if you suspect that your computer or network has been infected, you might want to run a virus checker or spyware remover to make sure that your systems are free of viruses and other spurious software.*
>
> *We apologize for the inconvenience, and hope we'll see you again on Google.*
>
> *To continue searching, please type the characters you see below:*

This was followed by a typical set of "CAPTCHA" characters to type. (CAPTCHA is an acronym that stands for "Completely Automated Public Turing test to tell Computers and Humans Apart." [6])

[6] More information can be found at http://www.captcha.net/.

Evidently, Google's software was smart enough to recognize a computer program repeatedly accessing similar searches within a very short period of time. A post on Google's blog explains the intent behind this feature. According to Niels Provos from Google's Anti-Malware Team:

> "we have seen self-propagating worms that use Google search to identify vulnerable web servers on the Internet and then exploit them. The exploited systems in turn then search Google for more vulnerable web servers and so on. This can lead to a noticeable increase in search queries and sorry is one of our mechanisms to deal with this." (Provos, 2007)

Certainly no one could fault Google for acting on their legitimate interest in protecting their servers from being overtaxed by viruses and possibly denial of service attacks. However, the automated approach they had chosen rendered it very difficult to convince them that I was a real person with an important task to complete. I attempted to get through to a human respondent in the hopes of obtaining special permission to carry on this research, but no one responded to my email and newsgroup postings.

Since the authentication was done on a per-browser basis, typing the CAPTCHA text in a browser such as Firefox did not solve the problem within the program. The project had moved beyond strict data mining issues and into the realm of internet security. I was reluctant to invest time in getting the program to answer the CAPTCHA challenge. So instead, I dealt with the problem first by physically switching IP addresses. For a few days I found myself transporting a laptop computer to various coffee shops and classrooms in order to grab data in brief bursts before being interrupted. In order to be considerate to Google's resources, I added a "sleep" routine so as to avoid hitting their servers too hard. This was only a temporary solution, however.

I am greatly indebted to a contact who uses the internet handle "nephlm," since he suggested the eventual solution to my problems. The Electronic Frontier Foundation sponsors a program called "Tor," which is described in this way:

> Tor is a toolset for a wide range of organizations and people that want to improve their safety and security on the Internet. Using Tor can help you anonymize web browsing and publishing, instant messaging, IRC, SSH, and other applications that use the TCP protocol. Tor also provides a platform on which software developers can build new applications with built-in anonymity, safety, and privacy features.
>
> Tor aims to defend against traffic analysis, a form of network surveillance that threatens personal anonymity and privacy, confidential business activities and relationships, and state security. Communications are bounced around a distributed network of servers called onion routers, protecting you from websites that build profiles of your interests, local eavesdroppers that read your data or learn what sites you visit, and even the onion routers themselves.

This proved to be an ideal solution to my problem. After installing Tor and configuring my program to access web sites through a local proxy port, I was able to make it appear as though my IP address changed periodically. Google lost the ability to track my activity, and I was able to continue collecting data uninterrupted.

This approach may perhaps raise some ethical issues. It is possible, though unlikely, that virus designers could learn from this approach. However, in order to use Tor, special software has to be knowingly installed on a client computer, and web traffic has to be redirected through a running proxy server. It seems unlikely that a virus could do all this work undetected. In any case, throughout the project I continued to instruct the program to sleep for a few seconds between each page access. This was no longer necessary, but I felt it was a considerate compromise to ease the load on Google's servers.

REDUCING THE SCOPE OF CLUSTERS SEARCHED

Once the program was running continuously, another limitation became apparent: time constraints. The Java database interface was designed to be as simple as possible on the development end, but the execution was not extensively optimized. The program uses an online database rather than a local file; this makes it possible to easily access the same data from any location without transferring files other than source code. The down side of this approach is that transferring data takes a noticeable amount of time. Even setting up the data in batch files and transferring several rows at once does not save very much time.

The program can identify new web page entries and associate them with clusters at the rate of about two per second. At this rate, scanning a single cluster of size 1,000 would take 8 minutes. The program was wasting far too much time scanning a few clusters with hundreds of stories in them; and it was also scanning a very large number of clusters that had only two or three stories each – most of which would not contain stories by any of the target sites.

Clearly, it would be preferable to scan a greater number clusters that are selected as being more likely to yield relevant results, rather than to scan all clusters sequentially regardless of their relevance. Therefore, I chose to reduce the search space.in the following way. Pick a range of cluster sizes – initially clusters containing between 50 and 100 stories – and only scan clusters within that range. The total number and size of all clusters may be read independently, but for identifying sites that are of interest to individual news organizations, these selected cluster sizes would presumably make a good representative sample from all the stories.

FOCUSING ON MORE RELEVANT CLUSTERS

Toward the end of the project, I began developing the web interface described in chapter 7 so that I could view monthly clusters at a glance and see what kind of data was being pulled in each month. The initial results were a bit discouraging. I discovered that many of the clusters were not genuinely related to the topic with which they were tagged.

In fact, apparently in the majority of cases, the first few stories in each cluster would have the search term appearing somewhere deep within the article, and then the remaining stories would have no obvious relation to the term at all.

To pick one example, for the topic "Paris Hilton" in March 2006, the largest cluster in the database contained 105 stories. Readers can currently view this cluster at < http://news.google.com/archivesearch?q=Paris-Hilton&cid=8616030813290510 >

The first story in this cluster is titled "Morrison or motivation?" It is a story in a blog associated with a paper in Louisville, KY. It begins with this sentence: "While I would argue IU point guard Earl Calloway has the mustache of all staches, it's a **Paris Hilton** no-brainer that Adam Morrison is one of, if not THE, best players in the country." Other stories in this cluster are similar stories about professional basketball player Adam Morrison.

