Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty

Evgeniy Gabrilovich [*]
CS Department
Technion
32000 Haifa, Israel
gabr@cs.technion.ac.il

Susan Dumais
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
sdumais@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
horvitz@microsoft.com

ABSTRACT

We present a principled methodology for filtering news stories by formal measures of information novelty, and show how the techniques can be used to custom-tailor newsfeeds based on information that a user has already reviewed. We review methods for analyzing novelty and then describe Newsjunkie, a system that personalizes news for users by identifying the novelty of stories in the context of stories they have already reviewed. Newsjunkie employs novelty-analysis algorithms that represent articles as words and named entities. The algorithms analyze inter- and intra- document dynamics by considering how information evolves over time from article to article, as well as within individual articles. We review the results of a user study undertaken to gauge the value of the approach over legacy time-based review of newsfeeds, and also to compare the performance of alternate distance metrics that are used to estimate the dissimilarity between candidate new articles and sets of previously reviewed articles.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering*; H.3.4 [**Information Storage and Retrieval**]: Systems and software—*user profiles and alert services*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*linguistic processing*; G.3 [**Probability and Statistics**]: *distribution functions, experimental design*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

News, novelty detection, personalization

1. INTRODUCTION

Just a decade ago, large-scale flows of information such as newsfeeds were owned, monitored, and filtered by organizations specializing in the provision of news. The Web has brought the challenges and opportunities of managing and absorbing newsfeeds to all interested users. We have pursued mathematical tools and user-interface

---

designs that can assist people with extracting the most relevant information from news sources. One approach to navigating a large corpus of news information is to identify differences and similarities between various decompositions of the collections. Such differencing machinery could be useful for distinguishing what is being said for the first time from what has already been mentioned, as well as in revealing differences of opinion on the issues covered.

Identifying "important" information has been an essential aspect of studies on Web search and text summarization. Search methods focus on identifying a set of documents that maximally satisfies a user's acute information needs. Summarization strives at compressing large quantities of text into a more concise formulation. In the absence of automated methods for identifying the deep semantics associated with text, prior work in summarization has typically operated at the level of complete sentences, weaving together the most representative sentences to create a document summary. Research on search and summarization has generally overlooked the dynamics of informational content arriving continuously over time.

We shall present methods for identifying information novelty and show how these methods can be applied to manage content that evolves over time. We start by describing a general framework for comparing collections of documents. We assume documents are organized into groups by their content or source, and analyze inter-group and intra-group differences and commonalities. Juxtaposing two groups of documents devoted to the same topic but derived from two distinct sources (*e.g.*, news coverage of an event in different parts of the world) can reveal interesting differences of opinions and overall interpretations of situations. Moving from static collections to sets of articles generated over time, we can examine the evolution of content. For example, we can seek to examine a stream of news articles evolving over time on a common story, with the goal of highlighting truly informative updates and filtering out a large mass of articles that largely relay "more of the same."

In contrast to prior work in text summarization, we work at the level of individual words rather than entire sentences. Working at this finer resolution, we gather detailed statistics on word occurrence across sets of documents in order to characterize differences and similarities among these sets. We further enhance the simple bag of words model by extracting *named entities* that denote names of people, organizations, and geographical locations. In contrast to phrases and collocations—whose discriminative semantic properties are usually outweighed by lack of sufficient statistics—named entities identify relatively stable tokens that are used in a common manner by many writers on a given topic, and so their use contributes a considerable amount of information. In fact, one type of analysis we describe below represents articles using only the named

entities found in them. We found that, for some topics, this analysis was comparable or even superior in performance to methods that manipulate the full bag of words.

We will focus on the analysis of live streams of news. Live news streams pose tantalizing challenges and opportunities for research. Newsfeeds span enormous amounts of data, present a cornucopia of opinions and views, and include a wide spectrum of formats and content from short updates on breaking news, to major recaps of story developments, to mere reiterations of "the same old facts" reported over and over again. We describe algorithms that identify significant updates on the stories being tracked, relieving the users from having to sift through long lists of similar articles arriving from different news sources. The methods provide the basis for personalized news portal and news alerting services that promise to minimize the time and disruptions to users who wish to follow evolving news stories.

The contributions of this paper are threefold. First, we present a framework for identifying differences in sets of documents by analyzing the distributions of words and recognized named entities. This framework can be applied to compare individual documents, sets of documents, or a document and a set (for example, a new article vs. the union of previously reviewed news articles on the topic). Second, we present a collection of algorithms that operate on live news streams and provide users with a personalized news experience. These algorithms are implemented in a system named Newsjunkie that presents users with maximally informative news updates. Users can request updates per every user-defined period or per each burst of reports about a story. Users can also tune the desired degree of relevance of these updates to the core story, allowing delivery of offshoot articles that report on related or similar stories. Finally, we describe an evaluation method which presents users with a single seed story and sets of articles ranked by different novelty-assessing metrics, and seeks to understand how participants perceive the novelty of these sets in the context of the seed story. As we shall discuss, the results of this study highlight several interesting issues such as what is considered a story "on a given topic," how people judge novelty, and how their judgments are influenced by the relevance of the information being read to the topic defined by the seed story.

