Mining the Web to Predict Future Events

Kira Radinsky
Technion–Israel Institute of Technology
Haifa, Israel
kirar@cs.technion.ac.il

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

ABSTRACT

We describe and evaluate methods for learning to forecast forthcoming events of interest from a corpus containing 22 years of news stories. We consider the examples of identifying significant increases in the likelihood of disease outbreaks, deaths, and riots in advance of the occurrence of these events in the world. We provide details of methods and studies, including the automated extraction and generalization of sequences of events from news corpora and multiple web resources. We evaluate the predictive power of the approach on real-world events withheld from the system.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems

General Terms

Algorithms, experimentation

Keywords

News corpora, predicting future news, learning from Web content

1. INTRODUCTION

Mark Twain famously said that "the past does not repeat itself, but it rhymes." In the spirit of this reflection, we develop and test methods for leveraging large-scale digital histories captured from 22 years of news reports from the New York Times (NYT) archive to make real-time predictions about the likelihoods of future human and natural events of interest. We describe how we can learn to predict the future by generalizing sets of specific transitions in sequences of reported news events, extracted from a news archive spanning the years 1986–2008. In addition to the news corpora, we leverage data from freely available Web resources, including Wikipedia, FreeBase, OpenCyc, and GeoNames, via the LinkedData platform [6]. The goal is to build predictive models that generalize from specific sets of sequences of events to provide likelihoods of future outcomes, based on patterns of evidence observed in near-term newsfeeds. We propose the methods as a means of generating actionable forecasts in advance of the occurrence of target events in the world.

The methods we describe operate on newsfeeds and can provide large numbers of predictions. We demonstrate the predictive power of mining thousands of news stories to create classifiers for a range of prediction problems. We show as examples forecasts on three prediction challenges: proactive alerting on forthcoming disease outbreaks, deaths, and riots. These event classes are interesting in serving as examples of predictions that can serve as heralds for attention for guiding interventions that may be able to change outcomes for the better. We compare the predictive power of the methods to several baselines and demonstrate precisions of forecasts in these domains ranging from 70% to 90% with a recall of 30% to 60%.

The contributions of this work include automated abstraction techniques that move the level of analysis from specific entities to consideration of broader classes of observations and events. The abstractions enlarge the effective sizes of training sets by identifying events as members of more general sets of evidence and outcomes at higher-levels of ontological hierarchies. For example, we can learn from news data about events in specific countries (e.g., Angola and Rwanda) to build classifiers that consider the likelihood of events on a continent (e.g., Africa) or to regions characterized by particular demographic and geological properties. The knowledge that Angola and Rwanda are elements of the broader set of countries comprising Africa is extracted from LinkedData.

As an example, the learning and inference methods can be used to provide alerts about increases in the likelihood of a forthcoming cholera outbreak within a specified horizon. Cholera is a fast-paced infection causing over 100,000 deaths per a year, with a mortality rate exceeding 50% for people with the ailment who do not receive treatment. With prompt rehydration therapy, the mortality rate drops to less than 1%. Alerts about inferred jumps in the likelihoods of future cholera outbreaks based on the monitoring of news stories could assist with the triaging of attention and planning effort. For example, inferred likelihoods of a cholera outbreak over specific periods of time could guide proactive designs for distributing fresh water in areas at raised risk. The methods we describe might one day be used to continue to monitor evolving news stories and to provide automated

WSDM'13, February 4–8, 2013, Rome, Italy.

alerting about the raised likelihood of outcomes of interest. Such predictions could serve as adjuvants to existing monitoring and communication services, such as the World Health Organizations (WHO) Global Alert and Response (GAR) system for coordinating responses to public health emergencies[1]. In tests, we found that automated predictions would have provided an alert a week in advance of several of the outbreaks of cholera (Figure 6).