These stories have nothing to do with Paris Hilton at all. The only relation is a thrown away remark at the beginning of one article which used the name as a colloquialism. Such was the case in a distressingly large number of the stories that had been scanned.

Obviously I had made a mistake. I had started searching for stories on Paris Hilton in March, and Google News returned 89 pages of results. However, Google sorts these clusters by relevance, so the first page contains the stories which are most definitely coverage of the appropriate subject. The first cluster returned is titled "Paris Hilton Hit With Restraining Order," which is obviously relevant. However, the first cluster on the last page is titled "News Briefs" and contains the phrase: "this program was set up to save money. But because of poor oversight, it has spent money like Paris Hilton at a shoe sale." Again, Paris Hilton's name appears as an easy pop culture reference, while the story itself is not about Paris Hilton, or indeed about celebrities at all.

Other topic searches yielded similar results. Searching for a politician’s name may yield relevant results in the first few pages, but later pages frequently contain stories that are generally about the politician’s home state, and only briefly mention the name itself; or they contain stories about the field of presidential candidates and not one politician in particular. Worse, many online news publications have side bars containing a selection of randomly chosen popular stories. A story may well be about labor relations, and yet contain the words “Britney Spears” in an unrelated link on one side. This story can show up in a “Britney Spears” cluster, and then all related stories are tagged by the scan as being about Britney Spears.

To remedy this problem, I decided to ignore a significant number of already explored clusters and rescan the search topics, this adding a “Relevance” score to clusters appearing on the first search page, in order of display. Because this resulted in fewer clusters, I expanded the exploration section to scan all clusters with a size between 25 and 100, rather than between 50 and 100.

DATA ANALYSIS

In this section, I will outline the analysis that was done on the final data. Some of this analysis was generated in the form of graphs on Microsoft Excel; some is analyzed using a combination of Java and the WEKA data mining package.

Chapter 9: Viewing Coverage Over Time

As shown in Figure 2, the web interface displays an abridged table showing the number of stories contained within collected clusters on the various topics each month. Once this component was working, I collected all the Google News Stories in an abridged table, and all the Digg stories in another table. Here are the results from Google News, broken down by total number of clustered stories appearing in the first 20 clusters in each month.

Site	Abu Ghraib	Anna Nicole Smith	Barack Obama	Britney Spears	Giuliani	Harry Potter	Hillary Clinton	John Edwards	Mitt Romney	New Orleans	Paris Hilton	Tiger Woods
Mar-06	524	97	199	108	83	211	478	403	148	883	161	1055
Apr-06	975	176	56	137	300	117	335	42	296	1025	83	1879
May-06	906	113	532	246	86	196	215	440	209	894	298	953
Jun-06	837	131	54	224	334	145	419	191	72	321	189	1221
Jul-06	55	146	54	78	1725	242	640	96	424	1239	63	2096
Aug-06	10	531	225	207	49	124	598	766	469	1446	152	2064
Sep-06	254	422	288	144	420	96	753	66	163	710	196	2544
Oct-06	462	198	185	68	632	306	420	115	150	520	264	375
Nov-06	319	127	382	481	590	341	1431	231	188	528	252	1104
Dec-06	622	89	189	544	203	632	358	338	414	982	263	798
Jan-07	454	117	202	260	288	234	1008	397	1359	1199	372	1256
Feb-07	49	698	261	306	397	245	717	142	548	314	58	550

Table 3: Google News clustered stories by month

The data from Digg is shown in a different way. While we can make a case that the total number of top search stories indicates the prominence of a news topic in a given month, this does not indicate the genuine popularity of a story on Digg. For Digg stories, our table will display two numbers: The total number of stories found, followed by the average reader-assigned score of each story.

Site	Abu Ghraib	Anna Nicole Smith	Barack Obama	Britney Spears	Giuliani	Harry Potter	Hillary Clinton	John Edwards	Mitt Romney	New Orleans	Paris Hilton	Tiger Woods
Mar-06	2, 11.50	0	0	6, 11.67	0	9, 11.00	0	0	0	9, 9.11	3, 7.33	2, 3.50
Apr-06	0	0	1, 6.00	2, 7.00	1, 3.00	2, 4.00	2, 5.50	0	0	3, 18.33	44, 4.16	0
May-06	0	0	0	1, 4.00	0	6, 9.83	4, 11.75	0	0	11, 3.91	121, 5.12	3, 3.67
Jun-06	0	0	0	24, 4.38	0	21, 4.00	9, 2.22	1, 8.00	0	12, 6.00	248, 4.03	3, 2.33
Jul-06	1, 25.00	0	1, 19.00	10, 2.20	3, 7.00	30, 5.57	12, 5.67	5, 4.20	0	21, 5.48	35, 5.31	10, 3.70
Aug-06	5, 6.80	0	1, 19.00	43, 4.49	8, 8.12	24, 6.08	22, 6.55	0	0	47, 6.79	103, 4.76	24, 4.83
Sep-06	9, 9.33	26, 5.31	8, 7.62	18, 4.17	9, 5.33	29, 4.52	10, 4.80	0	1, 5.00	24, 4.08	88, 4.62	19, 4.47
Oct-06	6, 6.67	7, 6.43	19, 6.26	16, 2.25	0	37, 5.11	18, 7.78	1, 1.00	1, 5.00	9, 3.00	52, 3.44	9, 3.44
Nov-06	7, 7.14	6, 1.67	15, 16.73	151, 4.50	20, 6.15	45, 4.24	9, 6.00	7, 11.57	8, 7.25	17, 5.24	87, 3.23	14, 4.29
Dec-06	2, 4.00	7, 2.14	31, 9.06	226, 3.28	9, 4.44	76, 5.38	13, 3.54	35, 6.94	0	32, 4.38	121, 3.37	16, 4.94
Jan-07	9, 11.33	10, 2.30	77, 6.99	158, 3.13	20, 8.25	38, 5.21	79, 7.82	12, 6.33	10, 4.50	91, 4.98	168, 4.61	29, 3.52
Feb-07	5, 9.00	451, 3.82	124, 5.78	373, 3.21	77, 8.47	116, 4.34	51, 6.06	41, 10.80	24, 5.21	63, 4.71	131, 2.63	19, 4.58

