Micro-Longitudinal Analysis of Web News Updates

Daniel O. Kutz and Susan C. Herring
School of Library and Information Science
Indiana University
Bloomington, Indiana
{dokutz, herring}@indiana.edu

Abstract

News sites on the World Wide Web pose challenges for information retrieval due to their dynamic content and a tendency to produce multiple versions of the same story. In this study, we advance a method we refer to as micro-longitudinal sampling that automates the capture and reduction of news site content at one-minute intervals to a manageable size for qualitative analysis. This method was used to mine text and images from the front pages of three major news sites over a three-week period, and the changes identified were manually analyzed using content analysis and critical linguistics methods. The results reveal multiple motivations for changes, only some of which, it is argued, need to be attended to in searching and archiving online news.

1. Introduction

News sites on the World Wide Web pose particular challenges for information retrieval. First, their content is dynamic: headlines, stories, and accompanying photographs are updated frequently [22], and some changes only persist for short amounts of time. This is because unlike in traditional news media such as newspapers and television, in which readers see only a final product which has undergone previous fact-checking and editing, news content on the Web is often first posted, and then refined once in the public eye. BBC News promises on its webpage that its news is "[u]pdated every minute of every day." However, we know little about what information in news sites actually changes, how often, and what motivates change (beyond, presumably, the imperative to keep news content "fresh").

A second challenge is related to the first: frequent changes often result in multiple versions of the same story, each containing slightly different information, even within a single site. These changes are not simply cumulative; information can be deleted as well as added, or modified to suggest a different interpretation of the reported events. What characteristics (other than chronological sequence) differentiate multiple versions, and which version best represents the story as a whole?

In order to address these questions, we advance a method we refer to as micro-longitudinal sampling that automates the capture and reduction of Web content at one-minute intervals to a manageable size for qualitative analysis. This method was used to mine text and images from the front pages of three major news sites representing different national cultures over a three-week period.

The findings reveal multiple motivations for updating site content, which we classify and analyze using methods of content analysis, and of critical linguistics, an approach traditionally applied to the identification of ideological bias in print newspapers. While evidence of ideologically-motivated change is found, it rarely introduces substantive new content. Moreover, a number of changes serve only to correct language mechanics and style, a phenomenon that is increasingly taking place in the public eye as online news producers hurry to make breaking news available to the public and thus to actualize the dynamic potential of the Web. Thus while frequent changes may give the appearance that a site is regularly providing new content, most such changes are non-substantive, and can effectively be overlooked for the purposes of retrieving information from news sites.

2. Literature Review

Information retrieval work concerned with news data has predominantly focused on text customization or summarization in an attempt to reduce the initial document set to a smaller or customized subset. For example, Makkonen et al. [21] categorize news based on shared events which are identified through co-occurrence topic detection and a developed ontology. Summary generation can reference pre-established categories [15], or attempt to develop new categories as current events get processed [1]. Other researchers use information retrieval methodologies to summarize data from multiple websites into one meta-site, allowing a user to go to one place to access news stories pulled from various sources [24, 26]. A popular user-centered approach is to employ machine learning to adapt the presentation of news to the specific needs of the reader [5, 18, 27]. One can also use clustering techniques to group documents; for example, Eichman and Srinivasan [10] examine various clustering heuristics to explore the categorization of news stories.

News media have also been studied from the perspective of media studies. According to Livingstone [19], media giants like CNN have become “media accelerants.” Online news, a genre which operates at an even faster pace than television news media, is known for its dynamic content [22]. In such media, there is even greater pressure in the newsroom to publish as soon as one is informed of a news event, a pressure that has carried over into print news media, which must now publish faster to keep up with online news [13]. The rapid acceleration and transmission of news also forces policy and decision makers to react quickly, and in some cases, unreflectively, to the media’s reports [19]. The risks of unthinking reactions are compounded by the findings of a number of researchers of bias in news reporting, despite the stated value of news providers on facts and objectivity [9].

News media bias has been investigated systematically by linguists in the Critical Linguistics tradition [12], who employ methods of language-focused analysis to reveal subtle bias in the presentation of news discourse, both in text and images. Bias in text can be encoded via word choice, metaphor, and syntactic structure—devices that might otherwise appear to be merely 'stylistic' to the unreflective observer [7, 8, 17]. Bias in images can be encoded in perspective, degree of closeness or distance to the subjects depicted, as well as by other means [3, 29]. Our methodology in the present study involves linguistic analysis, image analysis, and descriptive data gathering over a (short) longitudinal period of time.

3. Methodology

In an attempt to capture short-term changes to the content presented in webpages, our methodology retrieves data numerous times with a short time interval over successive visits. This differs from traditional information retrieval, where the focus is on capturing a large corpus of data (spanning a multitude of sites or documents) over a broad time period, where visits are separated by days or weeks. Instead, micro-longitudinal content analysis focuses on targeting a small number of individual pages (or sections of pages) for repeat visits that are separated by intervals of less than five minutes.