Experts such as epidemiologists who explore the relationships between the spread of disease and natural disasters make similar inferences. However, such studies are typically few in number, employ heuristic assessments, and are frequently retrospective analyses, rather than aimed at generating predictions for guiding near-term action. In contrast, a computational system has the ability to learn patterns from large amounts of data, can monitor numerous information sources, can learn new probabilistic associations over time, and can continue to do real-time monitoring, prediction, and alerting on increases in the likelihoods of forthcoming concerning events. Beyond knowledge that is easily discovered in studies or available from experts, new relationships and context-sensitive probabilities of outcome can be discovered by a computational system with long tentacles into historical corpora and real-time feeds. As an example, the methods we describe identified a relationship in Angola between droughts and storms that, in turn, catalyze cholera outbreaks. Alerts about a downstream risk of cholera could have been issued nearly a year in advance (Figure 1). Human experts who focus on shorter time horizons may overlook such long-term interactions. Computational systems can consider multiple time granularities and horizons in pursuing the probabilistic influences among events. Beyond alerting about actionable situations based on increased likelihoods of forthcoming outcomes of interest, predictive models can more generally assist by providing guidance when inferences from data run counter to expert expectations. It can be valuable to identify situations where there is a significantly *lower likelihood* of an event than expected by experts based on the large set of observations and feeds being considered in an automated manner. Finally, a system monitoring likelihoods of concerning future events typically will have faster and more comprehensive access to news stories that may seem less important on the surface (e.g., a story about a funeral published in a local newspaper that does not reach the main headlines), but that might provide valuable evidence in the evolution of larger, more important stories (e.g., massive riots).

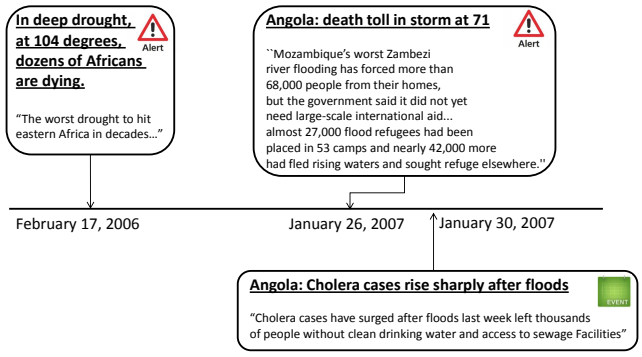

Figure 1: Example of likelihood of cholera rising after a drought followed by storms in Angola. Triangular alert icons represent inferences of significant upswings in likelihood of a cholera outbreak occurring within several days.

2. EVENT PREDICTION

We assume that events in the real-world are generated by a probabilistic model that also generates news reports corresponding to these events. We use the text of news stories to build an inferential model of the form $P\big(ev_j(\tau+\Delta)\big|ev_i(\tau)\big)$ for some future event ev_j at time $\tau+\Delta$ and past event ev_i happening at time τ (e.g., today). For example, the model learns that the probability of a news report about a drought (ev_j) happening after a news report about a flood (ev_i) to be 18%. This probability approximates the relationship between the two real-world events.

Given a target future event (such as cholera outbreak), calculating this probability for every possible future time $\tau+\Delta$ and every possible ev_i is an intractable problem. We simplify the analysis by focusing on a small subset of event sequence candidates that may be causally linked, and define sets of events ev_i that are linked to target events ev_j in this manner. In particular, we define and extract from the NYT archive news *storylines*—sets of topically cohesive ordered segments of news that include two or more declarative independent clauses about a single story. As an example, the following events form a storyline: {(drought in Africa, 02/17/2006), (storm in Rwanda, 01/26/2007), (flood in Rwanda, 01/27/2007), (cholera outbreak in Rwanda, 01/30/2007)}. We then use such storylines as a heuristic for identifying possible causal relationships among events. The process is performed by clustering news stories with similar text and semantic entities, as detailed in Section 2.1.

We show a componentized view of the method in Figure 2. At the start of the learning phase, the system mines the NYT news corpora and extracts storylines, using techniques adapted from well-known topic tracking and detection algorithms [8, 3, 7], that cluster similar texts together (Section 2.1). We next enrich the storylines with information extracted from Web knowledge sources via the Linked-Data project(Section 2.2). We extract a wide variety of facts, including such information as the population density in Rwanda, percentage of land in Rwanda covered by water, and the gross domestic product of the country. We generalize both features and events to increase the number of equivalent samples for constructing predictive models (Section 2.3). For example, we can learn from data about events in specific countries (e.g., Angola and Rwanda) to build classifiers that consider the likelihood of events of interest on a larger scale (e.g., larger continent of Africa) or to regions characterized by particular demographic and geological properties. At the end of the learning phase, the system estimates the probabilities $P\big(ev_j(\tau+\Delta)\big|ev_i(\tau)\big)$ and builds a probabilistic classifier for use in the prediction phase. The classifier can be used to provide real-time probabilities of events of interest, such as an impending "cholera outbreak in Angola" based on the previous knowledge obtained in a storyline about Angola or its generalization, Africa. The classifier we construct provides binary predictions of whether

[1] http://www.who.int/csr/alertresponse/en/

an event will occur following an observed event sequence. In experiments, we also evaluate both the occurrence of the target event and the mean time between the prediction and occurrence. We show results on providing such alerts nearly three weeks prior to the actual predicted event. We leave the prediction of the exact date of events of interest based on specific dynamics to future work.