The remainder of this paper is organized as follows. In Section 2, we review related research in language modeling, text summarization and topic detection and tracking (TDT). In Section 3, we present a framework for comparing text collections, and discuss its applications to finding differences both between and within groups of documents. We report the results of experiments with users in Section 4. In Section 5, we describe two new types of document analyses we developed that allow users to personalize the frequency and content of the news updates that they receive. Finally, we discuss future research directions in Section 6.

2. BACKGROUND AND RELATED WORK

The AT&T Internet Difference Engine (AIDE) [9] was one of the earliest attempts to develop a tool for comparing the content of material drawn from the Internet. The HtmlDiff system built in the course of this project looked for simple syntactic differences between Web pages similarly to what the Unix *diff* utility does; however, no attempt was made to capture semantic differences between the pages.

In an effort to characterize content differences beyond simple syntactic variations, several past studies focused on identifying words that are particularly characteristic of a given text. Kilgarriff [15] presents a good summary of these studies as well as their potential applications to characterizing text genres and differences

in male and female speech.

Comparing distributional properties of individual words naturally evolved into comparing language models for entire text collections, which was found useful in two ways. In corpus linguistics, researchers studied to what degree reasoning about one text collection can be based on the model computed for another collection [15]. In the realm of information retrieval, comparing the language models for documents returned by a query and the entire collection was found to be valuable for predicting query quality (*i.e.*, how well the query is formulated and how focused the results are expected to be) [7, 6].

In this work we reason about the similarity of document sets by comparing their language models using several distance metrics. Another way to compare document sets is to represent the documents as bags of words, and use established word similarity metrics that sum over pairwise distances between individual words. Prior work on *semantic* word similarity developed in two major directions, either using purely statistical analysis such as Latent Semantic Indexing (LSI) [8] and Hyperspace Analog to Language (HAL) [2], or capitalizing on background knowledge resources such as WordNet [20]. The two approaches have also been combined in a single framework [11]. Lee [17] investigated measures of *distributional* word similarity by using language models similar to ours to improve probability estimations for unseen word cooccurrences.

We seek to order news by novelty in an iterative manner, attempting at each step to identify the article that carries the maximum amount of new information, in the context of a background model composed from content that has been already reviewed. This approach is conceptually related to the notion of Maximum Marginal Relevance (MMR) [3], developed in the context of information retrieval. When selecting the next document to be returned in response to a query, the MMR criterion prefers relevant documents that are maximally different from documents that have been selected before.

Research on Topic Detection and Tracking (TDT) has led to investigations of a variety of problems related to novelty detection. Yang et al. [27] studied a specific form of novelty detection, namely First Story Detection (FSD). The authors used a Rocchio-style text classifier [26] to classify documents into predefined broad topics, and then measured the novelty of new documents given the topics predicted for them. Documents were represented as vectors of words and named entities weighted with a TF.IDF scheme [22].

Working in the domain of news, Swan and Jensen [24] automatically generated timelines from historic date-tagged news corpora (TDT). They used a χ^2 test to identify days on which the number of occurrences of a given word or phrase exceed some (empirically determined) threshold, and then generated timelines by grouping together contiguous sequences of such days.

Kleinberg [16] used randomized infinite-state automata to model burstiness and hierarchical structure in text streams. This work was not specifically focused on news, but experimented with other kinds of time-stamped text corpora such as personal email archives and conference proceedings spanning a number of years. Sample applications of this formal model include identification of increased activity bursts in email trails, as well as computing the most important words in conference paper titles over different time periods. In contrast to [24], this approach zooms into the bursts to determine their hierarchical structure, identifying short, intense bursts within longer but weaker ones.

Assessing the novelty of a quantity of text is related to analyzing its most informative sentences. The latter task has been the focus of research in extractive text summarization, which combines most important sentences of a collection of documents to produce

a summary. A study performed in the course of the Columbia Newsblaster project [23] identified key sentences by looking for importance-signaling words and high-content verbs obtained by analyzing large news corpora, as well as finding dominant concepts by consulting WordNet [10]. Allan et al. [1] computed usefulness and novelty measures at the sentence level by developing probabilistic models for news topics and events. Collins-Thompson et al. [4] selected sentences by using a C4.5 [19] classifier with a set of surface and semantic sentence features; they also compared pairs of sentences to check if one of them is a statistical translation of the other.

In contrast to prior research, our work focuses on developing a system that manipulates live newsfeeds and offers users personalized updates that are maximally novel in the context of information the user has reviewed before.

3. A FRAMEWORK FOR COMPARING TEXT COLLECTIONS

Given two sets of textual content, how can we characterize the differences between them? Answering this question is useful in a variety of applications, including automatic profiling and comparison of text collections, automatic identification of different views, scopes and interests reflected in the texts, and automatic identification of novel information.

In general, several aspects of "difference" may be investigated:

- Differences in *content* may reflect the different ways a particular person or event is described in two sets of documents. For example, consider analyzing differences in predefined partitions, *e.g.*, comparing US vs. European reports on various political issues, or comparing the coverage of a recent blackout of the East Coast of the United States in the news coming from sources based in the East Coast and West Coast.