This kind of analysis facilitates understanding of webpage evolution and makes the editorial process of a website more transparent. In traditional print media such as books, magazines, and newspapers, the reader is confronted with a final product that has previously undergone fact-checking and editing. On the web, messages are often first sent and then refined once in the public eye. A micro-longitudinal study captures minute changes that may only persist for short amounts of time. By analyzing these small changes, one can attain a better understanding of the functions, agendas, and the editorial processes characteristic of news website creation.

In order to facilitate this kind of analysis, we had to develop new tools for data retrieval. Available programs for harvesting or spidering web pages operate under the assumption that one would want to capture data on a website only once, at most repeating the process once per day, and on average a couple of times per month. Existing tools do not allow one to easily capture data at a more frequent rate. Due to the need for frequent access, manual retrieval of data by hand was also not an option, leading us to develop custom code that would meet our needs.

The program we created visits our requested set of webpages every 60 seconds. It identifies the timestamp of the page, to see when it last changed, and if new changes have occurred, it captures and saves the data locally. The websites we queried were hosted on machines that could handle a large number of requests, but in order to make our frequent connections as unobtrusive as possible, we only downloaded the full page if it had been updated. We also only downloaded those objects of the page that we intended to analyze. By only downloading the headline image, for example, as opposed to all the image components on the page, we were able to reduce data requests.

Once the data were downloaded, the specific components that we were interested in analyzing were extracted and saved to a database. This required us to develop scripts that would parse the HTML page and pull out the relevant text and images that were of interest to us for further analysis.

3.1. Data

Using our custom code, we retrieved and analyzed data from three news websites: the Cable News Network (CNN), the British Broadcasting Corporation (BBC), and the English-language version of Aljazeera. News sites were selected as a data source because online news is claimed to be especially dynamic, with constant updates reflecting late breaking events [22]. Large, mainstream news sites were selected with the expectation that they would show frequent content changes. Moreover, we also expected that the three sites would tend to represent news events in somewhat different ways, given the different cultural and political perspectives of the nations (the US, the UK, and Qatar) in which the sites are produced, and consistent with previous findings of nationalist bias in news reporting [2, 6, 25, 28].

Data were captured over a three-week period starting on February 2, 2004 and ending on February 24, 2004. To make the amount of data more manageable, we focused on the headline stories displayed on each news site. The headline story, as in a traditional newspaper, is the main story that is prominently displayed on the webpage. We captured and analyzed changes in four components that make up the headline story: the *title* (identified as the most prominent title to the news story), the *blurb* (the short paragraph accompanying the title that summarizes the news story), the headline *image* (the prominent image accompanying the news story), and the *caption* (the text describing the image).

3.2. Analytical Methods

Once the data were captured, a comparison was done to determine if there were any changes between subsequent page updates for each website that we captured. For all remaining pages, for each specific site, comparisons were done between webpage *n* and webpage *n+1* (where *n+1* is the next collected webpage that had changed its content). Any identified changes between these pages were recorded. Comparisons were only done between sequential pages (e.g., page, *n* was compared to page *n+1*, but not to page *n+2*).

We employed three methods of manual analysis to identify changes in content: descriptive analysis, content analysis, and critical linguistics. At the first stage we used basic descriptive analysis to track the duration and number of changes. At the next stage we employed content analysis to investigate the nature of the changes and to identify a set of categories under which changes could be classified. We identified four categories of changes for images, and six categories of changes for the text of a headline story (see section 4).

After classification, manual content analysis allowed us to identify change between subtle manipulations of content that cannot easily and reliably be identified through automated means. In order to facilitate the manual coding of change, we developed a system that allows us to annotate the data directly. Researchers can use this system to pull up the captured data for a specific time period, and study the evolution of the news events. They can then code the data to reflect the identified changes or add comments for future analysis. Once the data have been coded and changes have been identified and categorized, our system automatically calculates the changes and prints a report identifying types of changes in regard to text, images, duration, and percent changes.

At the third and last stage of analysis, we employed qualitative methods from critical linguistics to examine closely changes in the text of the titles and blurbs that appeared to be ideologically motivated; that is, that did not introduce new content or make a factual or stylistic correction, in order to determine the extent and nature of ideologically-motivated changes in the three online news sites.

By using content and linguistic analysis, we are not only able to classify and understand the types of changes taking place, but also to gain insight into the ideologies of the media [16]. By applying this approach to news sites on the World Wide Web, a better theoretical and pragmatic understanding of this evolving medium can be obtained.

4. Coding Categories

In coding our data, we used a grounded theory approach to create categories to describe the changes found in the headline stories of the news sites [14]. The following categories apply to changes to the textual content:

New: Replacement of the entire story text with text that has not been previously presented, as in the case of the introduction of a new story.

Clarification: A change that modifies a portion of the text in such a way that makes a correction or clarifies or otherwise "improves" the text, without suggesting any change in perspective. The news story is modified in order to make the text more readable or to remove a formatting, punctuation, spelling or grammatical error.

Retraction: The removal of a piece of information from the text of the news story, or the reinstatement of a previous piece of information. A retraction may be motivated by the author or editor having prematurely published information on the news site before the facts had been verified by another source.

