Patrick Polinski. Automated Monitoring of Online News Streams: Topic Detection and Tracking Considerations. A Master's paper for the M.S. in I.S. degree. November, 2003. 44 pages. Advisor: Robert M. Losee.

This paper describes the term frequency patterns found in online news summaries published over a seven-week period. The patterns are analyzed qualitatively and quantitatively to facilitate the refinement of algorithms used for the automatic detection and tracking of important topics appearing in streams of text. It is shown that a term's importance cannot be measured in raw frequency counts or significant increases in volume alone. The impact of these findings on existing algorithms is discussed, and new approaches for automated story detection and presentation are considered.

Headings:

Information retrieval

Information retrieval – filtering

Information retrieval – extraction

Really Simple Syndicate

Linguistics

AUTOMATED MONITORING OF ONLINE NEWS STREAMS: TOPIC DETECTION AND TRACKING CONSIDERATIONS

by
Patrick Polinski

**A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science**

Chapel Hill, North Carolina

November, 2003

Approved by:

_____

Advisor

1. Introduction

A recent study by the School of Information Management and Systems at the University of California at Berkeley estimated that 5 exabytes of new information was produced in 2002, enough to fill the Library of Congress 500,000 times (Lyman and Varian, 2003). The study found a 30% increase in stored data since 1999, with digital data marked by the greatest increase. The effect of this explosion in new content is often described as “information overload,” as it has become increasingly difficult to distill the meaningful information from the meaningless. Information overload, accompanied by reluctance amongst information seekers to spend time learning to use tools designed for information management, results in “information underutilization”, as potentially valuable information is never evaluated (Rao, 2002).

In this light, it is not surprising that techniques for the automated detection of the significant features within information streams have received increasing attention within the information and computer science research communities; notable efforts include the new information detection (NID) research conducted at Johns Hopkins University’s Center for Speech and Language Processing (Allan et al., 1999) and the topic detection and tracking (TDT) research conducted under the guidance of the National Institute of Standards and Technologies (NIST) (NIST Speech Group, 2003) . With news stories serving as the input, systems have been designed to extract the most important or most novel events and to monitor changes to these events over time. The results have been promising. A topic segmentation system built as part of the TDT research identified the

stories represented in a stream of 16,000 news articles (with topic boundaries removed) with 80% precision and 80% recall (Yamron, Carp, Gillick, Lowe, & van Mulbregt, 1998). A system designed by Swan and Allan (1999) identified topics and timeframes for most of the major stories appearing in a 175 days worth of news articles.

While Swan and Allan's system -- relying on a simple statistical model to detect significant changes in feature frequency -- was promising, it did miss some key stories throughout the coverage period. Explaining why some of the top stories were missed, the researchers noted:

> Stories 3 and 9 were covered in our corpus, but the coverage was spread out over several days (Story 9 was written about in one story per day for five consecutive days), and there was never a single day where the coverage appeared as significant. Stories that were covered well in the corpus but missed by our system were stories 10, 15, and 17. These stories were missed because the features that were distinctive about them (Haiti, Taiwan) were frequently in the news, and the occurrence of those features was not that different than their occurrence on any other day (Swan and Allan, 1999, p.44).

It appears a better understanding of feature distributions over time is required for more effective story detection models. Accordingly, the primary goal of this research is to examine the types of term frequency patterns that appear within a corpus containing many weeks worth of news articles. Graphical depictions of term frequencies are used to help answer the following questions:

- Which criteria must be met for story to be considered important? Which types of distributions meet the criteria?
- Are there interesting distributions that do not meet the requirement for an important story?

- What are the benefits and challenges of detection systems relying on raw frequency comparisons?
- What are the benefits and challenges of detection systems based on changes in term frequencies?

While formal recommendations of statistical models are not the intent of this research, some general approaches for detecting stories are considered.

The term frequencies were gathered from a collection of 50-days worth of Really Simple Syndicate (RSS) news feeds. RSS is a format enabling the syndication of web content through a publish and subscribe architecture. As a collection RSS documents is not the type of news corpus traditionally used in topic detection and tracking research, a secondary goal of this study is to introduce RSS, provide a rationale for its use in this type of research, and to present a methodology for RSS corpus development.

In Section 2, related work in automatic topic detection is reviewed in more detail. Section 3 provides an overview of RSS implementation today and describes an RSS collection procedure. In Section 4, selected term frequency distributions are described, and the impact of the different types of distributions on story detection approaches is considered. Conclusions are presented in Section 5, while Section 6 addresses areas of future work, with a focus on the story description task.

2. Related Work

This project is closely related to the Topic Detection and Tracking (TDT) body of research conducted through the NIST Speech Group. Sponsored by DARPA -- and with researcher teams from University of Massachusetts at Amherst, Carnegie Mellon University, and Dragon Systems -- the broad goal of TDT is to develop “algorithms for

discovering and threading together topically related material in streams of data such as newswire and broadcast news" (NIST Speech Group, 2003). While TDT research targets both text and audio streams, much of the work has been performed using a corpus of 16,000 news stories – in the form of text transcripts – run by the Reuters newswire and CNN broadcast news from July 1, 1994 through June 30, 1995. TDT research seeks to facilitate three main tasks: topic segmentation, topic detection, and topic tracking.

2.1 Topic segmentation

Topic segmentation is required for documents comprised of multiple topics with no clear topic boundaries, such as the audio recording of a television news program. The segmentation task detects topical shifts and divides the document accordingly. Reynar (1998) provides a nice review of the topic segmentation methods, including algorithms based on the first uses of terms (Youmans, 1991), term frequency weights (Richmond, Smith & Amitay, 1997), term repetition visualizations (Helfman, 1994), and vector space models (Hearst, 1994).

Within the TDT community, Yamron, et al. (1998) has approached topic segmentation by training a hidden Markov model (HMM) on the TDT corpus. While the TDT corpus comes pre-segmented, Yamron et al. removed story and paragraph boundaries in order to test their model. While the TDT corpus contained 15,863, the HMM algorithm discovered 16,139 topic boundaries. There were 10,625 exact matches resulting in 67% recall and 65.8% precision; given a cushion of 50 words, recall rose to 81.9% and precision jumped to 80.5%. More recently, Utiyama and Isahara (1999) proposed a comparable, domain independent model (not requiring supervised learning).

2.2 Topic detection and tracking

Topic segmentation is traditionally used to find the topical boundaries within a single, static document or continuous streams of text without metadata indicating topic demarcation (metadata). Topic detection and tracking research is focused on organizing streams of segmented stories into broader stories. It involves grouping incoming articles into topic bins and identifying when existing stories end and new ones emerge. Accordingly, the number of stories and the timeframes associated with stories evolve as new articles arrive.

Topic detection is related closely to text summarization research. For a corpus containing thousands of articles, grouping articles alone does not add much value, as the amount of information surpasses levels of human consumption. Rather, only the most important or novel events within an article should be collected. For example, Allan, Gupta, and Khandelwal (2001) proposed a model for extracting temporal summaries of the TDT corpus by extracting a single sentence from each event within the story, based on algorithms designed to detect sentences marked by usefulness and novelty.

Summarization is less critical when working with RSS feeds, which typically *are* summaries. Organizing these summaries along topic lines is the key task, and the approaches considered in this research are inspired by the work of Swan and Allan (1999) on time varying feature extraction. Swan and Allan employed a simple statistical model to extract the interesting and novel events from an incoming stream of textual news, where a stream is defined as “a collection of tokens arriving in a fixed order with each token having a time stamp” (p. 38). This approach was attractive due to its focus on time-based organization; key stories were extracted from 6683 news articles spanning 175

days in 1995, and the system was able to detect both the content of the key stories and the time periods involved.

In Swan and Allan's system, document frequencies (the number of documents containing the specified term) were calculated for the features (noun phrases and named entities) contained in the corpus. For each day, they determined a feature's significance by considering its theoretical or expected frequency versus its observed frequency. To determine when the frequency for a term on a specified day was not likely due to chance, chi-square tests were applied to each feature, based on a 2x2 contingency table comprised of a) the number of documents containing the feature on the specified date, b) the number of documents not containing the feature on the date, c) the number of documents containing the feature for the other days in the study, and d) the number of documents not containing the feature for the other days in the study.

Allan and Swan considered the feature's frequency for the day to be beyond the realm of chance when its chi square value is greater than 7.879 ($p<.005$). For example, the sample of the X^2 values for the term "Oklahoma City" appeared as follows:

April 17, 1995	**18th**	**19th**	**20th**	**21st**
1.38	2.31	1.10	617.96	170.49
22nd	**23rd**	**24th**	**25th**	**26th**
208.85	49.04	81.06	112.82	128.33
27th	**28th**	**29th**	**30th**	**May 1st**
95.01	83.85	21.11	7.26	.58
2nd				
17.79				

In this sample, the appearance of Oklahoma City was significant from April 20 through April 29 -- following the tragic bombing of the Federal building on April 19 -- and this block was considered one continuous story. Oklahoma City became significant again,

following a two-day break, on May 2nd. (Swan and Allan considered this second period of significance to represent a different story.)