- Differences in *structural organization* may go well beyond text and also consider link structure of Web sites, *e.g.*, comparing IBM Web site vs. Intel Web site.

- Differences in *time* (*i.e.*, temporal aspects of content differences) can reveal interesting topical changes in series of documents. This kind of analysis can be used to compare today's news vs. the news published a month or a year ago, to track changes in search engine query logs over time, or to identify temporal changes in topics in users' personal email.

Temporal differences are a particularly interesting case, and in this paper we focus on automatically assessing the *novelty* over time of news articles coming from live newsfeeds. Specifically, we formulated the following two research challenges:

1. Characterization of novelty in news stories, which allows us to order news articles so that each article adds maximum information to the (union of) previously read ones.

2. Studying topic evolution over time, which enables us to quantify the importance and relevance of news updates, granting end users control over these parameters and offering them a personalized news experience.

In the remainder of this section, we outline a methodology for analyzing newsfeeds and describe an algorithm for ranking news articles by predicting the amount of novel information they carry. The results of an empirical evaluation of this algorithm are reported in the next section. Section 5 studies topic evolution over time and proposes a new approach to analyzing different types of articles.

3.1 Methodology

We developed a software toolset named Newsjunkie that implements a collection of algorithms and a number of visualization options for comparing text collections. Newsjunkie represents documents as bags of words augmented with named entities extracted from the text. In-house extraction tools were used for this purpose, which identified names of people, organizations and geographical locations.

Document groups contain documents with some common property, and constitute the basic unit of comparison. Examples of such common properties can be a particular topic or source of news (e.g., blackout stories coming from the East Coast news agencies). We draw inferences about the differences between document groups by first building a language model for each group, and then comparing the models using some similarity metric (see Section 3.2). To facilitate exploring a variety of language models, Newsjunkie represents documents either as smoothed probability distributions[1] over all the features (words + named entities), or as vectors of TF.IDF weights [22] (in the same feature space).

3.2 Ranking news by novelty

Let us imagine the common situation where something interesting happens in the world, and the event is picked by the news media. If the event is of sufficient public interest, the ensuing developments are tracked in the news as well. Suppose you have read an initial report and, at some later time, are interested in catching up with the story. In the presence of Internet sites that aggregate thousands of news sources such as Google News or Yahoo! News[2], your acute information-seeking goal can be satisfied in many ways and with many more updates than even the most avid news junkie has the time to review. Automated tools for sifting through a large quantity of documents on a topic that work to identify nuggets of genuinely new information can provide great value.

Avoiding redundancy and overlap can help minimize the overhead associated with tracking news stories. There is a great deal of redundancy in news stories. For example, when new developments or investigation results are expected but no new information is yet available, news agencies often fill in the void with recaps of earlier developments until new information is available. The situation is further aggravated by the fact that many news agencies acquire part of their content from major multi-national content providers such as Reuters or Associated Press. Users of news sites do not want to read every piece of information over and over again. Users are primarily interested in learning *what's new*. Thus, ordering news articles by *novelty* promises to be useful.

We use a number of document similarity metrics to identify articles that are most *different* from the union of those read previously.[3]

We implemented the following distance metrics:

- Kullback-Leibler (KL) divergence [5], a classical asymmetric information-theoretic measure. Assume we need to compute the distance between a document d and a set of documents R. Let us denote the probabilistic distributions of words (and named entities if available) in d and R by p_d and p_R, respectively. Then, $dist_{KL}(p_d, p_R) =$

---

[1]Two smoothing options were implemented, using either Laplace's law of succession [21] or linear smoothing with word probabilities in the entire text collection [18]; the latter option was used throughout the experiments reported in this paper.

[2]http://news.google.com and http://dailynews.yahoo.com, respectively.

[3]In what follows, we use the term *distance* metrics to emphasize the fact that we are actually looking for documents that are most *dissimilar* from documents reviewed earlier.

$\sum_{w \in words(\{d\} \cup R)} p_d(w) \log \frac{p_d(w)}{p_R(w)}$. Note that the computation of $\log \frac{p_d(w)}{p_R(w)}$ requires both distributions to be smoothed to eliminate zero values (corresponding to words that appear in d but not in R, or vice versa).

- Jensen-Shannon (JS) divergence [5], a symmetric variant of the KL divergence. Using the definitions of the previous item, $dist_{JS}(p_d, p_R) = \frac{dist_{KL}(p_d, q) + dist_{KL}(p_R, q)}{2}$, where $q = \frac{p_d + p_R}{2}$.

- Cosine of vectors of raw probabilities (computation does not require *smoothed* probabilities).

- Cosine of vectors of TF.IDF feature weights.