Update: A change that informs the reader of the current status of a story as it is unfolding, and that incorporates new information. The author or editor changes the text of the headline story to inform the reader of the latest events and changes as supplied by the news agency.

Repeat: The verbatim reappearance at a later time of the complete text of a news item that was previously introduced as new, and subsequently removed.

Ideological change (more): Linguistic changes to the text of the news story, including transformations, nominalizations, and change in subject-verb-object order, modality, lexical structure, or desirability [12], that present the same news from a non-neutral perspective. Such changes typically ascribe or mask responsibility, and/or incorporate an evaluation of the reported events or persons as good or bad. This type of change does not update or clarify.

Ideological change (less): Changes to the text of the news story that reduce ideology by removing traces of a biased perspective and that make the text more neutral.

It should be noted that news samples can be coded as exhibiting more than one change. For example, a news blurb could contain changes that, over a single interval, remove ideology, but also update the event to the reader.

Similar categories are used to classify changes to images. Definitions are given below, ordered by increasing extent of transformation to the image:

Crop: We identified images as being cropped if the photo has been enlarged, or the scene has been scaled back. With a crop, the image still displays the same content and perspective, but the photo editor has decided to remove the border to include more information from the periphery or to enlarge the picture and remove peripheral information.

Perspective: A change in perspective signifies that the image is still from the same scene, but the perspective is different. The photo depicts the same subject or event, but from a different camera angle or vantage point.

Context: A change in context signifies that the photo is from the same news event, but from a different part of the event. For example, an image sequence that shows a presidential candidate in one photo and a close-up of participants in the audience in the next photo is coded as a change in context.

Repeat: Similar to the textual coding scheme, if the same photograph reappears at a later point in time after having been previously introduced as new and then removed, we code it as a repeat.

New: Similar to the textual coding scheme, we classify a change as being new if an image is displayed to the reader that has not been shown previously.

Miscellaneous: Images that did not fit in any of the above categories are coded as miscellaneous. This category includes non-photographic images (graphics).

Each image change was coded for only one of the codes listed above.

Both authors coded all the data, with an interrater agreement level in excess of 80%. Disagreements in coding were resolved through discussion.

5. Results

5.1. Changes over Time

For the three week period, a total of 2,244 changes to the text and images of headline news story were analyzed: 369 changes for CNN, 257 for the BBC, and 1,618 for Aljazeera. Out of all the changes, new news stories were posted 58 times by the BBC, 67 times by CNN, and 60 times by Aljazeera.

A micro-longitudinal study allows us to collate and track the number of changes that occur by hourly periods, from which we can then develop a timeline (Figure 1).

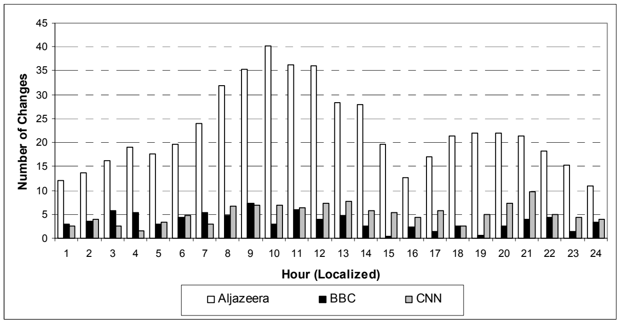

Figure 1. Average number of updates by hour

For Aljazeera, BBC, and CNN, Figure 1 shows that even though changes to the websites are made every hour, there is an increase in number of changes between 7:00 a.m. and 2:00 p.m., and another small increase after 7:00 p.m. (time being localized to the respective time zones of each news organization). These results indicate that rather than being updated at a constant rate, the frequency of changes in these news websites follows the rhythm of business hours in the countries in which they are based.

5.2. New Information versus Revision

With the six categories of changes to the textual content of the webpage, a distinction can be made between revisions that bring no newsworthy information to the reader (clarification, retraction, repetition, more ideology, less ideology), and changes that present new information (new, update). From a news reader's perspective, the latter are most interesting. A reader does not typically view a news site to see what news items have had minor grammatical or ideological adjustments. Rather, they are interested in reading about new or updated events. We analyzed our data to see how many changes over the three-week period result in the addition of substantive information, versus changes that do not provide substantive new information (for the purpose of this analysis, we ignored any repeats). The results are displayed in Table 1.

Table 1. Count of textual changes that add or revise

	New Information or Update	Revised Information	Percent New
Aljazeera	69	72	49%
BBC	129	123	51%
CNN	159	174	48%

Table 1 shows that out of the total number of changes to the text that accompany a major news story for each website over the three-week period, there is approximately a 50% chance that a change to textual content of a headline story by Aljazeera, the BBC or CNN would add substantial new information.

Headline news stories remain prominent on a page for an extended period of time. A histogram (Figure 2) shows that for CNN, the BBC, and Aljazeera, most news stories remain active for 250 to 500 minutes. In one extreme case, CNN's coverage of the proceedings of the U.S. Democratic caucus remained active as the main news story for 36 hours. During this period, updates, revisions, and change in ideology were presented only as they related to the Democratic nomination race, and no other news events were covered.