To identify which of the significant terms are representative of the same story, periods of significance were compared amongst the terms. If at least one day's overlap was apparent between two terms, a second chi-square test was applied to determine if the terms are related. This chi-square test was based on a 2x2 contingency table containing a) the number of documents containing both features, b) the number of documents containing feature one but not feature two, c) the number of documents containing feature two but not feature one, and d) the number of documents containing neither feature.

If the chi square value for any pair of terms with overlapping dates is greater than 7.879 ($p<.005$), the terms were grouped together. For example, Oklahoma City became part of the "Oklahoma City Bombing" story, grouped with: 'oklahoma', 'bombing', 'building', 'mcveigh', 'fbi', 'john doe', 'doe', 'timothy mcveigh', 'timothy', 'city'…

While the results of this system were promising, some of the major stories in the TDT corpus were not detected. The statistical model was designed to detect a specific type of term distribution, which was not the distribution associated with every important story. A stronger understanding of the types of feature distributions, and their impact on statistical approaches, is the goal of this research.

3. Corpus Development

This study requires a collection of news articles over a significant time-period, with each article accompanied by a time stamp. The standard corpus used for this type of research is comprised of news articles. Transcripts from CNN broadcast news and the Reuters newswire, for example, were used in the TDT studies. The news articles used

here come from a less conventional and arguably more appropriate source: a collection of RSS news feeds.

3.1 Why RSS?

3.1.1. Standard for online news publishing

RSS is an XML application (vocabulary) enabling the distribution of web content. RSS is typically used to publish headline information, a brief description of the actual content, and a link to the full content. Those interested in a site's content can subscribe to its RSS file, called a feed, which is automatically received when the new content is added. (RSS feeds are gathered and viewed in "newsreader" or "news aggregator" software packages.) Accordingly, RSS is considered a syndication technology – an easy way to distribute and receive summaries of new content. RSS has been described as "a distributable 'What's New'" for web publishers (King, n.d.), while newsreaders have been characterized as "TiVos for the web" (Dickerson, 2003).

The initial RSS format emerged in 1997, under the direction of UserLand Software's Dave Winer, in association with Netscape. The format of RSS has since become the source of considerable debate and four different versions of RSS are currently in use (Festa, 2003). The debate has been contentious to the point where there is even disagreement over the meaning of the RSS acronym. Depending on the community, RSS stands for Really Simple Syndication, Rich Site Summary or RDF Site Summary. A group is currently meeting with the goal of ending the debate through the establishment of a new syndication standard, most likely to evolve as an alternative to RSS than a fifth version.[1]

[1] See http://www.intertwingly.net/wiki/pie/FrontPage for information on the emerging standard.

3.1.2 Popularity

Despite the splintered nature of the RSS format, the adoption of RSS has not been hampered. RSS has been most widely employed by two communities: 1) content/news companies such as the BBC, Forbes, and Wired and 2) individuals maintaining web logs, or "blogs" – journal-like web sites with posted content organized by time. Popular "bloggers" include technology innovators, political pundits, and popular culture devotees.

Based on the users of Syndic8 (www.syndic8.com), a web site that maintains a searchable directory of RSS feeds, the top ten most popular feeds (shown in Table 1) are indeed a mix of main stream news providers and individual voices, with a general slant towards technology-related issues.[2]

Table 1. Popular RSS feeds

Feed Source	Feed URL	Description
Book Reviews	http://p.moreover.com/cgi-local/page?index_bookreviews+rss	Book reviews from major newspapers.
SlashDot	http://slashdot.org/slashdot.rdf	One of the earliest technology focused web logs.
Content Syndication and RSS – the Blog	http://rss.benhammersley.com/index.rdf	Site is down.
NPR News	http://www.newsisfree.com/HPE/xml/feeds/49/2449.xml	NPR news stories.
Wired News	http://www.wired.com/news_drop/netcenter/netcenter.rdf	Wired News stories.
Stephen Hebditch > Blog	http://www.hebditch.org/index.rdf	Commentary on content management.
Boing Boing Boing	http://boingboing.net/rss.xml	"The blog of wonderful things", with a slant towards technology issues.
CNN Top News	http://www.newsisfree.com/HPE/xml/feeds/15/2315.xml	CNN news stories.
USA Today: Newswire	http://www.newsisfree.com/HPE/xml/feeds/42/1842.xml	USA Today news stories.
Developer Shed	http://www.devshed.com/devshednews.rdf	The "open source web development site."

Source: Syndic8 (http://www.syndic8.com/boxpage.php?Box=ViewedFeeds&N=200)

[2] Based on the number of times the feed has been viewed by Syndic8 user

Not all of the popular RSS-enabled blogs have a hard-core technology focus. Table 2 describes a sample of syndicated, non-technology blogs based on Technorati's (www.technorati.com) listing of top 100 blogs.[3]

Table 2. Popular, non-technology blogs offering RSS feeds

Blog Name (ranking)	Incoming Links	Site URL	Description
InstaPundit.Com	3,664	http://www.instapundit.com	A law professor writing about politics.
MetaFilter	2,961	http://www.metafilter.com	A web log in which a community of users post comments and links on a wide variety of topics.
girlwithagun	2,194	http://www.livejournal.com/users/girlwithagun	"Confessions of an altruistic misanthrope."
Daily Kos	1,800	http://www.dailykos.com	"Political analysis and other daily rants on the state of the nation."
Wil Wheaton Dot Net	1,706	http://www.wilwheaton.net	A slice of the actor's ("Stand By Me", "Star Trek: The Next Generation") life.

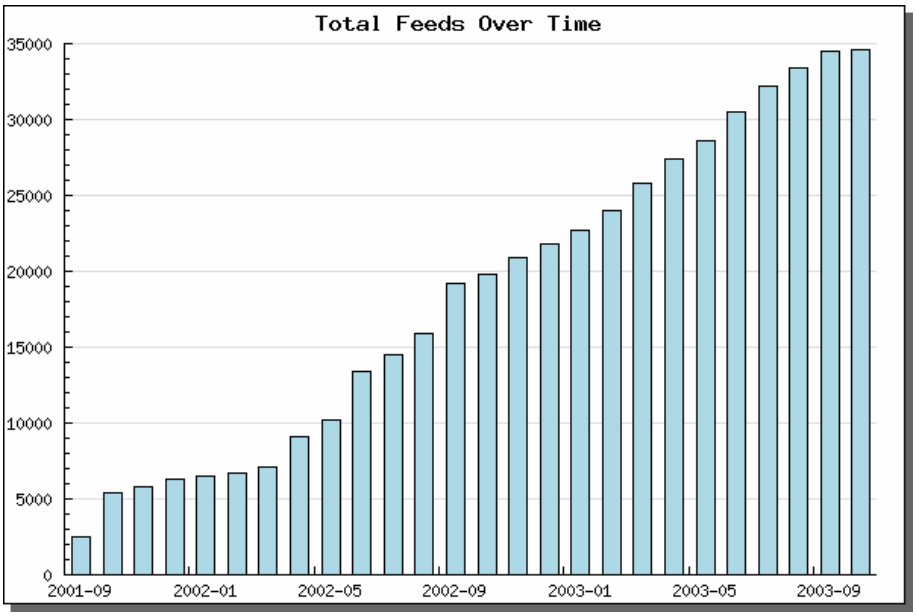
Source: Technorati (http://www.technorati.com/cosmos/top100.html)

Syndic8 provides a variety of other site-related statistics that offer insight into the nature of RSS adoption. As of October 2003, Syndic8's directory included over 16,000 active feeds, with thousands more feeds listed as dead (3,890), rejected (4,119), awaiting repair (3,317), or desiring syndication (2,097). Figure 1 depicts the growth of the total number of feeds added to Syndic8 over the last two years.[4] Figure 2 displays the growth in the number of registered Syndic8 users.

[3] Technorati's rankings are based on the number of incoming links from other web logs. While this is not an RSS feed-related measurement, it is likely that the number of people subscribing to a site's RSS content is relative to the number of people linking to the site.