- A metric we formulated to measure the density of previously unseen named entities in an article (referred as NE). The intuition for this metric is based on our conjecture that novel information is often conveyed through the introduction of new named entities, such as the names of people, organizations, and places. Using the notation of Figure 1, the NE metric can be defined as follows: Let $NE(R)$ be a set of named entities found in a set of documents R. Let $NE_u(R_1, R_2)$ be a set of *unique* named entities found in the set of documents R_1 and **not** found in the set R_2. That is, $NE_u(R_1, R_2) = \{e | e \in NE(R_1) \land e \notin NE(R_2)\}$. Then, $dist_{NE}(d, R) = NE_u(\{d\}, R)/length(d)$. Normalization by document length is essential, as, without normalization the NE score will tend to rise with length, because of the probabilistic influence of length on seeing additional named entities; the longer the document is, the higher the chance it contains more named entities.

These distance metrics can be harnessed to identify novel content for presentation to users. In the Newsjunkie application, we apply the novelty ranking algorithm iteratively to produce a small set of articles that a reader will be interested in. We employ a greedy, incremental analysis. The algorithm initially compares all the available updates to the *seed story* that the user has read, and selects the article least similar to it. This article is then added to the seed story (forming a group of two documents), and the algorithm looks for the next update most dissimilar to these articles combined, and so on. The pseudocode for the ranking algorithm is outlined in Figure 1.

Algorithm RANKNEWSBYNOVELTY($dist, seed, D, n$)
    $R \leftarrow seed$ // initialization
    for $i = 1$ to $\min(n, |D|)$ do
        $d \leftarrow \text{argmax}_{d_i \in D}\{dist(d_i, R)\}$
        $R \leftarrow R \cup \{d\}; D \leftarrow D \setminus \{d\}$
where $dist$ is the distance metric, $seed$ – seed story, D – a set of relevant updates, n – the desired number of updates to select, R - list of articles ordered by novelty.

Figure 1: Ranking news by novelty.

4. EMPIRICAL EVALUATION

To validate the algorithm presented in the previous section, we conducted an experiment that asked subjects to evaluate sets of news articles ordered by a variety of distance metrics.

4.1 Data

For the experiments described herein we used a live newsfeed supplied by Moreover Technologies[4], who aggregates news articles from over 4000 Internet sources. A simple clustering algorithm was used to group stories discussing the same events (called *topics* in the sequel). We used 12 clusters that correspond to topics reported in the news in mid-September 2003. The 12 topics covered news reports over a time span of 2 to 9 days, and represented between 36 and 328 articles. Topics included coverage of a new outbreak of SARS in Singapore, the California governor recall, the Pope's visit to Slovenia, etc. (Table 1 shows the full list of topics).

4.2 Description of the evaluation procedure

Judging novelty is a subjective task. One way to obtain statistically meaningful results is to average the judgments of a set of users. In order to compare different novelty-ranking metrics, we asked participants to read several sets of articles ordered by alternate metrics, and to decide which sets carried the most novel information. Note that this scenario requires the evaluators to keep in mind all the article sets they read until they rate them. Because it is difficult to keep several sets of articles on an unfamiliar topic in memory, we limited our experiment to evaluating the following three metrics:

1. The KL divergence was selected due to its appealing information-theoretic basis (KL).

2. The metric counting named entities was selected as a linguistically motivated alternative (NE).

3. The chronological ordering of articles was used as a baseline (ORG).

For each of the 12 topics, we selected the first story as the *seed story*, and used the three metrics described above to order the rest of the stories by novelty using the algorithm RANKNEWSBYNOV-ELTY (Figure 1). The algorithm first selects the most novel article relative to the seed story. This article is then added to the seed story to form a new model of what the user is familiar with, and the next most novel article selected. Three articles were selected in this manner for each of the three metrics and each of the 12 topics.

For each topic, the subjects were first asked to read the seed story to get background about the topic. They were then shown the three sets of articles (each set chosen by one of the metrics), and asked to rate the sets from most novel to least novel set. They were instructed to think of the task as identifying the set of articles that they would choose for a friend who had reviewed the seed story, and now desired to learn what was new. The presentation order of the sets generated by the three metrics was randomized across participants.

Originally, we had quite a few reservations about whether such an evaluation procedure would be feasible at all. Not only had the users to bear in mind quite a lot of information before pronouncing their decision, but the situation was further complicated by the varying *relevance* of articles to the seed story (we discuss this issue in detail in Section 5.3). The procedure described above was refined through a series of calibration experiments, in which we tried several techniques for eliciting judgments, and then thoroughly de-

---

Figure 2: Cumulative results.

Topic id	Topic description	#times most novel			Mean rank		
		KL	NE	ORG	KL	NE	ORG
topic1	Pizza robbery	5	4	1	1.7	1.6	2.7
topic2	RIAA sues MP3 users	2	7	0	1.8	1.2	3.0
topic3	Sharon visits India	2	3	4	2.6	1.7	1.8
topic4	Pope visits Slovakia	9	0	0	1.0	2.2	2.8
topic5	Swedish FM killed	5	4	0	1.4	1.6	3.0
topic6	Al-Qaeda	8	1	0	1.1	2.1	2.8
topic7	CA governor recall	4	2	3	1.7	2.2	2.1
topic8	MS bugs	3	5	1	1.9	1.6	2.6
topic9	SARS in Singapore	7	1	1	1.3	2.0	2.7
topic10	Iran develops A-bomb	3	5	2	2.2	1.7	2.1
topic11	NASA investigation	2	5	3	2.1	1.6	2.3
topic12	Hurricane Isabel	4	5	0	1.9	1.6	2.6

Table 1: Results by topic.

briefed the subjects after each experiment. One notable alternative would be to present the users with pairs of next articles, and ask them to give relative judgements as to which of the two articles carries more novel information given the seed story. We chose the set-based approach because it was more similar to the scenario we wanted to support, namely, given one article, to find a set of new articles to read. Furthermore, pairwise tests would also have involved more reading load for the participants.