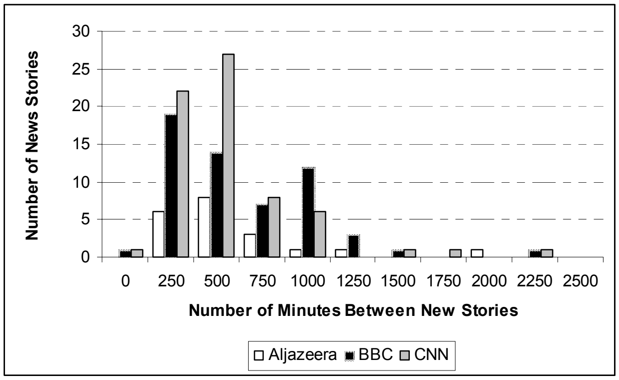

Figure 2. Histogram of story duration

5.3. New Information: Updates

As news information is processed by a news agency, subject updates to the news blurb occur. By tracking the number of updates, we identified that, on average, CNN updates a story six times before introducing a new story. BBC stories get updated an average of three times, and for Aljazeera, which updates published news blurbs infrequently, the number of updates is effectively two. The longest continuous sequence of updates by CNN is nine, starting on 2/3/04 at 9:05 p.m. and ending on 3/5/04 at 9:47 p.m.. Table 2 shows how the headline title evolves during this period.

Table 2. Example of title updates (CNN)

#	Headline Title
1	Kerry wins Arizona primary, CNN projects
2	Kerry wins North Dakota caucus, CNN projects
3	Kerry wins four states, Edwards one
4	Kerry wins five states, Edwards one
5-8	Kerry wins in East, Midwest, West; Edwards takes S.C.
9	Kerry rolls, wins five states; Edwards takes S.C.

In the BBC sample, the longest sequence of continuous updates (*n=8*) begins on 2/20/04 at 12:09 a.m. and ends on 2/21/04 at 9:46 p.m. The focus of this news story was on the Iranian elections (Table 3).

Table 3. Example of title updates (BBC)

#	Headline Title
1	Turnout will be key in Iran poll
2-3	Khamenei urges Iranians to vote
4	Iranians get extra time to vote
5	Polls close in key Iran election
6-7	High Iran poll turnout claimed
8	Iran's conservatives in the lead

By looking only at the titles, one can identify a progression of events. As new information is received, the corresponding title and blurb on the website are updated to reflect the latest information. For example, as John Kerry wins the Democratic caucus in various states, the titles are modified to show the tally of wins. Similarly, the BBC titles track the progression of the Iranian election from concerns about voter turnout, to initial predictions of vote leads.

5.4. Revision: Overview

Any changes that do not add substantive new content (clarification, retraction, repetition, more ideology, less ideology) were classified as revisions. In the case of revisions, although the webpage may inform the reader that the page has been updated, in reality, the change does not add new information. Table 4 shows the average number of revisions identified for a one-week period and the percentage value for each type of revision.

Table 4. Breakdown of textual revisions for one week

	Average Total Count		
	CNN	BBC	ALJ
Clarification	31.3	21.7	10.0
More ideology	21.7	14.7	5.0
Less ideology	3.3	1.3	1.0
Retraction	1.7	3.3	4.0
Repeat	0.0	0.0	6.0

	% of Revision		
	CNN	BBC	ALJ
Clarification	54%	53%	28%
More ideology	37%	36%	19%
Less ideology	6%	3%	4%
Retraction	3%	8%	15%
Repeat	0%	0%	23%

One difference between Aljazeera, CNN, and the BBC is that if Aljazeera makes a revision it is more likely to repeat a story (i.e., remove a story from the main story position and then redisplay it at a later date), or to retract a story (i.e., remove information that was published prematurely and subsequently determined to be incorrect). Aljazeera publishes almost the same number of new stories (*n=60*) as CNN (*n=67*) and the BBC (*n=58*), but once a new story is published it less likely that Aljazeera editors will go back and update or modify what they have published. This can be seen in Table 5, which tracks the number of times a new news item is posted in comparison to new textual information being added to a preexisting news item, or a revision being made that does not add any content. Aljazeera maintains the initial posted story, and is less likely to revise or update, while CNN and the BBC are more likely to publish a story, and then revise the structure over time and update the content to reflect the latest available news.

Table 5. Comparison of new news, updates and revisions

	New News	Updates	Revisions
Aljazeera	51%	25%	25%
BBC	23%	28%	49%
CNN	20%	28%	52%

5.5. Revision: Clarification

The majority of revisions identified for all three news sites modify the message in an attempt to remove grammatical errors or to improve the prose. As an example, CNN published the following title and accompanying news blurb:

> **Car bomb kills 47 would-be Iraqi soldiers**
> In the second deadly attack in as many days, a suicide car bomber killed 47 people today in Baghdad, most of them Iraqi men standing in line to join the Iraqi army, Iraqi medical officials said. On Tuesday, a truck bombing killed 53 people near a police station south of Baghdad**.** (CNN, 2/11/04, 11:54 a.m.).