2.1 Extracting Event Chains

We define and extract news storylines from the NYT archive as a heuristic for identifying potential causal relationships among events. A storyline is a set of topically cohesive ordered segments of news that includes two or more declarative independent clauses about a single story. As an example, a story line about the arrest of Carlos the Jackal includes the stories about verification of his identity, his transport to prison, and so on. Methods for extracting such storylines are referred to as *topic detection and tracking* (TDT) [8]. Topic detection involves identifying a series of linked events in streams of stories. To identify storylines, we modified the Inc.LM method, an approach to topic tracking found to be most successful for this task in several competitions [3]. Consider $Chains \in 2^{2^{|T| \times Time}}$ as the set of all possible storylines, where T is all the news articles and $Time$ is a discrete representation of time. We denote with $t_1 <<_c t_2$ an event represented by the news article t_1 occurring before an event represented by the news article t_2 in a chain $c \in Chains$. We use the notation $t(\tau)$ to represent an event as defined by the appearance of the text t of a news story at time $\tau \in Time$. Under the assumption that causality occurs only within storylines, the prediction challenge is reduced to calculating the probability $P\big(t(\tau_i > \tau) \big| t_j(\tau)\big)$ for $\{t_j | \exists c \in Chains, t_j <<_c t\}$.

Similar to other vector space approaches for topic detection [7], we first cluster documents with similar text. We consider news articles as documents and represent each news article as a vector $(\sigma_1^t \dots \sigma_n^t)$, such that

$$\sigma_i^t = \mathrm{tf}_{w,t} \cdot \log \frac{|T|}{|\{t' \in T \,|\, w_i \in t'\}|},$$

where $|T|$ is all the news articles, and $\mathrm{tf}_{w,t}$ is the frequency of the word w in article t. We then perform a nearest-neighbor analysis, where we find for each article the k closest (in our experiments $k = 50$) articles to it using a cosine similarity measurement, defined as

$$\mathrm{sim}(t_a, t_b) = \frac{\sum_{i=1}^{N} \sigma_i^{t_a} \sigma_i^{t_b}}{\sqrt{\sum_{i=1}^{N} {\sigma_i^{t_a}}^2} \sqrt{\sum_{i=1}^{N} {\sigma_i^{t_b}}^2}},$$

with a constraint on temporal proximity. Articles are either generated within a threshold time horizon or the date of an article is mentioned in a text of a more recent article in the chain. We performed several experiments using the time threshold on the TDT4 corpus[2], and reached our best performance when limiting the chains to 14 days. This type of analysis has a high recall for identifying articles that cover the same topic, referring to the fraction of relevant instances that are retrieved. However, the procedure has a low precision; a large fraction of the identified instances are false positives. We wish to enhance the precision, while maintaining the high recall. As an approach to reducing the false positives, we overlay a preference that the entropy of the entities $\{e \in Entities\}$ of the story articles C, defined as

$$StoryEntropy(C) = -\sum_{i=1}^{n} P(e_i \in C) \log P(e_i \in C)),$$

grows "slowly" as the story evolves over time. A similar approach has been shown to provide major improvements on a related topic-detection task [2]. We employ conditional random fields (CRF), trained on a heterogeneous corpus [9], to identify entities of the types *location*, *people*, and *organizations*. We define a vector of counts of entities and operations of addition and removal of an article from a chain. We use a greedy algorithm that selects at each step the next best document to add or decides to halt if all remaining documents increase the entropy by more than a threshold amount α, which we evaluate from a validation set. We performed experiments showing that this extension improves precision while maintaining levels of recall (see Section 3.4).

We performed the latter process on an offline corpus. We note that studies have addressed the usage of similar techniques for extraction from online streams of news (e.g., [2]). Such online approaches could be adapted in our approach to forecasting future events.