Table 4: Digg submissions by month

Viewing the table in this way reveals an important issue with the data. Notice that in early months, there are a large number of topics that contain either no stories, or a very small number of stories. This is because Digg is a relatively new site, and the number of stories submitted has increased as the site increased in popularity.

To deal with this issue and make sure that I had a relevant sample size across most topics, I decided to throw out the months before August in preparing my final analysis. (July had few topics with zero submissions, but it had several 1's, as well as a 3 and a 5.)

At this point I had two "scores" for each topic in each month: a magnitude of "relevant" stories from Google, and an average score from Digg. I further reduced the data by dividing topics into categories. Anna Nicole Smith, Paris Hilton, and Britney Spears were rolled into a single column labeled "Celebrity." Tiger Woods and Harry Potter became "Entertainment." (Tiger Woods is also a celebrity, but in fact most stories about him tend to focus more on the sport than the person.) New Orleans and Abu Ghraib were categorized as "News," and the rest were labeled "Politicians."

I then further normalized the monthly data by taking the maximum value from each column, and dividing the entire column by that number. For instance, in March 2006, the "Tiger Woods" column contains 1,055, so dividing the entire column by this value yields 0.497 (1055/524) for Abu Ghraib, and so on. This reduces all cells to a single value with a magnitude between 0 and 1, which makes it possible to make meaningful comparisons between the rows.

This approach yields a time sequence which is graphed below.