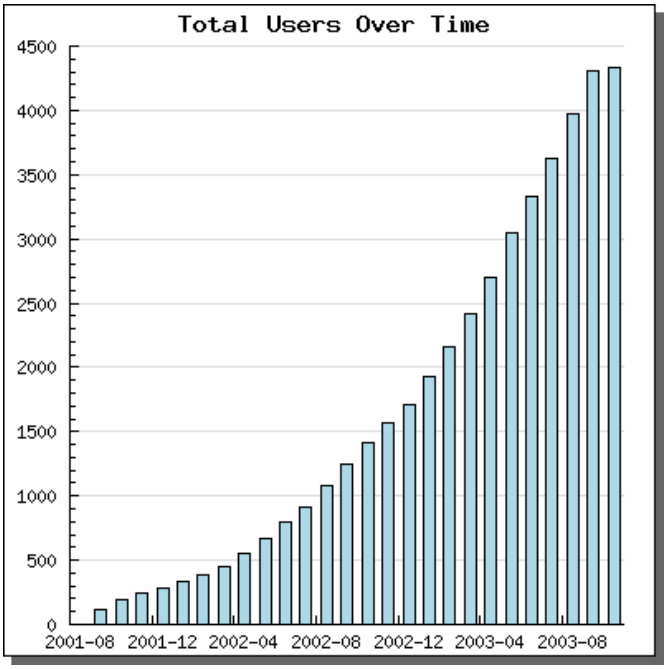
[4] Feeds are added based on user recommendations. All suggestions are reviewed before being added.

Figure 1. Feeds added to the Syndic8 directory



Source: Syndic8 (www.syndic8.com)

Figure 2. Registered users of Syndic8



Source: Syndic8 (www.syndic8.com)

3.1.3. Future Needs

The increasing popularity of RSS stems largely from the efficient information gathering it affords. The benefits of RSS, from both the publisher and subscriber perspective, are captured in the sentiments of O'Reilly Network's President and CEO Dale Dougherty. Speaking about O'Reilly's Meerkat service, which collects and presents technology related RSS feeds, Daugherty notes:

> What interests me about RSS is the ability to begin to monitor the flow of new information on the net. We all know what sites exist; what we really want to know is how often sites generate new information. As a writer and editor, I thought Meerkat would be valuable to watch what was happening in different technical communities. What I especially like about RSS and looking at feeds from hundreds of sites is that you can see the Web work at a grassroots level. I thought that Meerkat is the kind of tool I'd want to keep track of what is going on. We realized that this wasn't just useful to editors but to anyone who wants to be able to respond to new information.
>
> I'm not sure where Meerkat will take us, but it feels like it's opening up a remarkable new view of the Web. We'd really like to see more and more sites become RSS-enabled. RSS can do for them what Yahoo did for them in 1994, which is drive traffic by letting others know what you are doing. The difference is now we can notify others not just of a new site, but of new stories -- new activity on our site."

Chad Dickerson, the CTO of InfoWorld has also embraced RSS:

> In an age of spam and cold calls, this is just what the information-overload doctor ordered…Over the past few years, the Web itself has become like a blabbering acquaintance with a million fleeting and unconnected ideas, and e-mail has become a crowded cocktail party with a few interesting people whose words are obscured by the gaggle of others frantically trying to sell various unmentionables. With more and more traditional media companies supporting RSS every day and the unmediated voices of thought leaders such as Ray Ozzie and Tim Bray coming through my newsreader via RSS-enabled Weblogs, using my newsreader is like having a cocktail party for busy people where the conversation is lively and almost always to the point.

But does the RSS architecture really alleviate information overload? It clearly reduces the time spent *searching* for information. But this results in a mass of information that still needs to be *assessed*. With RSS becoming more ubiquitous, the information of interest can easily outpace the capacity for human consumption. Meerkat's channels, while focused specifically on technology news, still account for over a 1000 feeds a day -- far more than can be monitored by a single person. The technical writer Russ Lipton expresses the problem from the user's perspective:

> Ideally, I would like to subscribe to hundreds of Internet sources. Unfortunately, this imposes some performance constraints… as well as information management challenges. I don't yet have good tools for filtering, sorting, extracting and managing my subscriptions. Without those tools, I reach a limiting point where checking through my subscriptions is almost as tedious as proactively going to websites used to be. In a nutshell, we still need.... 'smarter' ways to tell the news aggregator which items within a feed are likely to be of most interest to me so they announce themselves appropriately when they arrive" (2002).

In light of Lipton's comments, the goals of TDT research appear to be highly congruent with the needs of RSS subscribers. Considering that RSS *is* the format for digital news exchange, it is an ideal corpus for TDT research.

3.2 Data Collection

There are two possible approaches for collecting a large body of RSS feeds.[5] The easiest method is to query a database of archived feeds, assuming such a database can be found for the content of interest. A second option -- manual collection via a news reader/aggregator -- allows the researcher to collect feeds from a handpicked set of sources. The downside here, however, is the time required.

[5] Regardless of the approach, the goal is to collect the RSS feeds in their raw (XML) format versus HTML representations; as XML, topic boundaries are easily recognized and any pre-processing requirements can be easily handled through Extensible Stylesheet Language Transformations (XSLT).

Given the time constraints associated with this study, a bulk download from an RSS archive is the desired approach. Unfortunately, there are very few publicly available feed archives. The aforementioned Meerkat Open Wire News Service, which maintains an archive of feeds going back sixty days, is perhaps the best example. Meerkat is accessible via the WWW (http://www.oreillynet.com/meerkat/) and enables users to filter the feeds based on source, category, time frame and a key word search. Meerkat also includes an API, allowing users to perform more advanced queries and to control the format of the output (HTML, RSS 1.0, RSS0.91, or a generic XML format). Unfortunately, Meerkat does not allow for bulk downloads of its stored feeds, but rather returns fifty feeds per page.

As an alternative, a number of RSS newsreaders were investigated to determine if collection from scratch provided a better data collection solution. Few newsreaders, however, offered archiving functionality, and those with this capability were buggy.

NewIsFree (www.newsisfree.com), a web-based RSS aggregator ultimately emerged as the best data collection tool. NewsIsFree (NIF) was selected because it maintains a three-week archive of its feeds. Given time constraints, an extra three weeks worth of the desired data was helpful. As was the case with Meerkat, the archived data could not be downloaded in bulk, as it was displayed one hundred stories at a time, per source, and in HTML format. Feeds were collected, then, one hundred at a time, using the Save As…Text File option of Internet Explorer. While this low-tech approach could have been used to collect feeds from Meerkat as well, NIF was chosen because of the normalized format of its stories. Because NIF “scrapes” its feeds – that is, converts the HTML from other sites into RSS -- they appear in a more standardized manner. The

Meerkat feeds, on the other hand, are produced by hundreds of different providers and, accordingly, take on many different shapes and sizes (some feeds have a esoteric title only). Given this study's emphasis on term frequency, feeds with a standardized "look and feel" were desirable.

3.3 Selection Criteria

Among the 6700 news channels built into NIF, sports-related feeds were selected for a number of reasons. For one, sports offer a nice mix of expected and unexpected events. Temporally speaking, many "high-level" sports events are expected, providing a built-in framework for evaluation. That is, in terms of media coverage, the World Series can be expected to dominate October, while the Super Bowl is featured in January, and the NCAA Basketball tournament monopolizes March.

Sports also provide a rich, dynamic set of unexpected events, ideal for a study of frequency distributions. Many major sports stories evolve with no past precedent and with no relationship to the seasonal (or temporal) expectations. Even sports competitions that do occur when expected are still fluid and generally unpredictable at the lower level. While some figures within sports tend to receive consistent attention, new actors and situations are always emerging. NCAA basketball is typically the major story in March, but the teams, players, and coaches making the news changes with each round of the tournament, and surprises often abound.

3.4 Selected Feeds

The final corpus contained sports-related feeds published over a 50-day span, from August 24th through October 14th, 2003. Th. It is a good time period to analyze sport's

news, as September and October are generally considered the greatest months for sports, highlighted, in order, by:

- US Open tennis finals
- The start of the NFL and NCAA football seasons
- The MLB pennant and wildcard races
- The MLB playoffs
- The start of the NHL season
- The start of the NBA season

Of the nearly 500 feed channels listed in NIF's sports category, 6 were selected for use. Channel selection was based on two criteria: 1) a national focus (as opposed to a city or region) and 2) the use of a headline and a description. The second criteria operated as the primary filter as the RSS feeds for many of the major sport's news providers (e.g. Sports Illustrated and ESPN) included headline information only. The six sources used to create the corpus are listed in Table 3. 9,173 feeds were collected over the 50-day period. Each feed contained a headline for the news story along with a brief description, with an average of 28 words per feed. The first sentence or two of the story served as the description for the USA Today, LA Times, New York Times, FaceOff and AllSports feeds. For MSNBC, the sub-headline was used as the description. Bylines appear in the LA Times and Allsports feeds, while FaceOff and USA Today use datelines. While bylines and datelines could have an impact on feature frequencies, in particular by over-inflating geographical locations, they were not removed.