2.2 Lexical and Factual Features

We seek to infer the probability of a predefined future news event of interest given a vector representing the news events occurring up to a certain time. To perform this task, we create training cases for each target event, where each case is represented using a set of observations or features. We define both *lexical* and *factual* features. We set the label for each case as true only if the text representing the future target event occurs in a document dated at a later time in the chain.

Let $w_1 \dots w_n$ be the words representing the concepts of the event at time τ and let $a_1 \dots a_m$ be additional real-world characteristics of the event concepts. We refer to these attributes respectively as *lexical and* factual features. The words w_i are extracted from the text of each news article using the Stanford Tokenizer, and filtered using a list of stop words. The factual characteristics a_i are extracted from the different LinkedData sources, specifically the properties under the type rdf:Property for the event concepts w_i. For example, given the text of the news story title, "Angola: Cholera Cases Rise Sharply After Floods," the feature vector contains both the tokenized and filtered words of the text (Angola, cholera, rise, sharp, flood), and other features describing characteristics of Angola (GDP, water coverage, population, etc.). We map each concept in the event text to a LinkedData concept. If several concepts are matched, we perform disambiguation based on the similarity between the concept text (e.g., its Wikipedia article) and the news article, using a bag-of-words representation.

We denote $f_1(ev) \dots f_{n+m}(ev)$ to be the features of the event ev (either lexical or factual features). We make a naive simplifying assumption that all features are independent, and describe the probability $P(ev_j(\tau + \Delta) \big| ev_i(\tau))$ as follows:

$$P(ev_j(\tau + \Delta) \big| ev_i(\tau)) \propto \prod_{k=1}^{n+m} P\big(ev_j(\tau + \Delta) \big| f_k(ev_i(\tau))\big).$$

[2] http://www.nist.gov/TDT

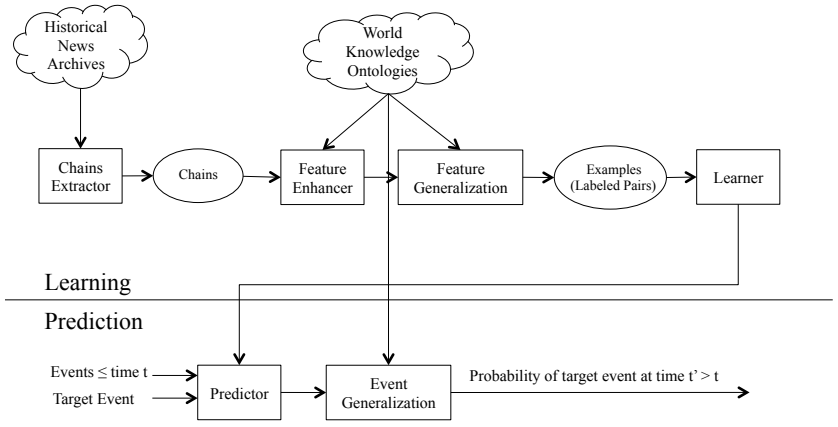

Figure 2: Main components and flow of analysis of event prediction pipeline.

Using Bayes rule, we can derive that

$$P\big(ev_j(\tau+\Delta)\big|f_k(ev_i(\tau))\big) = \frac{P\big(ev_j(\tau+\Delta), f_k(ev_i(\tau))\big)}{P(f_k(ev_i(\tau)))},$$

where $P\big(ev_j(\tau+\Delta), f_k(ev_i(\tau))\big)$ is evaluated from the data by counting how many times the event ev_j happens in the same storyline after an event having the value of feature f_k of the event ev_i at time τ. Similarly, $P(f_k(ev_i(\tau))$ is evaluated from the data by counting the portion of times that an event with this feature value happens in the corpus. We build a predictor based on these learned probabilities. The predictor outputs the probability that each future event represented in the system will occur. Changes in this probability can be noted and thresholds set for alerting.