4.3 Experimental results

Overall we obtained 111 user judgments on 12 topics, averaging 9–10 judgments per topic. Figure 2 shows the number of times each metric was rated the most, medium and least novel. As can be readily seen from the graph, the sets generated by the KL and NE metrics were rated more novel than those produced by the baseline metric (ORG).

Table 1 presents per-topic results. The three penultimate columns show the number of times each metric was rated the most novel for each topic. The last three columns show mean ranks of the metrics, assuming the most novel is assigned the rank of 1, medium novel – 2, and least novel – 3.

We used Wilcoxon Signed Ranks Test[5] [25] to assess the statistical significance of experimental results. Comparing the mean ranks of metrics across all the topics (as summarized in Figure 2), both KL and NE were found superior to ORG at $p < 0.001$. Considering individual per-topic results, the ORG metric never achieved the lowest (= best) rank of all three metrics. In 6 cases (topics 2, 4, 5, 6, 9, 12), the difference in mean rank between ORG and the lowest-scoring metric was statistically significant at $p < 0.05$, and in one additional case the significance was borderline at $p = 0.068$ (topic 8). Comparing the two best metrics (KL vs. NE), the difference in favor of KL was statistically significant at $p < 0.05$ for topics 4 and 6, and borderline significant ($p = 0.083$) for topic 9. The difference in mean ranks in favor of NE was borderline significant for topics 2 and 3 ($p = 0.096$ and $p = 0.057$, respectively).

The experiment we conducted should be considered a preliminary one as it only involved 12 different topics. Nevertheless, it allowed us to verify our evaluation procedure, as well as to observe the superiority of the two metrics tested (KL and NE) over the baseline (ORG). Based on these results, we do not observe significant

---

[5] This nonparametric test is used as an alternative to the familiar paired t-test when the underlying distribution cannot be assumed to be normal.

difference in performance between these two metrics. However, we believe that these preliminary results warrant further investigation of the potential content-sensitivity of the value of using KL versus NE metrics. In our future research we will attempt to characterize the properties of collections of articles on topics that could indicate where each metric performs the best for users.

5. PERSONALIZED NEWS UPDATES

Algorithm RANKNEWSBYNOVELTY presented and evaluated in the previous section works under the assumption that a user wants to catch up with latest story developments some time after initially reading about it. In this case the algorithm orders the recent articles by their novelty compared to the seed story, and then the user can read a number of highest-scoring articles depending on how much spare time he or she can allocate for the reading.

But what if the user wants to be updated continuously as the new developments actually happen? Some logistic support is needed to constantly keep track of the articles the user reads in order to estimate the novelty of the new articles streaming in the newsfeed. Based on user's personal preferences, that is, how often the user is interested in getting updates on the story, the server decides which articles to display. Therefore, an online decision mechanism is needed that determines whether any article contains sufficiently new information to warrant its delivery to the user. In a more general analysis of the benefits versus the costs of alerting, there are opportunities to balance the informational value of particular articles or groups of articles with the cost of interrupting users, based on a consideration of their context [14].

In what follows, we examine two scenarios of updating users with current news. The first scenario (discussed in Section 5.1) assumes the user is interested in getting updates once a day, while the second scenario (Section 5.2) updates the user continuously by monitoring incoming news for *bursts* of novel information. Finally, Section 5.3 introduces a mechanism that allows users to control the degree of *relevance* of the articles they are about to read.

5.1 Picking a single daily update

Let us first consider the simpler case when the user wants to see no more than a single daily update on the story.

One way to achieve this aim would be to use an algorithm similar to RANKNEWSBYNOVELTY, that is, accumulate the stories received on all the preceding days, and assess the novelty of each new story that arrived today by computing its distance from the accumulated set. The main problem with this approach is that the more stories are pooled, the less significant becomes the distance from any new story to the pool. After several days worth of articles have been accumulated, even a major update will be seen as barely

Algorithm PICKDAILYUPDATE($dist, Bg, D, thresh$)
$\quad d \leftarrow \text{argmax}_{d_i \in D}\{dist(d_i, Bg)\}$
$\quad$ if $dist(d, Bg) > thresh$ then $display(d)$
$\quad Bg \leftarrow D$
where $dist$ is the distance metric, Bg – the background reference set (union of all the relevant articles received on the preceding day), D – a set of new articles received today, $thresh$ – user-defined *sensitivity* threshold.

Figure 3: Picking daily updates.

new.