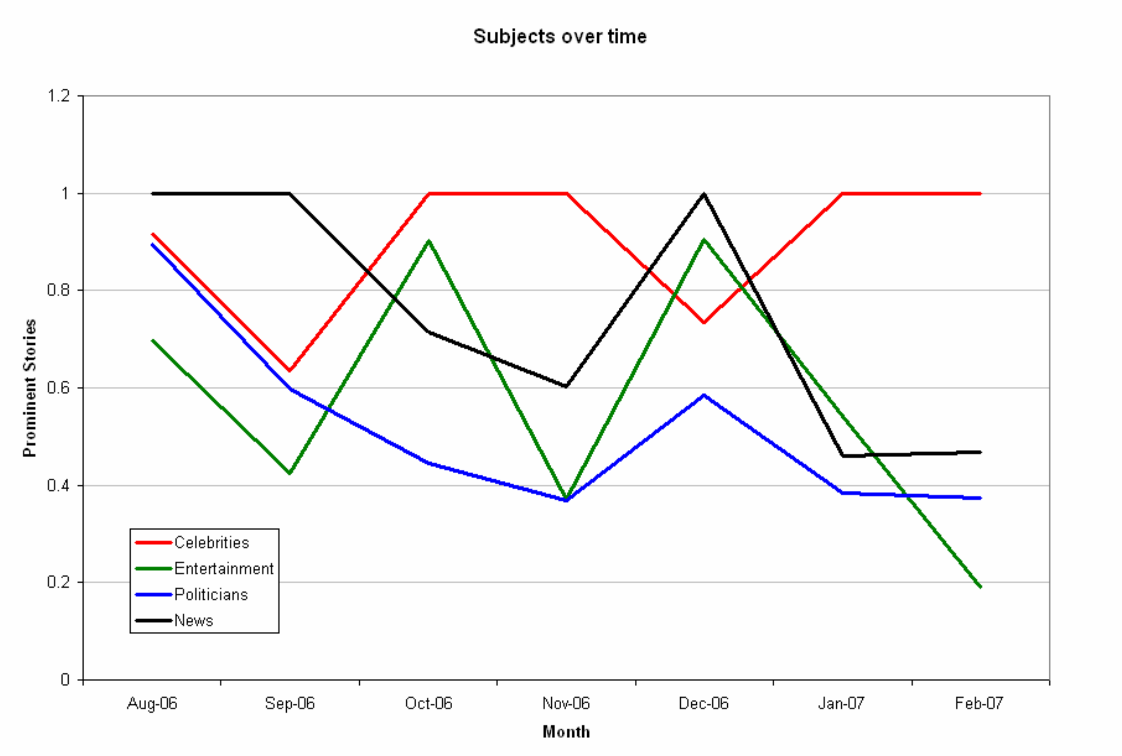


Figure 3: Monthly story priorities from Google News

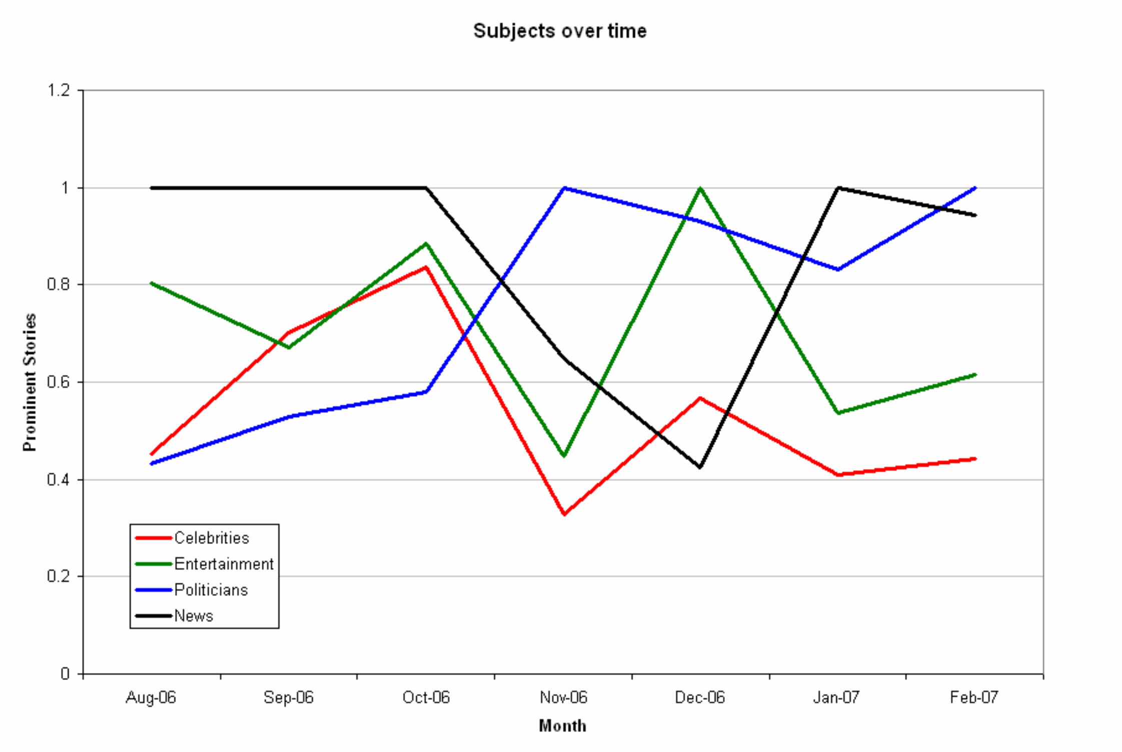


Figure 4: Monthly story priorities from Digg

Since the primary intent of this thesis was to determine how much of the news is devoted to "fluff", it may be instructive to single out the "Celebrity" category and show how Google News compares to Digg. The results follow.

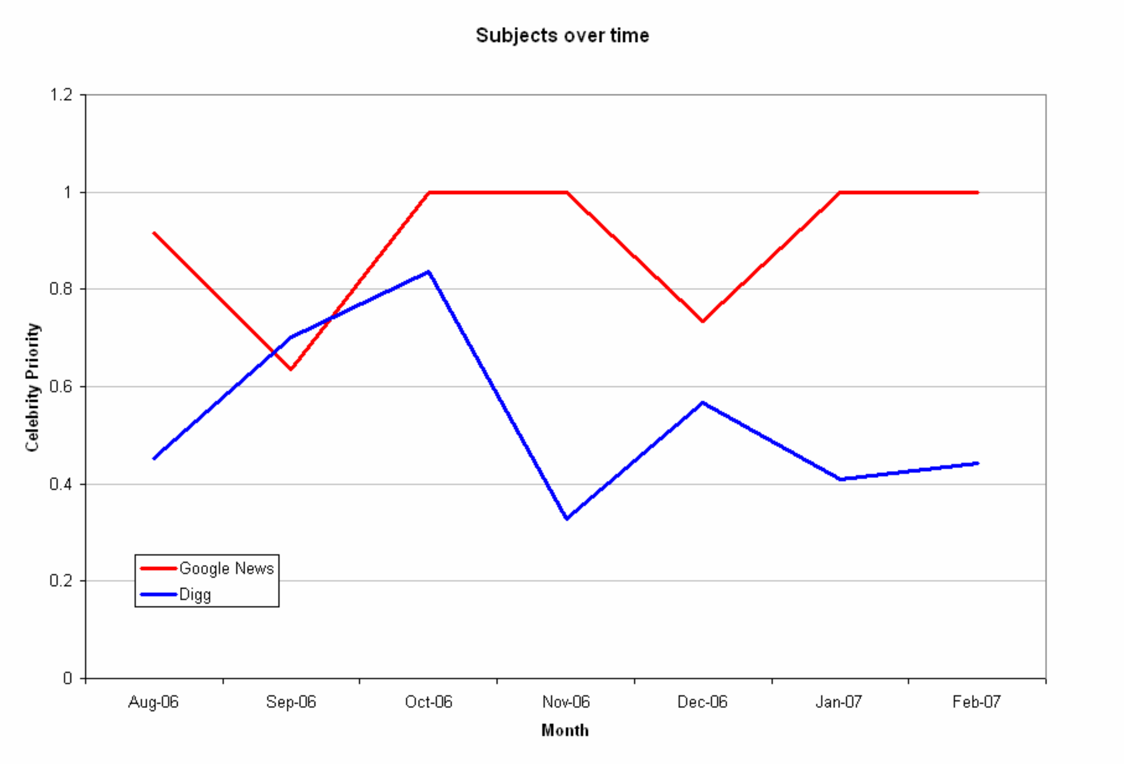


Figure 5: Comparison of focus given to celebrities

As you can see, in most months the media pays more attention to stories about the celebrities I selected than the Digg-reading public would like to see.