We may be interested in predicting various characteristics of a future event as *scalar* values. For example, beyond predicting the likelihood that deaths will be associated with a later accident or disruption occurring within a horizon, we may wish to predict the number of people who will perish given the occurrence of a target event that causes or is associated with deaths. To do this, we bin target predictions of numbers of deaths into a set of buckets capturing a mutually exclusive and exhaustive set of ranges of numbers of deaths, e.g., less than ten deaths, greater than 10 but less than 100 deaths, and greater than 100 deaths. We say that ev_j belongs to bin_k, if ev_j holds the bin_k relation. As an example, for the less than 10 deaths bin, we say that $ev_j \in bin_{0-10}$ if the text representing ev_j contained text indicating how many people died and this number was less than 10. We learn predictors that estimate the probability of the event to belong to a certain bin k, $P\big(ev_j(\tau+\Delta), ev_j(\tau+\Delta) \in bin_k\big|ev_i(\tau)\big)$, and output the bin with the highest probability. We present results on a study of the accuracy of inferring the number of deaths caused by an event in Section 3.2.

2.3 Learning to Predict with Abstractions

Event sequences are relatively sparse in the domains we explored. For example, the target event "Rwanda cholera outbreak" appeared only 33 different times in the news archive. Such sparsity may degrade classifier performance due to the poor estimation of both $P\big(ev_j(\tau+\Delta), f_k(ev_i(\tau))\big)$ and $P(f_k(ev_i(\tau)))$. In other cases, the feature values might not appear with high enough frequency in the data for deriving high-confidence inferences about the future. For example, there is not even a single mention of the African country of Comoros in the news corpus. Such sparsity in lexical features can lead to poor estimation of the probability $P\big(ev_j(\tau+\Delta), f_k(ev_i(\tau))\big)$, and therefore to poor predictors. As an example, for the target prediction of the likelihood of a forthcoming large-scale evacuation, a hurricane originating in Comoros might be important information for predicting a storm in nearby countries which might be useful in predicting evacuations in those countries. Similarly, if the system is focused on making predictions about upcoming storms in Comoros, i.e., "storm in Comoros" is the target event, there may not be enough data to evaluate the aforementioned probabilities.

We address the joint challenges of event and feature sparsity via employing procedures for automated abstraction. Instead of considering only "Rwanda cholera outbreak," an event with a small number of historical cases, we consider more general events of the form: "[Country in Africa] cholera outbreak." We turn to world knowledge available on the Web. Some LinkedData resources provide hierarchical ontologies. For example, Fabian et al.[24] created an *isA* ontology from Wikipedia content. This ontology maps Rwanda to the following concepts: Republics, African countries, Landlocked countries, Bantu countries, etc. Similarly, WordNet provides hypernym relations, that map Rwanda to the concept *country*.

We developed a an automated method for guiding abstraction. The method determines when to generalize events and features. As features are evaluated separately, estimations are made about the value of each feature abstraction to enhance the accuracy of predicting the target event. We evaluate for each feature and its abstractions the precision over the training data using cross validation. We note that it

is insufficient to measure the precision associated with using an abstracted feature without altering the target event. Consider the abstracted feature [Country in Africa], and the target event "Death in Kigali." The probability of a death in Kigali, the capital of Rwanda, caused by an event in some country in Africa, is small. Therefore, given an event in [Country in Africa], the probability of death being caused by the event in CapitalOf([Country in Africa]) may typically be more appropriate. We now formalize this intuition.

Let the semantic network graph G be an edge-labeled graph, where each edge is a triplet $\langle v_1, v_2, l\rangle$ and l is a predicate (e.g., "CapitalOf"). We look for a path of maximum length k (in our experiments $k = 3$)that connects the concept representing the abstraction and the concepts described in the target event. For example, given a chain where the first event discusses the large attendance at the opera "The Nose," and the second event discusses an award that the opera writer Dmitri Shostakovich receives, we find the following path in the Wikipedia graph, connecting the articles: "The Nose"$\xrightarrow{OperasBy}$Dmitri Shostakovich. Later, we can use a similar observation, observing a large attendance at the opera "The Murder of Comrade Sharik," to predict an award for the opera writer William Bergsma using the path "The Murder of Comrade Sharik"$\xrightarrow{OperasBy}$William Bergsma. Given two events' concepts c_1, c_2, represented by the nodes v_1 and v_2 in G, we call the labels of the k-sized path, connecting v_1 and v_2, an *abstraction path* $abs(c_1, c_2) = l_1, \ldots, l_k$. *Applying an abstraction* on a node v determines the node v' that satisfies the abstraction path, i.e.,
$ApplyAbs(v, abs(c_1, c_2)) = v'$,
s.t $\exists v_i \in V(G)(v, v_1, l_1) \ldots, (v_{k-1}, v', l_k) \in E(G)$.