Table 3. RSS sources used in this study

Source	URL	Example Text (Raw feed is in encoded in XML)
USA Today: Sports	http://www.usatoday.com/sports/briefs.htm	10/07/03 15:50 Ramirez, Red Sox advance OAKLAND —Manny Ramirez broke out of his postseason slump with a three-run homer, and former closer Derek Lowe reverted to his old role as the Boston Red Sox advanced to the American League Championshi..
MSNBC	http://www.msnbc.com/	10/07/03 07:34 Oh Manny! Red Sox hold off A's Ramirez ends slump, homers to lift Boston into ALCS
LA Times	http://www.latimes.com/sports/	10/07/03 09:08 Red Sox Appear Off-Curse Ramirez's three-run homer is the key hit as Boston completes a comeback from a 2-0 division series deficit with a 4-3 victory over Oakland in Game 5. (By Thomas Bonk)
New York Times	http://www.nytimes.com/pages/sports/index.html	10/07/03 06:48 After Collision, Course for Red Sox Is Bronx With key performances from two of their stars, Pedro Martinez and Manny Ramirez, and a gutty save by Derek Lowe, the Red Sox moved on to face the Yankees.
All Sports	http://www.allsports.com	10/09/03 05:04 ALCS Preview - Boston vs New York It's time to "Cowboy Up" again for the Boston Red Sox, as they get ready for the fiercest rivalry in baseball by taking on the New YorkYankees in the American League Championship Series. (Gary Dibert)
FaceOff	http://www.canada.com/sports/index.html	10/07/03 06:46 Red Sox beat back 'curse' to defeat Oakland 4-3 and advance to ALCS OAKLAND, Calif. (AP) - Pedro Martinez, Johnny Damon, Manny Ramirez and even that maligned Boston bullpen - they all were tougher than any curse.

3.5 Content Analysis

3.5.1. Pre-Processing

To prepare the corpus for analysis, some pre-processing was required. The news stories, initially grouped by source, were grouped by date. The corpus was then segmented into noun and noun phrases using the Phraser software package.[6] No normalization or stemming was performed on the noun phrases, although the concordance software used for content analysis (see 3.5.2) removed punctuation, such that "McNabb's" would be separated into "McNabb" and "s".

3.5.2. Feature Selection

A concordance software package called System Quirk[7] was used to determine feature (noun and noun phrase) frequencies and to identify the features (noun and noun phrases) to be included in the study. To determine the single terms, the entire corpus was added to System Quirk, and a word list was generated with no inclusion or exclusion (stop word) lists specified. The post-Phraser corpus consisted of 13,553 distinct terms. Terms appearing less than 10 times or greater than 1000 times were removed, resulting in 2408 nouns. These cut-off points were arbitrary. All nouns with frequencies over 1000 fell into the category of traditional stop words ('the", "an", "of"). Nouns appearing less than ten times throughout the course of fifty days were considered unlikely to be representative of important events. However, this cut-off could result in some noteworthy stories being missed – especially if all nine occurrences appeared in a single day. As a

[6] Available at: http://www.scientificpsychic.com/.
[7] Available at: http://www.mcs.surrey.ac.uk/SystemQ/.

final filter, numbers and dates were removed, resulting in the final list of 2,316 single nouns, sorted by frequency.

The selection of noun phrases was more complicated. While System Quirk does not have a built-in phrase frequency function, phrases can be entered into its inclusion list. To create the list, the original lists of feeds were run through the Phraser software again, this time with the options to "Decompose Phrases into Components" and to "Isoloate components separated by conjunctions and prepositions" turned on. For example, the story:

COLUMBUS, Ohio The attorney for suspended Ohio State tailback Maurice Clarett met with NFL executives Monday for about an hour in Washington to discuss whether he will be able to enter the 2004 draft...

is decomposed into:

- COLUMBUS
- Ohio
- attorney
- suspended Ohio State tailback Maurice Clarett
- suspended Ohio State tailback Maurice
- Ohio State tailback Maurice Clarett
- suspended Ohio State tailback
- Ohio State tailback Maurice
- State tailback Maurice Clarett
- suspended Ohio State
- Ohio State tailback
- State tailback Maurice
- tailback Maurice Clarett
- suspended Ohio
- Ohio State
- State tailback
- tailback Maurice
- Maurice Clarett
- suspended
- Ohio
- State
- tailback
- Maurice
- Clarett
- NFL executives Monday
- NFL executives
- executives Monday
- NFL
- executives
- Monday
- Washington
- able
- draft

The list was then reduced to two-word phrases, as instances of these phrases would be sufficiently descriptive and likely to occur most frequently. The two-word phrase list was

de-duped and then applied as an inclusion list against the full corpus. Any phrases appearing more than once were selected, resulting in a list of 1335 phrases.

3.5.3. Feature Frequencies Per Day

In this study, frequency numbers reflect the total number of times a feature (single noun or phrase) appears on a given day, not the number of documents containing the feature. To generate the frequency counts, the single noun and phrase lists were added to the System Quirk as inclusion lists and applied to each day's worth of news feeds.

4. Results

4.1 Feature Analysis

The frequencies are studied to answer three main questions:

1. Which features (nouns and noun phrases) are representative of important events or stories?
2. How can these features be detected?
3. How do you detect the topic boundaries within a feature?

The answers to these questions are dependent upon how frequency patterns are interpreted. Do the distributions indicate a single definition of an important story or are there "flavors" of important stories? Is story detection simply a matter of raw frequency counts, where the term with the highest number wins? Or should a term's frequency for a day be considered in light of its frequency on other days (such as Swan and Allan's consideration of the probability of a term's density). In other words, does the detection of frequency "spikes" effectively identify the key events? Finally, does the approach (inter-

feature raw frequency comparisons versus measurements of intra-feature distribution changes) depend upon the type of feature?

To help address these questions, consider figures 3-8, showing the frequency distributions of six features. To help place these distributions in context of the overall feature set, note that the average frequency is .97 occurrences per day, with 11 occurrences representing the 99th percentile and 25 representing the 99.9th percentile.

Figure 3. Feature A

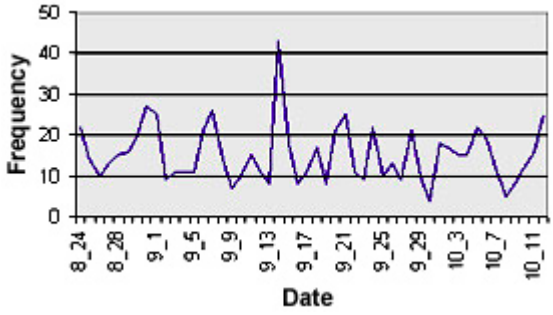


Figure 4. Feature B

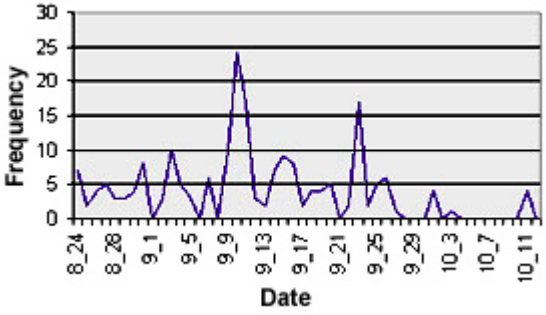


Figure 5. Feature C

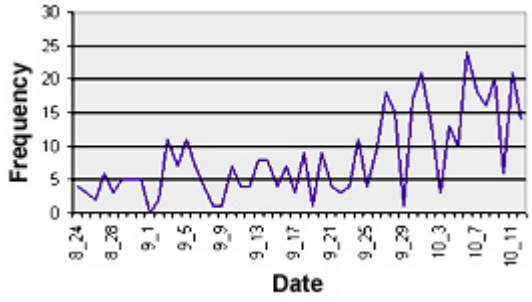


Figure 6. Feature D

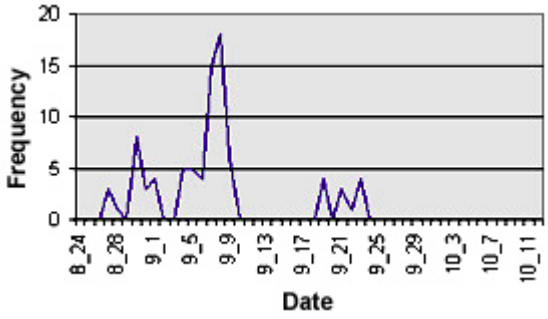


Figure 7. Feature E

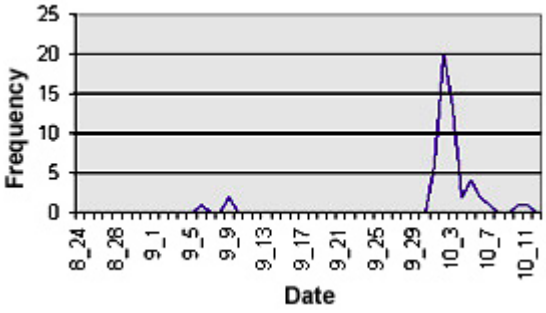
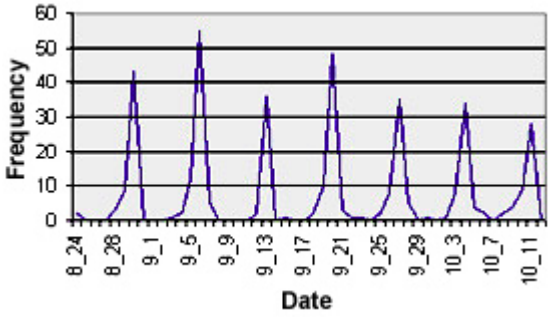


Figure 8. Feature F



While there are many other important patterns represented within the 2315 single terms and 1334 phrases, the distributions above highlight many of the challenges involved in story detection and deserve close inspection. Without knowledge of the actual features, is it clear which graphs represent the most significant stories? The features represented by Figure 3 and Figure 8 account for the most instances overall and are marked by some high-volume spikes. Figure 5 depicts a term that gradually builds in use, while figures 6 and 7 depict terms with significant use on a few days and with very little or no use on most days. A case could be made that an important story is evident in each graph shown above.