Chapter 10: Topic-specific results in Digg

The Digg data is crucial for testing hypotheses about what topics people find interesting. All search terms are associated with two numbers: the total number of stories identified, and the average score (number of Diggs) per story. The results are indicated below. They cover a period from August 2006 through February 2007, as in the results above. The topics are sorted by average score. "Count" indicates the total number of stories submitted within the designated time period.

Topic	Score	Count
Giuliani	10.01	243
Abu Ghraib	8.22	49
John Edwards	7.67	189
Barack Obama	7.24	348
Hillary Clinton	6.29	291
Mitt Romney	5.36	72
New Orleans	4.95	340
Harry Potter	4.84	511
Tiger Woods	4.13	150
Paris Hilton	3.71	891
Anna Nicole Smith	3.62	632
Britney Spears	3.41	1238

Table 5: Priorities found in Digg topics

Several facts are immediately clear when we view the data in this format. Our three "celebrity" subjects all have the highest story counts and the lowest average scores. The large "Count" fields indicate that many people found stories about those people that they considered were worth submitting. However, the scores tell a different story. It appears that, while many people submit stories about Hilton, Smith, and Spears, the general Digg population does not regard these stories favorably. This supports the hypothesis that news readers do not prefer to read stories about fluff topics; whereas they do tend to favor stories about politicians and other more weighty news subjects.

Rudy Giuliani stories received the highest average score, 10.01. This includes only data through February. When expanding the data range to include stories in all times, the preference for the "Giuliani" topic becomes even more pronounced, with Giuliani stories gaining an average score of over 18.

Thus, we might be tempted to assume that Giuliani must be the most popular candidate. However, if we look at the Digg page with the all-time highest rated stories on Giuliani[7], we see a very different story.

- Mr Giuliani Please Stop Mentioning 9/11
- Rudy Giuliani Constitutionally Ineligible To Be President
- Anger at Giuliani 9/11 fundraiser "$9.11 for Rudy" in poor taste
- America's Worst Nightmare: President Giuliani
- Giuliani: "For Me Every Day Is An Anniversary Of Sept. 11" GET OFF IT!
- Rudy Giuliani: "Freedom is Slavery"
- Rudy Giuliani's daughter is supporting Barack Obama
- DIGG this! Soldier to Giuliani: Have you done your foreign policy homework?
- Reporter Arrested on Orders of Giuliani Press Secretary
- Giuliani Closed Off Streets to Avoid 9/11 Victims' Families

Notably, these stories are all negative. There is not a single friendly news story in the first page of Giuliani links. (It bears mentioning that these stories did not all occur in or before February; they are taken from the set of all Digg data and include stories up to early October 2007.) One interpretation of this result is that what people really want from the news is more stories that speak badly of Rudy Giuliani. Many top stories regarding other politicians are similarly negative, but no other candidates generate nearly as much negative interest as Giuliani does.

[7] http://digg.com/search?s=giuliani&submit=Search&section=all&type=title&area=all&sort=most
Visited 10/02/2007

Chapter 11: Correlating Results with Individual Media Sources

I used the Java API from the WEKA package to analyze the final results. My Java program has a class called FinalResults which is responsible for reading significant information out of the database and converting it to an arff file, which is the format used by WEKA.

The format of the arff file for clusters is expressed in this way:

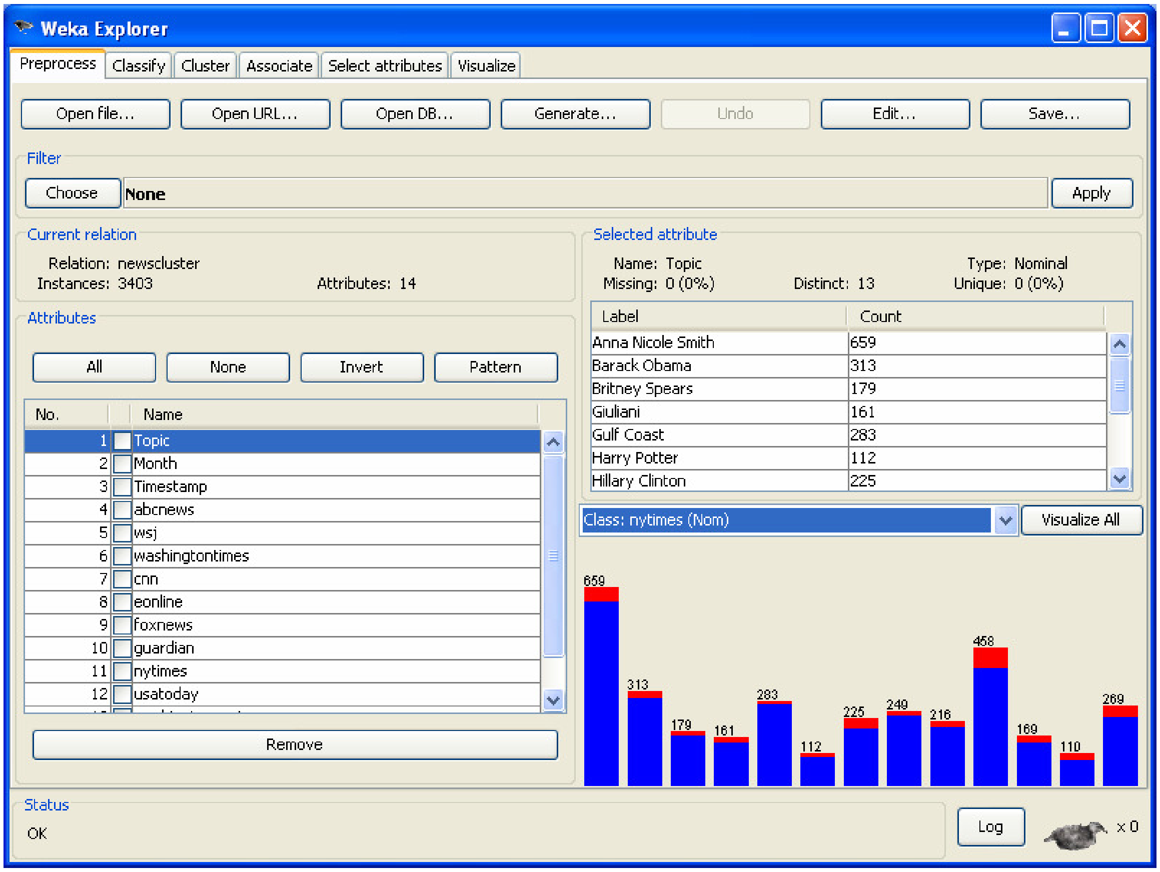
```
@RELATION newscluster
@ATTRIBUTE Topic {'Anna Nicole Smith','Barack Obama','Britney
Spears','Giuliani','Gulf Coast','Harry Potter','Hillary Clinton','John
Edwards','Mitt Romney','New Orleans','Paris Hilton','Rupert
Murdoch','Tiger Woods'}
@ATTRIBUTE Month Numeric
@ATTRIBUTE Timestamp DATE "yyyy-MM-dd HH:mm:ss"
@ATTRIBUTE abcnews {0,1}
@ATTRIBUTE bbc {0,1}
@ATTRIBUTE wsj {0,1}
@ATTRIBUTE washingtontimes {0,1}
@ATTRIBUTE cnn {0,1}
@ATTRIBUTE foxnews {0,1}
@ATTRIBUTE guardian {0,1}
@ATTRIBUTE nypost {0,1}
@ATTRIBUTE nytimes {0,1}
@ATTRIBUTE salon {0,1}
@ATTRIBUTE usatoday {0,1}
@ATTRIBUTE washingtonpost {0,1}
@ATTRIBUTE digg {0,1}
```

A row in this table takes this form:

```
"Hillary Clinton",3,"2006-03-02 00:00:00",1,0,0,0,0,0,0,1,0,0,0,0
```

In this example, we are looking at a cluster about Hillary Clinton, from month 3 (counting up from January 2006, this means March of 2006), which contains stories from both ABC News and the New York Times.

WEKA allows visualization of data attributes, which makes it immediately possible to inspect the data results and see graphs such as this one:



Filter: None

Relation: newscluster
Instances: 3403 Attributes: 14

Name: Topic Type: Nominal
Missing: 0 (0%) Distinct: 13 Unique: 0 (0%)

No.		Name
1	□	Topic
2	□	Month
3	□	Timestamp
4	□	abcnews
5	□	wsj
6	□	washingtontimes
7	□	cnn
8	□	eonline
9	□	foxnews
10	□	guardian
11	□	nytimes
12	□	usatoday

Label	Count
Anna Nicole Smith	659
Barack Obama	313
Britney Spears	179
Giuliani	161
Gulf Coast	283
Harry Potter	112
Hillary Clinton	225

Class: nytimes (Nom)

OK

Figure 3: Data visualization

By selecting the "Topic" line on the left it is possible to see bars representing the number of clusters covering each topic. After selecting "nytimes" as the class, the stories are highlighted in red (indicating clusters that include a story by the New York Times) and blue (indicating a cluster in which the New York Times did not participate.)

In the above figure you can see some general features of the data. An individual news source such as nytimes has a small overall presence in the clusters, but higher red bars indicate greater coverage. For instance, the cluster on the left represents stories about Anna Nicole Smith, and the red portion fills about 7% of the bar, indicating New York Times stories about Anna Nicole Smith. In this example, one of the highest proportions is given to the topic of New Orleans, the topic with 458 clusters. The New

York Times has a presence of 15% in stories about New Orleans. It is important to note that this graph comes from an early iteration of the data scan, and may not reflect the current final data.

Once the data is placed in this format, it is then transformed into a series of new .arff files, each of which contains the results from a single news source. Since, as mentioned above, the New York Times has a story in 7% of 648 clusters, the intermediate table will contain a row value of “nytimes”, “Anna Nicole Smith”, 7%, 648. In this way, a data file is generated for each site, and each file contains one row for every topic, revealing what percent of clusters have stories from the selected source.

In the final step, the data was normalized again. The reason this was done is that we are not merely interested in the overall presence or absence of a source in a topic; we are interested instead in the priority. If the New York Times is present in 15% of all clusters about New Orleans, but only 10% of clusters overall, then obviously New Orleans has a higher priority for nytimes than most clusters. Therefore, I normalized the values of all results from each site so that they fit into a range from 0 (lowest priority) to 1 (highest priority). These are inclusive values: Every row in the final table contains at least one “1” and at least one “0”.

Like the Google News sources, the Digg topics were also normalized to bring them within a range from 0 to 1, though these values were based on scores rather than story counts. As seen in Table 5, Britney Spears was the lowest scoring Digg topic and Giuliani was the highest. The final results of normalized topics across all sites are tabulated below.