One hour later, at 12:46 p.m., “On Tuesday” in the second sentence was changed to “Yesterday.” Here one can identify a revision that attempts to clarify the time of the news event, without adding any bias or new information. As another example, the BBC has the following news item:

> **Many dead in Moscow metro blast**
> About 40 people die and more than 100 are injured in a suspected suicide attack on a packed Moscow underground train. (BBC, 2/6/04, 3:46 p.m.).

Two minutes later the blurb was modified to:

> **Many dead in Moscow metro blast**
> About 40 people die and more than 100 are injured in a suspected suicide attack on a packed Moscow subway train. (BBC, 2/6/04, 3:48 p.m.).

The news item remains the same, except for a change of wording from *underground train* to *subway train.* This change was reversed 28 minutes later at 4:16 p.m. back to *underground train*, then changed back to *subway train* at 4:39 p.m., back to *underground train* at 5:38 p.m., and finally back to *Subway train* at 5:49 p.m.

In this example, over a period of two hours, no new information is supplied to the reader; rather the editor tries to determine the most appropriate term to describe the underground transportation system that was bombed.

5.6. Revision: Ideology

A more subtle type of revision to the text of headline news stories that we identified deals with ideology, or perspectival shift. Such changes do not make the content more readable, but rather insinuate an evaluation or bias into the reporting of supposedly neutral facts. Table 4 shows that the second most common type of revision (after clarification) adds ideology. This suggests that the presentation of news stories tends to evolve from more to less neutral over time. As an example of increased bias, CNN changed the following news headline:

Ricin find closes 3 Senate office buildings (CNN, 2/3/04 10:29 a.m.)

to

Ricin scare closes Senate buildings (CNN, 2/3/04 10:49 a.m.)

From a relatively neutral wording indicating that ricin was found, the title changes 20 minutes later to “ricin scare,” a more emotionally manipulative noun phrase suggesting that the affected participants were frightened. The title containing “scare” remains active until 4:08 p.m., when it is changed to:

Senate awaits ricin tests (CNN, 2/3/04, 4:08 p.m.)

Here a decrease in ideological loading has occurred, along with an update on the current state of events with regard to the ricin investigation.

Another example from the BBC in which it is possible to identify an increase in ideology is when a headline news blurb is changed from:

> Thirty-nine people die in a Moscow train blast which President Putin says was an attack by Chechen rebels. (BBC, 2/6/04 10:24 a.m.)

to:

> The Russian leader blames Chechen rebels for a Moscow underground blast which killed at least 39 people. (BBC 2/6/04 11:25 a.m.)

No new information is conveyed by this change; both versions identify the scene of the event, the number of people who died, and the fact that leadership attributes the cause of the blast to Chechen rebels. However, the change from a verb of saying to a verb of blaming represents

Putin as stronger in mind and disposition—more leader-like—and the rebels as more blameworthy (cf. [23]).

5.7. Image Analysis

The results of the image analysis show that multiple pictures typically accompany a single headline story. Table 6 indicates that Aljazeera uses the most images to accompany a story—on average 3.23 images per new story. The BBC is the most reserved when it comes to displaying multiple images, using fewer than two images per story.

Table 6. Average number of headline images per week

	Number of stories	Number of new images	Number of Images per Story
Aljazeera	8.7	28.0	3.2
BBC	19.3	35.3	1.8
CNN	22.3	63.7	2.9

Table 7, obtained by calculating the type of image changes over a one week average, shows that if the BBC or CNN website is modified with a change in image, there is a strong likelihood that the image will be new. On the Aljazeera website one is most likely to encounter repeat images; that is, images that have previously been displayed. It is our impression that Aljazeera's content management system takes a set of images that accompany a news story, and then rotates through the images over random or predetermined intervals. A change in image context (or repetition of images) accompanying the same story is an easy way for the news agency to provide little new information to the reader, while at the same time making it seem as though the webpage is dynamic. This is in contrast with the CNN and BBC websites, which do not rotate a set of images and which are more likely to permanently replace a previous image with a new one.

CNN and BBC are also more likely than Aljazeera to show a different context or perspective in their images. As a news event unfolds, these two sites show multiple, related images associated with the story. An example from CNN of a progression that involves shifts in context and perspective is given in Images 1-6. These photos accompanied a news story that appeared on February 24, 2004, reporting discussion of a proposed constitutional ban in the United States on same-sex marriages.