To avoid this pitfall, we modify our original algorithm as shown in Figure 3. Given the user and her choice of the topic to track, algorithm PICKDAILYUPDATE compares the articles received today with the union of all the articles received *the day before*. The algorithm picks the most informative update compared to what was known yesterday, and shows it to the user, provided that the update carries enough new information (*i.e.*, its estimated novelty is above the user's personalized threshold). Such conditioning endows the system with the ability to relay to the user truly informative updates and to filter out articles that only recap previously known details. The algorithm can be trivially generalized to identify n most informative updates per day.

It could be argued that by ignoring all the days before the immediately preceding one, algorithm PICKDAILYUPDATE might also consider novel those articles that recap what was said several days ago. In practice this rarely happens, as most of the articles are written in the way that interleaves new information with some background on previous developments. In our future work we plan to consider more elaborate distance metrics, that consider all previous articles relevant to the topic but decay their weight with age.

5.2 Reporting breaking news

The algorithm presented in Section 5.1 is still largely an "offline" procedure, as it updates users at predefined time intervals. Hard-core news junkies might find it frustrating to wait for daily scheduled news updates! For some, a more responsive form of analysis may be desired.

Taking the previous idea to the extreme and comparing every article to the preceding one will not work well, as the system will predict nearly every article as novel. Instead, we use a sliding window covering a number of preceding articles to estimate the novelty of the current one. Observe that estimating distances between articles and a preceding window of fixed-length facilitates the comparison of scores. We evaluated different window lengths of 20–60 articles. We found that lengths of approximately 40 typically worked well in practice.

In contrast to the algorithm PICKDAILYUPDATE, the background reference set now becomes much shorter, namely, 40 articles instead of a full day's content. This increases the likelihood that the window is not long enough to cover delayed reports and recaps that follow long after the story was initially reported. In order to filter out such repetitions, we first need to better understand the nature of news reports.

When an event or information update about an event of importance occurs, many news sources pick up the new development and report it within a fairly short time frame. If we successively plot the distance between each article and the preceding window, such arrival of new information will result in peaks in the graph. We call such peaks a *burst of novelty*. At the beginning of each burst, additional articles tend to add new details causing the graph to rise.

Algorithm IDENTIFYBREAKINGNEWS($dist, D, l, fw, thresh$)
$\quad Window \leftarrow \bigcup_{i=1}^{l} d_i \in D$
$\quad$ for $i = l + 1$ to $|D|$ do
$\quad\quad Scores_i \leftarrow dist(d_i, Window)$
$\quad\quad Window \leftarrow (Window \setminus d_{i-1}) \cup d_i$
$\quad Scores^{filt} \leftarrow MedianFilter(Scores, fw)$
$\quad$ for $j = 1$ to $|Scores^{filt}|$ do
$\quad\quad$ if $Scores_j^{filt} > thresh$ then
$\quad\quad\quad display(d_{j+l})$
$\quad\quad\quad$ *skip to the beginning of the next burst*
where $dist$ is the distance metric, D – a sequence of relevant articles, l – sliding window length, fw – median filter width, $thresh$ – user-defined *sensitivity* threshold.

Figure 4: Identifying breaking news.

As time passes, the sliding window covers more and more articles conveying this recent development and the following articles do not have the same novelty; as a result, the computed novelty heads downward signifying the end of the burst.
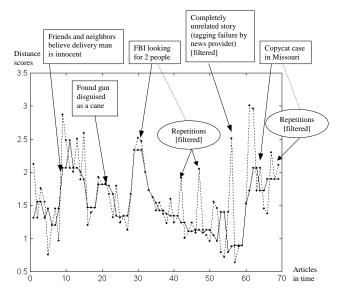
Delayed reports of events as well as recaps on a story are less likely to be correlated in time between different sources. Such reports may appear novel compared to the preceding window, but since they are usually isolated they only cause narrow spikes in the graph. In order to discard such standalone spikes and not to admit them as genuine updates, we need to filter the novelty signal appropriately.

The *median filter* [12] provides exactly this functionality by reducing the amount of noise in the signal. The filter successively considers each data point in the signal and adapts it to better resemble its surroundings, effectively smoothing the original signal and removing outliers. Specifically, a median filter of width w first sorts the w data points within the window centered on the current point, then replaces the latter with the median value of these points.

After computing the distance between every article and a sliding window covering the preceding ones, we pass the resultant signal through a median filter. We considered filters of width 3–7; the filter of width 5 appears to work well in the majority of cases.

We note that the use of a median filter may delay the routing of novel articles to users, as we need to consider several following articles to reliably detect the beginning of a new burst. However, we found that such delays are rather small (half the width of the median filter used), and the utility of the filter more than compensates for this inconvenience. If users are willing to tolerate some additional delay, the algorithm can scan forward several dozens of articles from the moment a burst is detected, in order to select the most informative update instead of simply picking the one that starts the burst. Combination approaches are also feasible such as the rendering of an early update on breaking news, and then waiting for a more informed burst analysis to send the best article on the development. Figure 4 shows the pseudocode the algorithm IDENTIFYBREAKINGNEWS that implements burst analysis for news alerting.