Source	Abu Ghraib	Anna Nicole Smith	Barack Obama	Britney Spears	Giuliani	Harry Potter	Hillary Clinton	John Edwards	Mitt Romney	New Orleans	Paris Hilton	Tiger Woods
abcnews	0.42	0.00	0.35	0.00	0.74	0.40	0.51	1.00	0.51	0.33	0.17	0.03
bbc	0.42	0.00	0.22	0.68	0.31	0.71	0.01	0.02	0.55	0.18	0.23	1.00
cnn	0.23	0.00	0.43	0.30	0.66	0.00	0.72	0.72	1.00	0.02	0.09	0.07
foxnews	0.51	1.00	0.37	0.36	0.71	0.16	0.71	0.47	0.63	0.24	0.45	0.00
guardian	0.40	0.34	0.29	0.09	0.61	0.10	0.51	0.76	1.00	0.30	0.00	0.02
nypost	0.15	0.97	0.00	0.47	1.00	0.50	0.44	0.40	0.23	0.15	0.74	0.10
nytimes	0.63	0.00	0.55	0.09	1.00	0.00	0.98	0.10	0.30	0.30	0.16	0.27
salon	0.69	0.71	0.72	0.12	0.85	0.20	0.97	1.00	0.60	0.18	0.53	0.00
usatoday	0.00	1.00	0.50	0.15	0.25	0.16	0.29	0.10	0.71	0.93	0.27	0.22
washpost	0.42	0.13	0.94	0.19	0.87	0.20	1.00	0.72	0.99	0.38	0.00	0.11
washtimes	0.65	0.02	0.96	0.45	0.60	0.24	0.74	1.00	0.85	0.30	0.00	0.01
wsj	0.70	0.00	0.52	0.90	0.99	0.68	0.62	0.33	1.00	0.20	0.00	0.11
digg	0.73	0.03	0.58	0.00	1.00	0.22	0.44	0.65	0.30	0.23	0.04	0.11

Table 6: Final analysis

How do the various news sites stack up in their coverage of fluff news? As noted before, the three celebrity topics – Paris Hilton, Britney Spears, and Anna – were the lowest scoring Digg topics. Anna Nicole Smith received extensive coverage from Fox News, New York Post, and USA Today. Paris Hilton tended to be a generally low scoring topic, aside from receiving a priority of 0.74 from New York Post. Britney Spears also tended to score low, except in the Wall Street Journal (0.90) and the BBC (0.68).

As for presidential candidates, Giuliani does seem to compel some of the highest levels of coverage, receiving priorities of { 1, 1, .99, .87, .85, .74, .71 } from 7 of the 12 media sites. Hillary Clinton and John Edwards are also clear favorites of the media, each receiving a priority above 0.7 from 6 sites, and above 0.9 from 3 of the sites. However, Clinton's best coverage comes from "mainstream" news sites – Washington Post and New York Times; whereas John Edwards appears to do better with TV news and the right wing Washington Times. (This may have something to do with played-up stories about his expensive haircuts.)

As a final analytical step, I used WEKA's clustering tool to see which sources are most aligned with each other, and also to see which news sites are most closely aligned with Digg. I picked three as the number of clusters, and simple K-Means as the clustering algorithm. By this measurement, sites would be considered similar to each other based on the least mean sum of the squares for all priorities.

The clusters I got as a result were:

1. ABC News, Wall Street Journal, Washington Times, CNN, Guardian, Washington Post

   Key topics: John Edwards (0.76), Giuliani (0.75)
2. BBC, Fox News, New York Post, Salon, USA Today

   Key topics: Anna Nicole Smith (0.74), Giuliani (0.62)
3. Digg, New York Times

   Key topics: Giuliani (1.0), Hillary Clinton (0.71)

Tentatively, this indicates that the New York Times is closest to Digg in terms of delivering content that people are interested in reading. Of course, it is worth bearing in mind that Giuliani and Clinton are both New York politicians, so it is natural that a New York paper would cover them; it may just be a coincidence that those happen to be topics of special interest to Digg readers. The second cluster, containing both of the Murdoch run publications and the notoriously fluffy USA today, seem to take interest in Giuliani stories, but spend too much time on a topic like Anna Nicole Smith. Of course, these numbers do not indicate whether the coverage on a given topic is generally positive or negative, which is obviously an important consideration.

POTENTIAL OBJECTIONS TO METHODOLOGY

This study would not be complete without acknowledging and addressing some of the shortcomings of the methods used. I would like to present a few possible objections to the methodology, and consider future avenues for studying this topic.

Digg users do not represent the general public.

Even though I earlier raised the point that people responding to media surveys are a self-selected group, this may be still truer of Digg users. First of all, all Digg users have internet access, which implies that they are more likely be people of means than the entire population. Second, Digg users are active participants in an online community, which implies a higher level of engagement than the average TV watcher.

Although these are valid concerns, at worst we can say that the Digg community recommends a distinct demographic of the population. Considering that this is a demographic which is particularly involved in reading and discussing the news, it seems to me that these results would be significant to media marketers, even though they cannot be universally applied. In order to gather better information, it might help to collect web traffic statistics directly from major news sites, rather than relying on a voluntary scoring system.

It is not appropriate to normalize the data. The total number of news clusters / Digg submissions is a significant factor, yet the data selection disregards this factor.

Clearly, mine are not the only acceptable methods for breaking down the data. My reasoning in focusing on the presence of a news source in each news cluster is this: A news cluster indicates that *something has happened* with regard to a particular topic. The sample space I used, clusters sized between 25 and 100 stories, insures that the event is relatively noteworthy but not universally covered. The extent to which a media source chooses to cover or not cover this event definitely indicates their priorities.