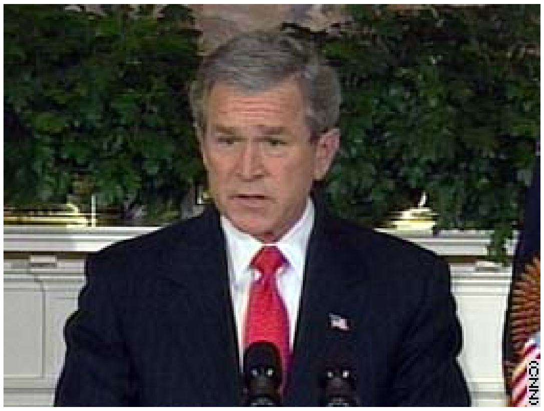

Image 1. CNN, 2/24/04 11:01a.m. Duration: 42 minutes (new)

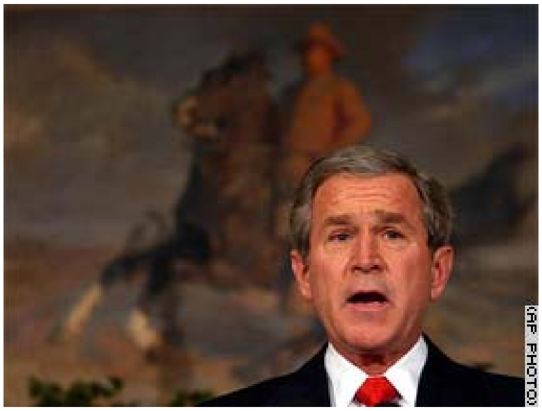

Image 2. CNN, 2/24/04 11:43a.m. Duration: 130 minutes (change of context)

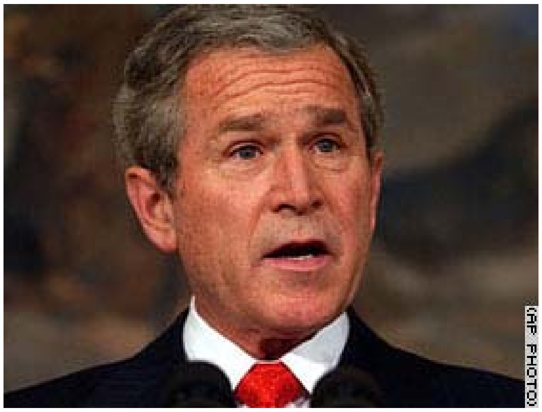

Image 3. CNN 2/24/04 13:53a.m. Duration: 122 minutes (change of perspective)

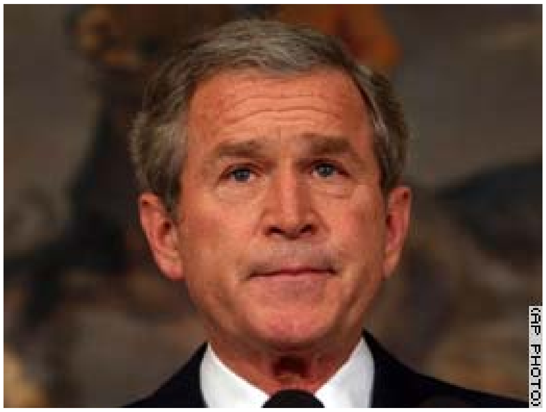

Image 4. CNN 2/24/04 15:55a.m. Duration: 109 minutes (change of perspective)

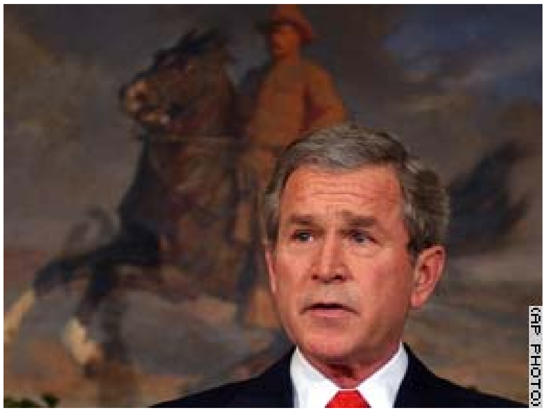

Image 5. CNN 2/24/04 17:44 a.m. Duration: 77 minutes (change of perspective)

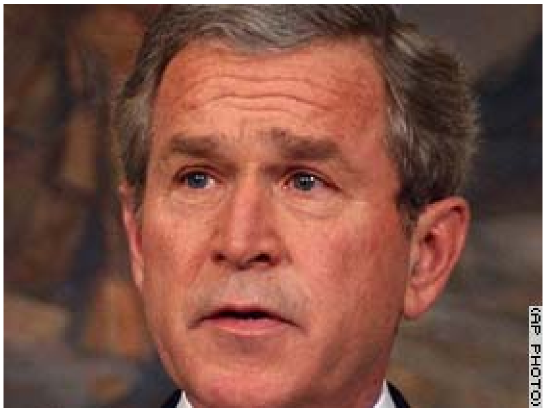

Image 6. CNN 2/24/04 19:01 a.m. Duration: 327 minutes (crop)

Despite the fact that it consisted mostly of discussion by U.S. lawmakers with a common point of view, this news story remained active for almost 14 hours, during which period the story was revised, updated, and corrected multiple times. On three occasions (images 4-6), the only changes were to the image accompanying the story, without any modification of the textual content (news title, blurb, or image caption). Thus the only new "information" added by those three changes was different views of U.S. President George W. Bush's face.

A small percentage of changes to the images on CNN and BBC are crops. On occasion this leads to a modified image sequence that is barely perceptible to the user, without a side-by-side comparison of the old and new images. That the news sites go to the effort of making such subtle changes shows that the presentation of images, just as the presentation of textual information, is perceived as important to the staging of news stories on the website.