Secondly, within each graph, is it clear when the feature is noteworthy and when it is not? Each graph is marked by a series of spikes (except for Figure 7, which has a single spike only.) Is each spike representative of a distinct story, or are spikes separated by only one or two day's worth of relative inactivity part of the same story? Finally, should the spikes be the only area of focus when considering a term's potential importance?

To answer these questions, consider the features behind the graphs.

Feature A: 'victory'

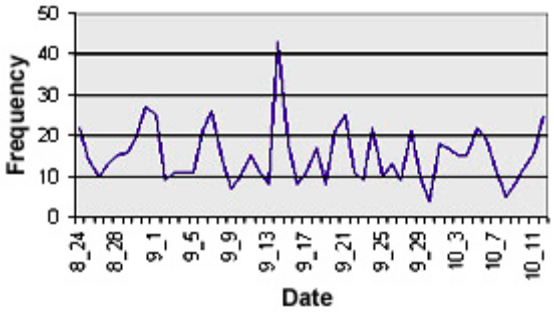
Is it important?

For a term to be representative of an important event, two criteria must be met. 1) The term must appear at significantly high levels, and 2) the term must be helpful as a story identifier. (At the story-level, it must have discrimination value[8].)

[8]Discrimination value is used in the indexing task of information retrieval to measure a term's ability to distinguish the document in which it is contained from other documents. Very high frequency terms are considered to have negative discrimination value, so importance decisions based on high frequency *and* high discrimination may seem misguided. However, given that the analyzed corpus contains nouns only, and that the most frequent and least frequent nouns were not included, the terms being considered are best described as having medium frequency. It does reason, however, that the greater the frequency, the greater the likelihood that the term is a poor discriminator.

Considering the frequency benchmarks previously provided, from a raw frequency perspective, ‘victory’ (shown again in Figure 9) meets the first criteria nearly every day. In fact, throughout the entire 50-day period, ‘victory’ appears more often than only four other terms – ‘season’, ‘over’, ‘first’ and ‘night’.

Figure 9. ‘victory’



But does ‘victory’ offer enough discrimination value to be considered important? The likely answer is no. Terms that appear at consistently high levels over an extended timeframe are almost certain to fall into one of two categories: 1) common terms, useful in describing many events or 2) very significant terms during the time frame covered. Of course, ‘victory’ falls into the first category, as it is a generic sports term associated with many distinct stories.

How should it be detected?

The better question here is whether it should be detected. Given its low discrimination value, ‘victory’ may belong on a stop word list, with traditional stop words (those serving grammatical purposes only like ‘the’, ‘an’, and ‘of’) and generic sports terms (‘game’, ‘season’, ‘win’). Considering that ‘victory’ does not appear at steady levels--its distribution is marked by a series of spikes, with a major spike on September 14--it appears that the feature does mean more on some days than others, unlike traditional stop words like ‘the’. Accordingly, on some days – and hopefully on days in which victory is integral to a story, such as a player or coach recording a record-setting number of victories – it may be a helpful or even critical descriptor for the story.

(If Swan and Allan's feature grouping approach to story generation is used, for example, the extraction of 'victory' could be helpful.)

If 'victory' were considered for extraction, a detection system searching for significant changes in a term's frequency would be the best approach. A poor choice for detection would be an algorithm based on the comparison of raw frequency counts, in which case 'victory' would be returned as important nearly every day.

Closer inspection indicates that the frequency spikes associated with 'victory' are a function of time of week, appearing most often on Sunday-- the day after the busy Saturday sport's schedule. Based on a review of the articles containing the feature on 9/14, the major spike appears to be a function of an exceptionally busy schedule the day before rather than an individual event receiving extraordinary coverage. A system that considered the relationship between the frequency of this term and the day of the week could prove beneficial. For this feature, it may be more useful to ignore the large weekend spikes in favor of more moderate (but nevertheless rarer) weekday spikes.

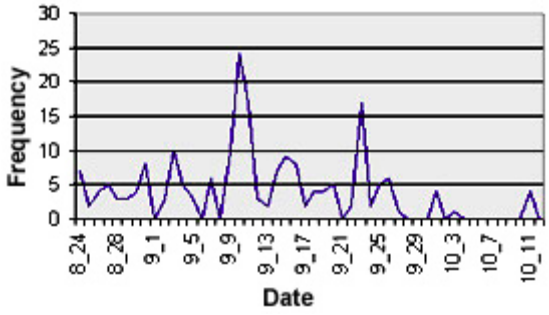
Where are the story boundaries?

With knowledge of its low discrimination value and relationship to the time of the week, victory would only be useful on relatively large weekend spikes or on weekday spikes, such as the 22 occurrences on Wednesday, September 24 (see Figure 9). There would be little reason to believe that any stories identified by this feature would last longer than the spike in volume.

Feature 2: 'Clarett'

Is it important?

Figure 10. 'Clarett'



The chart for 'Clarett'[9] (shown again in Figure 10) appears very similar to 'victory', in particular in its spikiness. Based on the graphical similarities to 'victory' and the relatively consistently high frequencies – with appearances on 34 out of the 50 days and as the 91st most frequently appearing term (out of 2316) – the feature's discrimination value could be assumed to be low. In this case, this assumption would be incorrect, as 'Clarett' represents the actions of a single actor, and, accordingly has high discrimination value. The combination of high discrimination and high frequency clearly makes 'Clarett' an important feature.

How should it be detected?

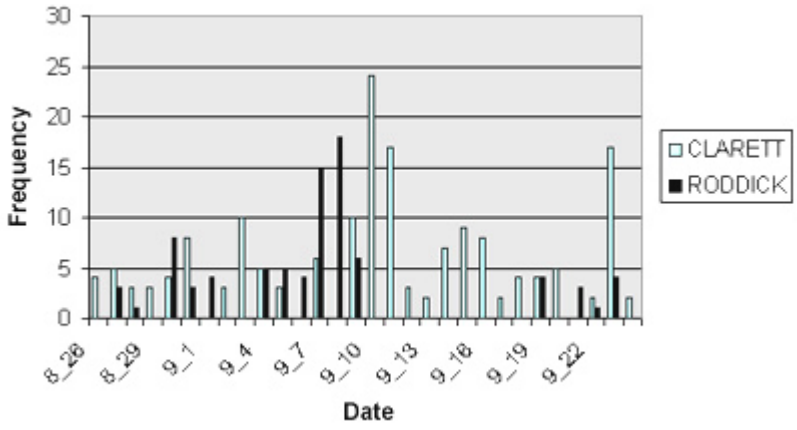
Despite the similarities, 'Clarett' and 'victory' are best treated by different models. Obviously, 'Clarett' should not be considered a stop word. Extracting the feature only on its extraordinarily high frequency days (relative to its other days) is also a poor option. With this approach, 'Clarett' would likely only be flagged as significant during its major spikes (September 10 and September 23) and the significance of these spikes relative to

[9] Maurice Clarett was a high-profile freshman tailback for Ohio State University in 2002. His appearance in the news in the timeframe covered in this study stems from many off-field incidents, including an accusation of academic impropriety and a charge for misdemeanor falsification, regarding items he reported stolen from his car. The media closely followed the University's response (which included a one year suspension delivered on 9/10) and Clarett's reaction, which included a lawsuit against the NFL (reported on 9/23), which bars anyone under the age of 20 from entering the draft.

the spikes of other features (such as feature D, shown in Figure 6) could be tempered given the relative high volume on other days.

In the case of 'Clarett', comparisons of its raw frequency to the raw frequency of other terms would work best. Why raw frequencies? Throughout the course of the fifty day period, 'Clarett' appears more times (200) than any other last name, with 'Bryant' (as in Kobe Bryant[10]) in second (167). From the period of August 24 through September 9, *before* the two major spikes in 'Clarett' frequency, 'Clarett' is the second most frequently occurring last name, sandwiched between 'Agassi' and 'Roddick', two high-profile American tennis players playing in the high profile U.S. Open. A measure of intra-frequency change would certainly flag 'Agassi' and 'Roddick' as significant prior to September 10, as they are rarely active afterward, but such approaches would likely miss 'Clarett', which is effectively penalized for receiving even more coverage on other days (see Figure 11). Of course, 'Clarett' will not compare well to other terms with low discrimination value like 'victory', so the efficacy of the raw numbers approach is dependent upon the system's ability to ignore the generic terms.