When inferring probabilities for each entity en in the target event and a feature of the causing event, we iteratively abstract each feature f to a more general concept $gen(f)$, using a semantic hierarchical graph G^H, calculating the abstraction path $abs(f, en)$ (based on the semantic graph G), and instead of $P\big(ev_j(\tau + \Delta), f_k(ev_i(\tau))\big)$, we calculate the probability for the more abstract event,

$$P\left(ApplyAbs\Big(ev_j(\tau + \Delta), abs(f_k, en)\Big), gen(f_k)(ev_i(\tau))\right).$$

A similar process is conducted when generalizing the target event. In this case, the probabilities are calculated for the abstracted target event, and the precision is calculated on the concrete target event. For each entity en in the target event and a feature of the causing event, we wish to iteratively abstract each feature f to a more general concept $gen(f)$, using a semantic hierarchical graph G^H (in our experiments we used the IsA and InCategory relations).

Figure 3 shows pseudocode for the abstraction process. Given a target and a possible causative event, the goal of the procedure is to estimate the probability of the causative event or any of its abstractions to cause the target event. The algorithm is given as input several parameters: a target event (e.g., cholera Outbreak in Rwanda), denoted as $target$, and an event occurring at time τ ($cause$), the storylines the system extracted, denoted as $Chains$, the hierarchical graph G^H, the semantic graph (G), and some parameter specifying a maximum degree of abstraction (k). The system evaluates the probability

$$P\left(ApplyAbs\Big(ev_j(\tau + \Delta), abs(f_k, en)\Big), gen(f_k)(ev_i(\tau))\right).$$

At stages 1-2, the system builds a classifier estimating the probability that any of the entities of the lexical features of the causative event precede an appearance of the target event in a text of an event in a storyline. For example, the $bestClassifier$ at this stage will have estimations of the probability of "cholera Outbreak in Rwanda" given the entity Kigali (Rwanda's capital). At stage 3, the algorithm iteratively estimates the probability of the target event happening given any of the abstracted features (extracted using the hierarchical graph G^H). For example, one iteration can be the evaluation of the number of times the entity "Capital of Africa" preceded "cholera Outbreak in Rwanda" in our storylines. Stages 3.1-3.2 evaluate the needed transformations to the target event given the abstracted cause entity. For example, instead of looking for cases where an event with an entity belonging to "Capitals in Africa" occurred and an event regarding "cholera Outbreak in Rwanda" followed, we look for examples where an event of the type "cholera Outbreak in Africa" followed. We then train and evaluate new classifier using the transformed training data. If its performance, as measured by cross validation on the training data, is superior to that of the classifier in advance of the abstraction, we update the best classifier found.

3. EXPERIMENTAL EVALUATION

We now describe the experiments that we conducted to test the methodology, and present the results of the studies of inferences performed on a test portion of the news archive held out from the training phase.

3.1 Experimental Setup

In this Section we outline the data we obtained for the experiments, the experimental methodology, and the baselines we compared against.

3.1.1 Data

We crawled and parsed the NYT archive containing news articles for the years 1986–2007. We say that a chain of events belongs to a domain D, if it consists one of the domain relevant words, denoted as $w_i(D)$. For example, for the challenge of predicting future deaths, we consider the words "killed," "dead," "death," and their related terms.[3] For the challenge of predicting future disease outbreak, we consider all mentions of "cholera, " "malaria, " and "dysentery."

During prediction, we hold out from the learning phase a test set of a decade of events for the period of 1998–2007 (the *test period*). We say that a chain is a *test-domain chain* if (1) the dates of all of its events occurred in the test period dates, and (2) the first chronological event in the chain does not contain one of the domain terms, e.g., the first event did not contain a mention of death (otherwise the prediction might be trivial). Formally, let $C = \{e_1 \ldots e_k\}$ be a test chain, thus $\forall i : w_i(D) \notin e_1$.

[3]We consider all the similarity relations in Wordnet: Synonyms, pertainyms, meronyms/holonyms, hypernyms/hyponyms, *similar to*, *attribute of*, and *see also* relations.