Table 7. Weekly average of types of image change

	CNN	BBC	Aljazeera
Crop	5.1%	0.9%	0.0%
Perspective	13.8%	3.5%	0.7%
Context	31.1%	29.0%	4.2%
Repeat	2.6%	7.0%	80.3%
New	46.9%	58.8%	14.8%
Misc	0.5%	0.9%	0.0%

6. Discussion

Traditional information retrieval is concerned with mining a large amount of data over an intermittent time period. Micro-longitudinal content analysis attempts to focus on a smaller domain, retrieving detailed information about changes that are taking place as a webpage evolves. The application of this method demonstrates in a tangible manner how dynamic a webpage can be, how information that we may naively imagine to be fairly static (even for a website genre that is assumed to change frequently) actually is in constant flux. Moreover, the taxonomy we developed from the systematic study of changes occurring to news websites shows that the changes are of different types, some conveying substantive new content, while others revise the content of news stories for grammatical, stylistic, or ideological reasons, and others simply recycle previously presented content.

This taxonomy and the quantitative findings derived from its application can be used to guide future work, for example in developing effective tools for data capture, such as news gathering engines. Our findings show that if one were to want to develop a headline news aggregation engine that broadly captures news stories across multiple sites, such as that made available by Google News, one could aggregate data every four hours, and still capture a majority of the headline news stories. Aggregating every one-and-a-half to two hours would capture a majority of the images accompanying such stories, at least on CNN and the BBC. However, if one wanted, e.g., for archival reasons, to capture accurately all the information presented on a news site for an event, one would need to capture data at a very high rate, or else risk missing the evolution of the story.

While today's storage technology allows for such a method of data capture, exhaustive archiving raises issues of how to accommodate future access and retrieval. For example, when retrieving an archived news story, how does one determine which version represents the "true" story? How should one identify, retrieve, and display the evolution of a news story? Traditionally, text summarization has focused on synthesizing a story from various sources [20, 24], but one could also use the editorial process of these news websites as rendered transparent by micro-longitudinal analysis to attempt to synthesize a news report from a single news producer. In constructing a summary of a story's contents, one might wish to exclude versions containing spelling errors that were later corrected, as well as versions that were later revised to add or remove an ideological nuance.

From the human perspective, by capturing all the changes taking place on organizational websites, we render their document creation process transparent, and gather insight into the organizations' focus and political stance. This can be of particular value when no other ways of gaining deep access are available. Rendering the editorial process transparent also has the potential to raise readers' awareness of the nature of news production and of the ephemeral nature of information on news websites, contrary to what readers may naively assume are stable facts. A website credibility study done in 2002 [11] found that for the genre of news sites, 30.2% of respondents identified concerns about information bias as relevant to judging the credibility of a website. To facilitate a deeper understanding, one can directly catalog bias and trivial grammatical mistakes. This could help to humanize the newsmaking regime, and empower users to evaluate the quality and creditability of stories presented on a website by a news provider.

Finally, micro-longitudinal content analysis validates previous research findings that news reporting reflects bias, on the web as in traditional mass media [6, 9]. This bias may be ideological in origin, reflecting the agendas and worldviews of the editorial staff or of the media conglomerate of which they are a part, or they may be motivated by the drive to sell the media through the use of provocative or emotionally rousing news copy. In either case, rhetorical manipulation to project a particular perspective is typically superfluous to the events of the story, although it may have a persuasive impact on readers.

7. Summary

In this study, we have presented micro-longitudinal content analysis as a methodology to facilitate new ways of examining and retrieving information from web-based documents. We applied our methodology to the genre of news websites, which are traditionally assumed to be dynamic in nature. From this we developed a classification system that can be used to identify changes to a news based website. Using this classification we analyzed a set of websites at micro intervals over a three week period, and showed that even though these pages are dynamic in nature, on average 50% of the time the changes taking place are of no news value to the reader, but instead are changes that attempt to clarify the message, or give the story an ideological "spin." We also showed that the editorial process taking place on western news sites (BBC, CNN) is open and transparent, in that revisions are continuously being made to the online news story. This is a contrast to Aljazeera, where news stories are not as likely to evolve and change in the public eye.

8. Future Work

Having identified the categories of change taking place in text and images, we can now apply our classification system to entire news sites (rather than just the headline stories), e.g., to analyze minor news stories and so-called "soft news" [4]. In addition, in future work, we plan to capture changes on other types of websites that (purportedly) do not change as frequently, for example corporate sites. We are also planning to apply the methodology to weblogs, where, as with news sites, multiple revisions of an entry are possible, but not immediately transparent, hence the editorial process can be hidden if readers do not frequently review the site.

References

[1] Allan, J., Papka, R., & Lavrenko, V. (1998). On-line news event detection and tracking. *Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 37 – 45.

[2] Al Shabbab, O., & Swales, J. (1986). Rhetorical features of Arab and British news broadcasts. *Anthropological Linguistics*, 28(1), 31-42.

[3] Bell, P. (2001). Content analysis of visual images. In T. van Leeuwen & C. Jewitt (eds.), *Handbook of Visual Analysis* (pp. 10-34). London: Sage.