Figure 11. 'Clarett' v. 'Roddick'



[10] The media followed Kobe Bryant's sexual assault charge closely.

Where are the story boundaries?

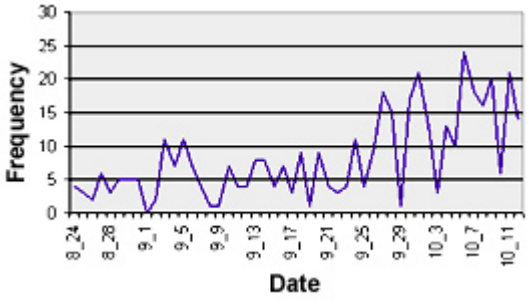
Properly detecting the topical boundaries associated with ‘Clarett’ is a difficult task. Do spikes in frequency represent new stories, or are they indicative of new developments within a broader story? With knowledge that ‘Clarett’ represents a surname (in particular a unique surname, unlike ‘Smith’) and with the assumption that it is unlikely that an individual appears in the news on consecutive days for unrelated reasons, the graph is perhaps best interpreted as a single story marked by twists, turns, peaks, and valleys. That ‘Clarett’ appears in the off-season, without the structure of a built-in schedule, is another argument for a single-story interpretation.

Feature 3: ‘Cubs’

Is it important?

The feature ‘Cubs’[11] (shown again in Figure 12) shares many of the same traits as ‘victory’ and ‘Clarett’, most notably in that it appears on the majority of days. As was the case with ‘victory’ and ‘Clarett’, the high frequency of ‘Cubs’ offers some evidence that it could be a generic sports term. The “ramping effect” evident in the distribution does set it apart from the other features, but does not shed much light on the nature of the term. For example, the graph could be indicative of a team reaching the playoffs after a late season push or could be representative of a generic term like ‘goal’ (see Figure 13),

Figure 12. ‘Cubs’

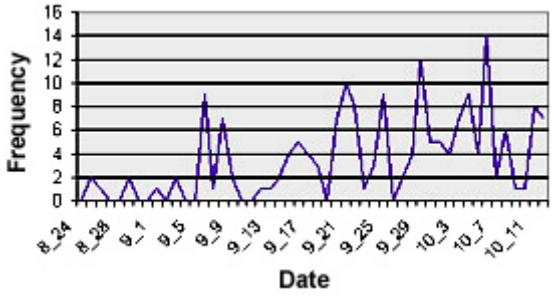


[11] The media coverage of the Chicago Cubs focused on their push for a playoff birth and their participation in the playoffs.

which increases in frequency as news sports using the term begin their seasons. In this case, the former represents the actual story, making ‘Cubs’ an important feature.

How should it be detected?

Figure 13. ‘goal’



The discrimination value for ‘Cubs’, which is an umbrella for many players, coaches, and fans, falls somewhere in between the low discrimination power of ‘victory’ and the high value of ‘Clarett’. Its value is likely closer to ‘Clarett’, however, and, like ‘Clarett’, raw frequency counts are likely a better gauge of significance than statistics designed to detect exceptionally high frequencies only. For example, if the latter approach is used, the period from September 13 through September 26 will appear as relatively uneventful. In truth, the feature ranks 48th overall in this time-period, and third amongst terms with clear discrimination value. (Top ranking terms are generic: ‘over’, ‘night’, ‘first’, ‘season’, ‘victory’, ‘win’, etc.)

Where are the story boundaries?

It is reasonable to expect a larger entity like the Chicago Cubs to produce more stories than individual entities like Maurice Clarett. That the Cubs are appearing in-season, with different pitchers going each night and playing different teams each week, also lends credence to a multi-story interpretation. But what level of granularity is appropriate? The articles prior to September 27 describe a collection of important cub’s games as they fight to make the playoffs. The September 27 spike reflects the playoff clinching game, the September 30 to October 5 volume reflects the first playoff series,

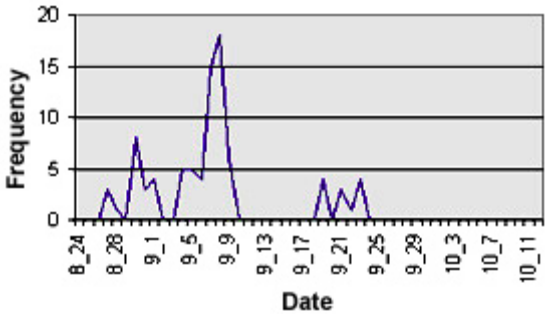
while the October 6 to October 12 volume reflects the second playoff series. These events could easily be merged into one broad story or could be treated as four separate stories.

Feature 4: 'Roddick'

Is it important?

The distribution of 'Roddick'[12] (shown again in Figure 14) is much easier to interpret than the preceding three. This feature meets the high frequency criteria on a small subset of days. Its low frequency on most days raises the likelihood that the term demonstrates specificity. Accordingly, it meets the criteria for importance on the days in which its volume is sufficiently high relative to the other terms in the corpus. Given that 'Roddick' is the second most frequent surname on September 7 and the most frequent surname on September 8, its importance is unmistakable.

Figure 14. 'Roddick'



How should it be detected?

In this case, either an algorithm designed to detect significant changes in the feature's distribution or a raw frequency comparison (with other features) approach would work. With 'Clarett' and 'Cubs', the former approach could be problematic as these features are still important on their non-spike days. This is not the case with 'Roddick'.

[12] The coverage of Andy Roddick centers around the U.S. Open tennis tournament, which he won on 9/7/2003. The smaller story beginning on 9/19 reflects Andy Roddick's participation in the Davis Cup tournament.

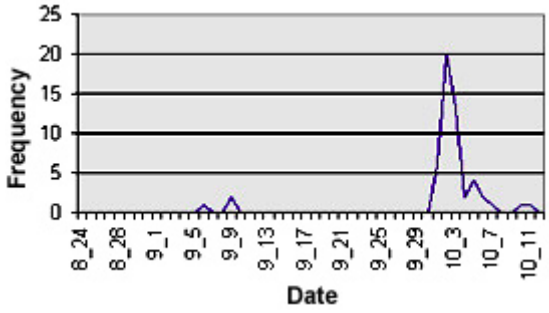
Where are the story boundaries?

If each spike in term frequency were interpreted as a story, ‘Roddick’ would be associated with one big story, one very big story, and three or four little stories. As mentioned in the discussion of Maurice Clarett’s story boundaries, it seems unlikely that an individual would be in the news in consecutive days for unrelated reasons, even when allowing for one or two day’s worth of inactivity. There is also (limited) evidence that a natural boundary for a term like ‘Roddick’ involves many day’s worth of inactivity, as the two periods of activity are separated on both sides by stretches of zero frequency.

Figure 5: ‘Limbaugh’

Is it important?

Figure 15. ‘Limbaugh’

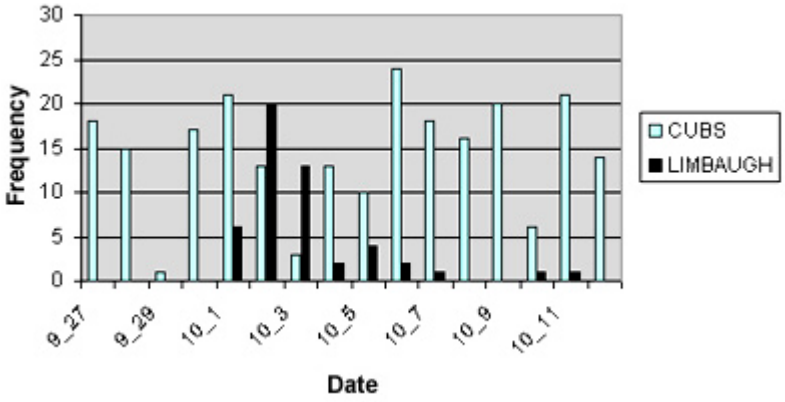


Like ‘Roddick’, the interpretation of the distribution of ‘Limbaugh’[13] (shown again in Figure 15) is straightforward. There is one major spike, which is significant given the term does not appear at all throughout the vast majority of the 50-day period.

An interesting question is whether Limbaugh should be considered more important than a feature that appears on a more consistent basis, while sharing a similar peak. For example, compare the graphs for ‘Cubs’ and ‘Limbaugh’ (see Figure 16).

[13] Rush Limbaugh received media coverage due to the racial implications of his comments about quarterback Donovan McNabb.