Figure 5 shows the application of the algorithm IDENTIFYBREAKINGNEWS to a sample topic. The topic in question is devoted to a bank robbery case in Erie, Pennsylvania, USA, where a group of criminals apparently seized a pizza delivery man, locked a bomb device to his neck and, according to statements made by the delivery man, forced him to rob a local bank. The man was promptly apprehended by police, but soon afterwards the device detonated and killed him. The bizarre initial story and ensuing investigation were tracked by many news sources for some two weeks in September 2003. The *x*-axis of the figure corresponds to the sequence of articles as they arrived in time, and the *y*-axis plots (raw

Dotted line represents raw distance scores, solid line — median-filtered scores.

Figure 5: Identifying breaking news — sample plot of raw and filtered novelty signals.

and median-filtered) distance values for each article given the preceding sliding window. Raw distance scores are represented by a dotted line, and filtered scores are plotted with a solid line. The text boxes accompanying the figure comment on the actual events that correspond to the identified novelty bursts, and show which potentially spurious peaks have been discarded by the filter. The smoothed novelty score, which incorporates the median filter, does a good job of capturing the main developments in the story (interviews with friends, details about the weapon, FBI bulletin for two suspects, and a copycat case), while at the same time filtering out spurious peaks of novelty.

5.3 Characterization of article types

When we debriefed several users who completed the experiments described in Section 4, several of them reported that it was difficult to judge the novelty of articles because of their varying *relevance* to the seed story. In some cases this had to do with errors in the tagging of the news stories by the newsfeed we relied upon, while in one or two extreme cases a failure of the Moreover parser caused grossly unrelated articles to be glued to the relevant ones. In other cases the variance in relevance was due to the differences in writing styles and policies among different publications.

These comments led us to believe that the novelty scores we compute should not be relied upon as a sole selection criterion; some articles are identified as novel by virtue of changing the topic. To further refine the analysis of informational novelty, we have formulated a classification of types of novelty, based on different relationships between an article and a seed story. These classes of relationships include:

1. **Recap** articles are those that are clearly relevant, but only offer reviews of what has already been reported and carry little new information.

2. **Elaboration** articles add new, relevant information on the topic set forth by the seed article.

3. **Offshoot** articles are also relevant to the mainstream discussion, but the new information they add is sufficiently different from that reported in the seed story to warrant a new trail.

4. **Irrelevant** articles are those due to clustering and parsing mistakes.

Of these classes, relationship types 2 and 3 are probably what most users want to see. But how can we identify them automatically, and how can we empower the users themselves to exercise control over this spectrum?

To achieve this aim we suggest a new type of document analysis that scrutinizes *intra-document dynamics*. As opposed to previous kinds of analysis that compared entire documents to one another, the new technique "zooms into" documents estimating the relevance of their parts.

We start with building a language model for every document, and fix a distance metric to use, *e.g.*, KL divergence. Then, for each document, we slide a window over its words and plot the distance scores of each such window versus the seed story. We construe the score of a window of words as a sum of pointwise scores of each word vs. the seed story, as stipulated by comparing the language model of the current document with that of the seed story using the selected metric. Several different window lengths were considered, and the value of 20 was found to work well in practice.
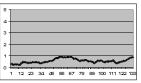
An important property of this technique is that it goes beyond the proverbial *bag of words*, and considers the document words in their original *context*. We opted for using sliding contextual windows rather than apparently more appealing paragraph units, since using a fixed-length window makes distance scores directly comparable. Another obvious choice of the comparison unit would be individual sentences. However, we believe that performing this analysis at the sentence level would consider too little information, and the range of possible scores would be too large to be useful. A recent study in novelty detection [13] corroborates this reasoning — when the importance of individual sentences (deemed relevant) is considered, 93% of them are classified as carrying novel information.

Figure 6 shows sample results of intra-document analysis. The seed story for this analysis was a report on a new case of SARS in Singapore. Articles that mostly **recap** what has already been said typically have a very limited dynamic range and low absolute scores. **Elaboration** articles usually have higher absolute scores that reflect the new information they carry. One elaboration for this story reported that the patient's wife was being held under quarantine. Further along this spectrum, articles that may qualify as **offshoots** but are still anchored to the events described in the seed story have a much wider dynamic range. One offshoot was a story that focused on the impact of SARS on the Asian stock market, and another was on progress on a SARS vaccine. Both offshoot articles used the recent case as a starting point, but were really about a related topic. We believe that analyzing intra-document dynamics such as the dynamic range and patterns of novelty scores are useful in identifying different types of information that readers would like to follow.
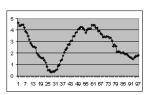
6. CONCLUSIONS AND FUTURE WORK

The Web has been providing users with a rich set of news sources. It is deceptively easy for Internet surfers to browse multitudes of sources in pursuit of news updates, yet sifting through large quantities of news can involve the reading of large quantities of redundant material.
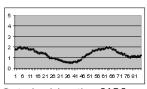
We presented a collection of algorithms that analyze live newsfeeds and identify articles that carry most novel information given a model of what the user has read before. To this end, we extend
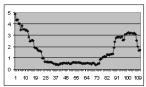
On-topic, recap

On topic, elaboration: **SARS patient's wife held under quarantine**

Offshoot: **SARS impact on Asian stock markets**

Offshoot: **Swiss company developed SARS vaccine**

The x-axis shows contextual word windows ending at word i; the y-axis plots the distance from each window to the seed story.