[4] ben-Aaron, D. (2003). When news isn't news: The case of national holidays. *Journal of Historical Pragmatics, 4* (1), 75-102.

[5] Billsus, D. & Pazzani, M. J. (2000). User modeling for adaptive news access. *User Modeling and User-Adapted Interaction,* 10 (2/3), 147-180.

[6] Brookes, H. J. (1995). 'Suit, tie and a touch of juju'-The ideological construction of Africa: A critical discourse analysis of news on Africa in the British press. *Discourse & Society*, 6 (4), 461-494.

[7] Cameron, D. (1996). Style policy and style politics: A neglected aspect of the language of the news. *Media Culture & Society*, 18 (2), 315-333.

[8] Davis, H. H., & Walton, P. A. (1983). Sources of variation in news vocabulary: A comparative analysis. *International Journal of the Sociology of Language*, 40, 59-75.

[9] Dunsky, M. (2001). Missing: The bias implicit in the absent. (News reporting of the Israel-Arab conflicts). *Arab Studies Quarterly*, June 22. Retrieved June 22, 2004 from http://static.highbeam.com/a/arabstudiesquarterlyasq/june222001/missingthebiasimplicitintheabsentnewsreportingofth/

[10] Eichmann, D., & Srinivasan, P. (2002). Adaptive filtering of newswire stories using two-level clustering. *Information Retrieval*, 5 (2/3), 209-237.

[11] Fogg, B.J., Kameda, T., Boyd, J., Marshall, J., Sethi, R., Sockol, M., & Trowbridge, T. (2002). *Stanford-Makovsky Web Credibility Study 2002: Investigating what makes Web sites credible today.* http://captology.stanford.edu/pdf/Stanford-MakovskyWebCredStudy2002-prelim.pdf

[12] Fowler, R. (1991). *Language in the news: Discourse and ideology in the British press.* London New York: Routledge.

[13] Froomkin, D. (2004). Ideas for online publications: Lessons from blogs, other signposts. *Online Journalism Review*, May 26. http://ojr.org/ojr/workplace/1085603014.php

[14] Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.

[15] Hayes, P; Knecht, L.; & Cellio, M. (1997). A news story categorization system. Originally appeared in *Proceedings of the 2nd Conference on Applied Natural Language Processing*, 1988 (pp. 518-526). San Francisco: Morgan Kaufmann Publishing.

[16] Jolliffe, L. (1993). Yes! More content analysis! *Newspaper Research Journal*, 14 (3-4).

[17] Jucker, A. H. (1992). *Social stylistics. Syntactic variation in British newspapers*. Berlin: Moutonde Gruyter.

[18] Kurtz, A.J., & Mostafa, J. (2003). Topic detection and interest tracking in a dynamic online news source. *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, Houston, TX, 122-124.

[19] Livingston, S. (June, 1997). *Clarifying the CNN Effect: An Examination of Media Effects According to Type of Military Intervention*. The Joan Shorenstein Center for Press Politics. http://www.ksg.harvard.edu/presspol/publications/pdfs/70916_R-18.pdf

[20] Klavans, J. L. (2004). *Current research in multilingual, multi-document text summarization: The Columbia University Newsblaster experience.* Keynote talk, Digital Documents Track, Thirty-seventh Hawaii International Conference on System Sciences (HICSS-37).

[21] Makkonen, J.; Ahonen-Myka, H., & Salmenkivi, M. (2004). Simple semantics in topic detection and tracking. *Information Retrieval*, 7 (3/4), 347-357.

[22] Massey, B., & Levy, M. (1999) Interactivity, online journalism, and English-language Web newspapers in Asia. *Journalism and Mass Communication Quarterly*, 76 (1), 138-151.

[23] McAdams, K. C. (1990). Power prose: The syntax of presidential news. *Journalism Quarterly*, 67 (2), 313-322.

[24] McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Sable, C., Schiffman, B., & Sigelman, S. (2002). Tracking and summarizing news on a daily basis with Columbia's Newsblaster. *Proceedings of the 2002 Human Language Technology Conference* (HLT). San Diego, California.

[25] Meeuwis, M. (1993). Nationalist ideology in news reporting on the Yugoslav crisis: A pragmatic analysis. *Journal of Pragmatics*, 20 (3), 217-237.

[26] Rasolofo, Y., Hawking, H., & Savoy, J. (2003). Result merging strategies for a current news metasearcher. *Information Processing & Management*, 39 (4), 581-609.

[27] Sakagami, H., & Kamba , T. (1997). Learning personal preferences on online newspaper articles from user behaviors. *Proceedings of the 6th International WWW Conference*, Santa Clara, CA. 291-300.

[28] Sidiropoulou, M. (1995). Causal shifts in news reporting: English vs Greek press. *Perspectives: Studies in Translatology*, 1, 83-98.

[29] van Leeuwen, T. (2001). Semiotics and iconography. In T. van Leeuwen & C. Jewitt (eds.), *Handbook of Visual Analysis* (pp. 92-118). London: Sage.