Figure 16. ‘Cubs’ v. ‘Limbaugh’



Given their nearly identical spikes, should the significance of ‘Cubs’ on October 1 be considered equivalent to the significance of ‘Limbaugh’ on October 2? Given its steadily increasing volume, the high frequency of ‘Cubs’ on October 1 is not a surprise. On the other hand, ‘Limbaugh’, with no build-up to its October 2 frequency appears to be unexpected. Should a rapid rise following a long stretch of zero occurrences make a feature more or less significant than peak associated with a steadier incline?

How should it be detected?

Like ‘Roddick’, this feature could be accurately detected by either intra-feature change analysis or inter-feature raw number comparisons. The former approach would likely report ‘Limbaugh’ as being a more significant event than the one represented by ‘Cubs’ on the prior day, while the latter approach would consider these features as nearly equal (with a slight edge given to ‘Cubs’).

Where are the story boundaries?

Aside from a minor blip in early September, all of the occurrences are concentrated in a period from October 1 to October 12. Like ‘Roddick’, story boundaries appear to be represented by many days of zero frequency. So, while the story clearly declines in

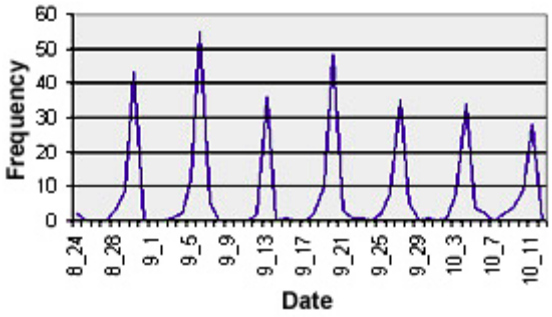
significance after October 3, it should probably not be considered completely dead until a period of inactivity beyond one or two days is witnessed.

Figure 6: 'Friday'

Is it important?

Based on its distribution (shown again in Figure 17) 'Friday' meets the high-frequency requirement, with higher peaks than any of the features discussed thus far. The feature appears to provide discrimination value as well, as it does not appear every day. However, it does appear in a very predictable manner. Not surprisingly, each of the major increases in this term's frequency falls on a Saturday, when the news from the previous day is recounted. The feature, then, is not important, as it does not help identify the topic of an important story, but merely describes when many different stories occur.

Figure 17. 'Friday'



How should it be detected?

The best approach is to treat 'Friday' as a stop word. Both raw number comparisons and measures of the distribution's change would be likely to incorrectly classify the feature as important. In fact, this scenario was encountered by Swan and Allan, who acknowledged "[e]very day of the week except Wednesday appeared as a significant term at some point in the analysis, even though these terms offer no clues as to what the story is about" (1999, p. 42). A feature's frequency, then, may be statistically significant for reasons beyond the topics covered in the news.

5. Conclusions

RSS provides a rich collection of topical terms and appears to be a suitable input for topic detection and tracking processes. Topic detection is clearly not a simple task, however, as it has been shown that feature importance is not a reflection of raw frequency counts or frequency spikes alone. Of the six features studied, in fact, the graphs with the highest volume and the tallest spikes are associated with the most insignificant terms (‘victory’ and ‘Friday’, respectively). To measure importance and to effectively determine story boundaries, other questions should be considered:

- Is the feature’s discrimination value high, moderate or low?
- Is the feature representative of a scheduled or unscheduled event?
- Is the feature a topical discriminator or temporal discriminator? Or, more broadly, could a frequency spike be explained by reasons other than topical significance?

Unfortunately, answers to these questions are not easily extrapolated from the shape of a feature’s distribution. It has been shown, for example, that the distributions for features with low discrimination can appear very similar to features with high discrimination power. It should be noted, however, that these graphs would likely diverge given more week’s worth of data. For example, ‘Cubs’ and ‘goal’ share similar distributions across the 50-day span, but only because this span coincided with the critical stretch of the baseball season. In December, with the Chicago Cubs no longer playing games and with ‘goal’ still highly applicable to the hockey, football, and basketball games being played, the charts for the features would look dramatically different.

Its been demonstrated that both inter-frequency comparisons (of raw numbers) and measures of intra-frequency change have strengths and weaknesses. The former approach is more meritocratic. The feature with the highest frequency wins, regardless of whether the feature's count is relatively low or not significantly high compared to its frequencies on other days. This seems correct: the person or team in the news the most on a given day is very likely the most important person or team for the day.

There are a number of exceptions, however. If terms with low discrimination, such as 'win', 'night', and 'season', cannot be identified and filtered, they will be selected as big stories nearly every day. Similarly, terms with moderate discrimination (like 'Cubs', which may be associated with multiple stories) may trump features specific to a single story. Finally, this approach may unfairly weight in-seasonal stories, where a percentage of media coverage is rote. (Of five stories on the Chicago Cubs, two may be nothing more than mechanical, box-score type wrap-ups of the previous night's game).

While expecting an extraction system to make discrimination judgments and detect expected and unexpected events is overly ambitious, some practical steps would improve the efficacy of a system based on raw number comparisons:

- Employ a large, domain-specific stop word list. The 50 days worth of data collected in this study alone provides good evidence of the sports terms marked by high frequency and low discrimination.
- Use two or three word phrases instead of single nouns. Phrases offer stronger discrimination, differentiating 'John *Smith'* from 'Steve *Smith*' and the 'New York *Giants'* from 'San Francisco *Giants'*. In this study, features with high discrimination ranked much higher on the phrase list than on the single word list.

For example, ‘Maurice Clarett’ was the 41st most frequent phrase, while ‘Clarett’ was ranked 91st among single nouns.

Approaches designed to detect when a feature’s frequency is abnormally high, such as the probability tests employed by Swan and Allan, also provide a solution to the interference caused by generic, high frequency features. Because these approaches are not based on raw numbers, terms appearing with great frequency are only extracted when the frequency on a given day is significantly higher than on other days.

The downside of this approach, though, is its potential to miss the important features that occur with consistent frequency. Given the similarities in the ‘victory’ and ‘Clarett’ distributions, an approach designed to ignore ‘victory’ on all but its most significant days would almost ignore ‘Clarett” on many important days. As mentioned earlier, time will certainly improve the efficacy of this approach. With more data comes a better understanding of the theoretical frequency of a term. Just as ‘goal’ will remain in the news when the ‘Cubs’ does not, ‘victory’ will continue to appear with consistency long after the Maurice Clarett saga has ended. With many more days of data, the ‘Clarett’ distribution will look much more like “Limbaugh” than ‘victory’, and frequency probability approaches will prove more effective. Keep in mind, however, that even with 175 days worth of data, Swan and Allan’s system missed stories “because the features that were distinctive about them… were frequently in the news, and the occurrence of those features on that specific day was not that different than their occurrence on any other day” (1999, p.44).

6. Future Work

6.1 Algorithms

The research implies that more sophisticated algorithms are required to ensure that important stories are not missed and insignificant stories are not selected. A system, for example, that compares the frequencies of similar entities only, could prove helpful. While it is difficult to determine the expected frequency for a term, the expected frequencies (and discrimination value) for terms belonging to the same class (baseball teams, football teams, tennis players, etc.) could be assumed equal[14]. In this scenario, importance could be determined based on significant deviations upwards from the class average. If this approach is pursued, automated clustering of terms will need to be added to the primary TDT tasks. Research will also be required to determine if such an approach would scale to the general news environment, where classes of terms are not so obvious.

6.2 Interface

While much of TDT research is still at the algorithm level, creating software that automatically discovers important stories and presents them in a human-friendly manner is the ultimate goal. Two interfaces have been demonstrated. Swan and Allan (2000) designed a simple, but effective automatic timeline generator while Frey, Gupta, Khandelwahl, Lavrenko, Leuski, and Allan (2001) developed a more sophisticated system based on the Lighthouse information retrieval system (Leuski & Allan, 2000).

[14] Many believe there is an east-coast bias in sports coverage, which would have to be studied and quantified.

More research into how the information extracted by TDT systems should be displayed is needed, however.

In particular, interfaces that grant users a high level of control over the type of information extracted, including tools for content and "story type" filtering, should be developed. For a technology publisher such as O'Reilly's Dale Daugherty, the concepts gaining momentum within the cutting-edge, grass roots blogging community could be the inspiration for the next book. Accordingly, an interface enabling the detection of upwardly ramping term distributions (such as the distribution for 'cubs') would be beneficial. On the other hand, for the intelligence community monitoring suspicious activity (including hidden messages sent through the media), isolated spikes in frequency (such as 'Limbaugh') or spikes appearing with unexplained predictability (a 'Friday' spike on Wednesday) may be of interest. Other interactive features supporting user-defined filtering, such as a 'stop tracking this term' option, would also be helpful, especially if automatic detection of stop words proved impossible.