Figure 6: Examples of relationships of articles with a seed story.

the conventional bag of words representation with named entities extracted from the text, and use a variety of distance metrics to estimate the dissimilarity between each news article and a collection of the previously read ones. The techniques underlying the algorithms analyze inter- and intra-document dynamics by studying how the delivery of information evolves over time from article to article, as well as within each individual article at the level of contextual word windows.

News browsers incorporating these algorithms can offer users a personalized news experience, giving users the ability to tune both the desired frequency of news updates and the degree to which these updates should be similar to the seed story, via exercising control over the novelty constraint.

To evaluate the algorithm for ranking news articles by novelty, we propose a new evaluation scheme that asks users to read several sets of articles ordered by different metrics, and rate them from the most to least novel. Although at first glance this task appears very hard as it requires the users to keep in mind all the articles they read, in practice the scheme was feasible. Debriefing the users after the experiments offered interesting insights into how people judge novelty and relevance issues.

This research can be extended in several directions. We plan to investigate more sophisticated distance metrics that incorporate some of the basic metrics we described herein, as well as use complex weighting windows that vary document weights over time. We also plan to combine information-theoretic measures such as KL and JS with a representation based solely on named entities, to estimate the amount of novelty carried by the latter in a more principled way. Of particular interest and practical use is the characterization of which metrics are best for predicting the novelty of articles for different types of topics. The notion of intra-story patterns of novelty is a rich area for exploration. In this work we only scratched the surface by modeling a few article types of particular interest in the context of news. However, we believe a much richer ontology would be necessary to capture all the possible variety of existing article types. We hope that further research in these directions will provide additional insights into principles and applications for personalizing news.

Finally, we observe that techniques similar to those we described can also be used for other types of content, such as blogs and news-

groups, assessing the novelty of postings within threads or topic-focused discussions. Although technically possible, we believe that using these techniques for estimating the novelty of email messages would be less useful. This is because in personal email information is usually transmitted in an extremely compact way, so that even the most novel piece of information may be conveyed in a single word or only a few words. Therefore, much deeper language understanding is required to adequately reflect the importance of information, while more statistically-oriented approaches relying on a bag of words might overlook such distinctions.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] J. Allan, V. Khandelwal, and R. Gupta. Temporal summaries of news topics. In *Proceedings of the 24th International Conference on Research and Development in Information Retrieval*, pages 10–18, 2001.

[2] C. Burgess, K. Livesay, and L. Kevin. Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25(2&3):211–258, 1998.

[3] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st International Conference on Research and Development in Information Retrieval*, pages 335–336, 1998.

[4] K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan. Information filtering, novelty detection, and named-page finding. In *Proceedings of the 11th Text REtrieval Conference*. National Institute of Standards and Technology, 2002.

[5] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[6] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. In *Proceedings of the Human Language Technology Conference (HLT-2002)*, pages 94–98, March 2002.

[7] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th International Conference on Research and Development in Information Retrieval*, pages 299–306, August 2002.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society For Information Science*, 41:391–407, 1990.

[9] F. Douglis, T. Ball, Y.-F. Chen, and E. Koutsofios. The AT&T internet difference engine: Tracking and viewing changes on the web. *World Wide Web*, pages 27–44, January 1998.

[10] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[11] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, January 2002.

[12] R. Gonzalez and P. Wintz. *Digital Image Processing*. Addison-Wesley, second edition, 1987.

[13] D. Harman. Overview of the TREC 2002 novelty track. In *Proceedings of the 11th Text REtrieval Conference*, pages

46–55. ACM Press, 2002. NIST Special Publication 500-251.

[14] E. Horvitz, C. Kadie, T. Paek, and D. Hovel. Models of attention in computing and communication: From principles to applications. *Communications of the ACM*, 46(3):52–59, March 2003. `http://research.microsoft.com/˜horvitz/cacm-attention.htm`.

[15] A. Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133, 2001.

[16] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, 2002.

[17] L. Lee. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the ACL*, 1999.

[18] C. D. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 2000.

[19] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[20] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[21] E. Ristad. A natural law of succession. Technical Report Technical Report CS-TR-495-95, Princeton University, 1995.

[22] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[23] B. Schiffman, A. Nenkova, and K. McKeown. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference (HLT-2002)*, March 2002.

[24] R. Swan and D. Jensen. Timemines: Constructing timelines with statistical models of word usage. In *Proceedings of the ACM SIGKDD 2000 Workshop on Text Mining*, pages 73–80, 2000.

[25] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[26] Y. Yang, T. Ault, and T. Pierce. Combining multiple learning strategies for effective cross-validation. In *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 1167–1182, 2000.

[27] Y. Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned novelty detection. In *Proceedings of the Internaltional Conference on Knowledge Discovery and Data Mining*, pages 688–693, 2002.