As for the number of total stories per topic: We might suggest that the New York Times should have covered more stories about Blackwater in 2006, because Blackwater eventually proved to be a newsworthy topic. However, since almost no sources covered Blackwater in 2006, the fact that the New York Times did not have this foresight should not necessarily be considered an issue with their reporting.

In the future, it would be a relatively simple matter to come up with and apply more techniques for breaking down the data in various ways. The program development and data collection were the most time consuming phases; now that this groundwork is complete, many other analyses could be applied.

Your method relies too much on Google's algorithms. You are assuming that they have done their work correctly.

This is a fair criticism. Part of the point of this thesis was to investigate a method for gathering existing data, which can be applied in other work. Google functions as a "black box" in this experiment, and it is assumed that the search engine prioritizes stories in an appropriate way. However, Google may have hidden biases that color the data. We can probably leave the analysis of Google's bias up to future studies.

To address this concern, we might use Lexis-Nexis instead of Google. Or, future iterations of this project might require implementing a program that directly monitors news sites to obtain new stories. An advantage of this would be that a story's prominence within the site could be scanned and recorded, so front page stories could be distinguished from minor stories.

CONCLUSION

The introduction to this paper raised the possibility that media sources might give an undue amount of attention to sensationalist news. The analysis provided in this report appears to have confirmed this as the true situation. Data shows that news readers in general, as represented by Digg users, are distinctly uninterested in stories about celebrities and far more interested in coverage of more serious topics. Yet media sources consistently devote a significant amount of resources to covering sensationalism.

What are some possible explanations for the discrepancy? We can speculate on various reasons. Here are a few ideas.

1. **Reporting on celebrities is cheaper.** Serious journalism requires man power, travel expenses, and educated research. By contrast, celebrity news only requires a small number of dedicated photographers and interviewers, many of whom are paid for the material they produce but not salaried. Therefore, despite a desire for hard-hitting news among readers, perhaps it is more cost-effective to produce a lot of stories that generate minor interest.
2. **Serious journalism generates serious enemies.** Reporting serious news often involves coming into conflict with people who have power and money. Journalists often criticize multinational corporations which employ large legal teams, or politicians who have the power to pass punitive legislation. Writing about celebrities is less risky, even if they are wealthy celebrities.
3. **Fluff reporting is presented as a deliberate distraction.** This is the "bread and circuses" model of the media. Perhaps media conglomeration has put a

> large number of news corporations in the hands of a small group of owners with a specific political agenda, and this agenda runs contrary to responsible journalism. Under this theory, the media has hidden incentives not to report serious stories, and therefore misdirects the public with stories about celebrities and similar irrelevancies.

I am not a student of journalism, and so I cannot do much more than guess which answer is correct, if any. Nevertheless, I believe this research may help to demonstrate that consumers are being underserved in the delivery of serious news.

Appendix

This space is reserved for possible future appendices before the final submission. Appendix candidates may include include code samples, SQL queries, or more .arff file readouts.

Glossary

HTML

HyperText Markup Language. It is the predominant is the predominant markup language used in web pages.

DOM

Document-Object Model. Most programs that analyze HTML and XML documents rely on the creation of a well-formed tree of document elements, which can be analyzed for content and presentation layouts.

URL

Uniform Resource Locator. It is the predominant is the predominant markup language used in web pages.

SQL

Standard Query Language. Executing SQL statements is the most common method for searching and updating tables in a database.

WEKA

Waikato Environment for Knowledge Analysis. WEKA is a free software package used for analyzing data.

API

Application programming interface. An API is a source code interface that an operating system or library, such as WEKA, provides to support requests by computer programs.

CGI

Common Gateway Interface. CGI is a standard protocol for interfacing external application software with a web server.

References

The Daily Show. (June 11, 2007). “Paris Hilton Gets In a Car.” Video reproduction retrieved 11/1/2007, from http://www.spikedhumor.com/articles/110359/ The_Daily_Show_Paris_Hilton_Gets_in_a_Car.html

Severin, W. and Tankard, J. 2000. Communication Theories: Origins, Methods and Uses in the Mass Media. New York: Allyn & Bacon

Tan, P., Steinbach, M., and Kumar, V. 2005. Introduction to Data Mining. Boston: Pearson Education, Inc.

Witten, I. and Frank, E. 2005. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann Publishers

Niels Provos. (July 9, 2007) “The reason behind the ‘We're sorry...’ message.” Retrieved September 29, 2007, from http://googleonlinesecurity.blogspot.com/2007/07/ reason-behind-were-sorry-message.html

Tor: Anonymity Online. http://tor.eff.org/

Weinberger, D. 2007. Everything Is Miscellaneous: The Power of the New Digital Disorder. New York: Times Books

Mindich, D. 2005. Tuned Out” Why Americans Under 40 Don't Follow the News. Oxford University Press USA

Vita

Russell Glasser was born in Princeton Junction, New Jersey in 1974, to Alan and Sheryl Glasser. Russell graduated from Los Alamos High School in Los Alamos, NM, in 1992. He attended the University of California at San Diego from 1992 to 1997, obtaining a Bachelor of Science Degree in computer science in June of 1997. Russell has worked at a variety of software development firms as well as teaching classes at the computer learning center. He has been working in his current position as a software engineer at IBM since November of 2000.

Permanent address: 8121 Rimini Trail
Austin, TX
78729

This dissertation was typed by the author.