6.3 Story description

Providing story descriptions is a key TDT task, and was not fully addressed in this paper. Single nouns or noun phrases can indicate the actor in a story, the location of a story or the name of an event, but do not describe a story in sufficient detail to be useful to a human. An important story, for example, would not be reported as "Cubs", but perhaps "Chicago Cubs defeat the Atlanta Braves to win first playoff series since 1908".

Swan and Allan's approach to story generation --the grouping of terms statistically determined to be dependent – has some clear weaknesses. The named entities groupings were accurate but not descriptive, while the noun groupings were descriptive, but noisy.

Perhaps the largest weakness of this approach, however, is its inability to indicate how the story's information changes over its period of significance. For any story spanning several days, the key descriptive terms on the first day of the story are likely to be quite different than the key descriptive terms on the last day. This type of evolution is not evident in stories described through a list of key terms. This weakness is acknowledged by Frey et al., where it is noted that "[b]etter approaches might generate a summary from the multiple documents or summarize the changes from the previous day" (2001, p. 353).

The nature of RSS feeds--in particular that they serve as temporal summaries to begin with--affords some powerful alternatives to story description that deserve greater attention. For example, instead of using a probability-based approach to group terms, the full-text of the feeds containing the feature of interest could be returned (if there are only a handful of occurrences) or summarized (if there are many occurrences). For example, the 6 most frequently appearing nouns in the feeds containing the key feature could be reported for each day the feature is considered important, as demonstrated for 'U.S. Open', on 8/28 and 9/7.

U.S. Open – 8/28	U.S. Open – 9/7
1. Davenport	1. Roddick
2. Clijsters	2. Henin-Hardenne
3. Wednesday	3. Final
4. Federer	4. Ferrero
5. Williams	5. Clijsters
6. Henin-Hardenne	6. Agassi

This approach is advantageous as it ensures that there are no spurious terms included in the summary, and it demonstrates how the important pieces of a story change over time. It doesn't, however, provide sufficient context. There is no indication of the relationship between the story's participants (did Roddick beat Agassi?) and the term

‘tennis’ does not appear. Methods for summarizing the context for the key features of a story would be helpful.

6.4 Other research questions

The effect of more heterogeneous RSS input…

The RSS feeds used in this study were prepared in a standardized manner, with each feed including a descriptive news headline along with a one to three sentence introduction or summary. Because there are no enforceable rules for the amount of information included in RSS feeds, formatting within the broader RSS feed population is variable. Among the feeds gathered by Meerkat, some offer a very limited description of the content, such as single word title and no description. On the other hand, there are feeds that offer very detailed titles and very rich descriptions. (In fact, some feeds include the full text of the article they are “summarizing”.) Accordingly, detection based on term frequency --or even document frequency-- can be problematic, as this number is influenced directly by the richness of the feed. It is possible that, given enough feeds, the inconsistencies in feed production will cancel each other out. The effects of heterogeneous RSS input, however, must be tested.

The effect of metadata…

The story extraction system adopted in this paper assumes that the time stamp is the only metadata available. With RSS feeds, however, title, description, and link information is tagged. The emerging syndication standard is expected to be fully extensible, allowing users to add their own metadata without hampering interoperability. Accordingly, metadata-aware statistical models should be explored.

How does sports news compare to other types of news…

The extent to which the findings in this paper are generalizable to other types of news is not clear. In particular, the highly scheduled nature of sports news appears to result in "spiky" frequency distributions, which may not be as prevalent in a general news corpus.

The relationships between stories…

Just as terms can be grouped together to form stories, stories can often be grouped into broader stories or themes. For example, wars in Afghanistan and Iraq serve as distinct stories, but they can also be considered parts of a bigger story, perhaps best labeled the "war on terror." Or consider Frank DeFord's (2002) list of the top ten sports stories of 2002 (see Table 4). His top three stories are not specific events, but high-level themes extracted from a collection of important stories. Arguably this type of "theme extraction" is best left to humans, and the actual benefits to information seekers are difficult to enumerate. However, this type of meta-story generation offers an interesting challenge for researchers in artificial intelligence.

Table 4. Frank Deford's top three sports stories of 2002

Story	Description
"Debased Ball"	From steroids to a strike threat, from the All-Star Game fiasco to the lowest-rated World Series ever, baseball has a brutal year.
"Net Gain"	Thanks to growing enthusiasm for MLS and a strong World Cup showing, U.S. soccer's popularity is on the rise.
"Cinderella Stories"	The Patriots, Hurricanes, Nets and Angels are surprise winners in the "big four" sports.

Source: http://sportsillustrated.cnn.com/features/2002/year/essays/deford/

References

Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lavrenko, V., et al. (1999). Topic Based Novelty Detection 1999 Summer Workshop at CLSP Final Report. Retrieved November 4, 2003, from http://www.clsp.jhu.edu/ws99/projects/tdt/final_report/report.pdf.

Allan, V., Gupta, R. & Khandelwal, V. (2001). Temporal Summaries of News Topics. *Proceedings of SIGIR*, 10-18. Retrieved October 25, 2003 from http://www-ciir.cs.umass.edu/~allan/Papers/2001-sigir.pdf.

Deford, F. (2002). Big dogs and underdogs. *Sports Illustrated*. Retrieved September 5, 2003, from http://sportsillustrated.cnn.com/features/2002/year/essays/deford/.

Dickerson, Chad. (2003, July 3). RSS Killed the Infoglut Star. *InfoWorld*. Retrieved August 31, 2003, from http://www.infoworld.com/article/03/07/03/26OPconnection_1.html.

Festa, Paul. (2003, August 4). Blogs locked in a bitter battle. *ZDNet*. Retrieved August 5, 2003, from http://zdnet.com.com/2100-1104-5059458.html.

Frey, D., Gupta, R., Khandelwal, V., Lavrenko, V., Leuski, A., & Allan, J. (2001) Monitoring the News: a TDT demonstration system. *Proceedings of the Human Language Technology Conference (HLT)*, pp. 351-355.

Hearst, M.A., (1994). Multi-paragraph segmentation of expository text. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 9-16. Retrieved November 4, 2003 from http://www.sims.berkeley.edu/~hearst/papers/tiling-acl94/acl94.html.

Helfman, J.I. (1994). Similarity patterns in language. *IEEE Symposium on Visual Languages*, pp.173-175. Retrieved November 4, 2003 from http://citeseer.nj.nec.com/helfman94similarity.html.

King, A.B. (n.d.). Introduction to RSS. *WebReference*. Retrieved August 31, 2003, from http://www.webreference.com/authoring/languages/xml/rss/intro/.

Leuski, A. & Allan, J. (2000). Lighthouse: Showing the Way to Relevant Information. *Proceedings of the IEEE Symposium on Information Visualization, IEEE Computer Society*, pp. 125-130.

Lipton, R. (2002). What is Publish and Subscribe? Retrieved September 15, 2003, from http://radio.weblogs.com/0107019/stories/2002/02/25/whatIsPublishAndSubscribe.html.

Lyman, P. & Varian, H.R. (2003).How much information. Retrieved November 4, 2003, from http://www.sims.berkeley.edu/research/projects/how-much-info-2003/.

NIST Speech Group. (2003). Topic detection and tracking (TDT). Retrieved November 4, 2003, from http://www.itl.nist.gov/iad/894.01/tests/tdt/.

Rao, R. (2002). Rich interaction with content. Retrieved November 4, 2003, from http://www.ramanarao.com/articles/2002-09-richtech.pdf

Reynar, J.C. (1998). Topic Segmentation: Algorithms and Applications. PhD. Thesis, University of Pennsylvania, Department of Computer Science. Retrieved October 27, 2003, from http://www.cis.upenn.edu/~jcreynar/research.html.

Richmond, K., Smith, A. & Amitay, E. (1997). Detecting subject boundaries within text: A language independent statistical approach. *Exploratory Methods in Natural Language Processing*, 47-54. Retrieved October 27, 2003, from http://acl.ldc.upenn.edu/W/W97/W97-0305.pdf.

Swan, R. & Allan, J. (1999). Extracting Significant Time Varying Features from Text. *Proceedings CIKM, ACM Press*, 38-45.

Swan. R. & Allan, J. (2000) Automatic Generation of Overview Timelines. *Proceedings of SIGIR*, pp. 49-56.

Utiyama, M. & Isahara, H. (1999). A statitical model for domain-independent text segmentation. *Proc. of the 39th Annual Meeting of the Association for Computational Linguistics*, 491-498. Retrieved November 4, 2003 from http://acl.ldc.upenn.edu/P/P01/P01-1064.pdf.

Yamron, J.P., Carp, I., Lowe, S., & van Mulbregt, P. (1998). A hidden Markov model approach to text segmentation and event tracking. *Proc. ofICASSP-9*8.

Youmans, G. (1991). A new tool for discourse analysis: the vocabulary management profile. *Language*, 67 (4), pp. 763-789.