Procedure ABSTRACT($target, cause, Chains, G^H, G, k$)
(1) **Foreach** $\{entity \in Entities(cause)\}$
(1.1) $PositiveExamples \leftarrow \{(ev_1, ev_2) | ev_1 <<_{c \in Chains} ev_2, entity \in ev_1, \forall e \in Entities(target) : e \in ev_2\}$
(1.2) $NegativeExamples \leftarrow \{(ev_1, ev_2) | ev_1 <<_{c \in Chains} ev_2, entity \in ev_1, \exists e \in Entities(target) : e \notin ev_2\}$
(2) $bestClassifier \leftarrow Build(PositiveExamples, NegativeExamples)$
(3) **Foreach** $\{entity \in Entities(cause), absEntity \in Abstractions(entity, G^H)\}$
(3.1) $absPaths \leftarrow FindPaths(absEntity, Entities(target), G, k)$
(3.2) $absTargets \leftarrow ApplyAbs(absEntity, absPaths, G)$
(3.2) **Foreach** $absTaret \in absTargets$
(3.2.1) $PositiveExamples \leftarrow \{(ev_1, ev_2) | ev_1 <<_{c \in Chains} ev_2, absEntity \in ev_1, \forall e \in Entities(absTarget) : e \in ev_2\}$
(3.2.2) $NegativeExamples \leftarrow \{(ev_1, ev_2) | ev_1 <<_{c \in Chains} ev_2, absEntity \in ev_1, \exists e \in Entities(absTarget) : e \notin ev_2\}$
(3.2.3) $absClassifier \leftarrow Build(PositiveExamples, NegativeExamples)$
(3.2.4) **If** $CV(bestClassifier, Chains) < CV(absClassifier, Chains)$
$bestClassifier \leftarrow Update(absClassifier)$
(4) **Return** bestClassifier

Figure 3: Procedure for generalizing features via abstraction. *Build* takes as input positive and negative examples and estimates the probability of our target event. *FindPaths* finds all predicate paths of size k between two nodes in the graph given as input. *ApplyAbs* applies the predicate path on a node, returning nodes that are connected to the given node via the predicates of the directed paths. *CV* calculates the precision via cross validation of a classifier on the training data.

3.1.2 Experimental Methodology

For each prediction experiment we first select a target event e_{target} from a test-domain chain. The procedure differs depending on the type of the experiment:

1. Predicting general events in the period 2006–2007. In this type of experiment, a target event is any news headline published during 2006–2007, i.e., we build a classifier for each possible headline.

2. Predicting events in specific three domains: deaths, disease outbreaks, and riots. In this case, any news story containing one of the domain words is selected. Additionally, we validate manually that those events actually contain an event from the domain. If several of the target events exist, we choose the first one appearing chronologically to be the identified target event, i.e., $e_{target} = argmin_j\{e_j | \exists i : w_i(D) \in e_j\}$. As e_{target} is selected from a test-domain chain $j > 1$, i.e., it is not the first event in the chain. That is, we consider only event chains that are not observed by the system during the years 1998–2007, and do not contain words implying the target event within a domain (e.g., the word death) during the first event chain. The first event of the chain is given as input to the system.

In summary, the general events predictions represents prediction of *all* the events in 2006-2007. The system is given an event from 2006–2007 as input, and we measure the success in predicting the event. For the domain-specific predictions (death, disease outbreak, and riots), we manually check to see if the event occurs using the domain representative words or their synonyms as filters. We consider only event chains that are not observed during the years 1998–2008, and do not contain words implying the target event within a domain (e.g., the word death) during the first event chain. The first event of the chain is given as an input.

We train from the data via evaluating the probabilities of e_{target} happening for events occurring up until the date of the first event in the chain. During the test, the algorithm is presented with the first event of the chain e_1 and outputs its prediction about e_{target}. In the experiments, we consider the predictor as indicating that the target event will occur if

$$P(e_{target} | e_1) > P(\neg e_{target} | e_1),$$

i.e., the probability of the event happening given e_1 is bigger than the probability of it not happening. We perform these experiments repeatedly over all the relevant chains, and evaluate for each:

$$\text{precision} = \frac{|\{\text{events reported}\} \cap \{\text{predicted events}\}|}{|\{\text{predicted events}\}|}$$

and

$$\text{recall (sensitivity)} = \frac{|\{\text{events reported}\} \cap \{\text{predicted events}\}|}{|\{\text{events reported}\}|}.$$

3.1.3 Comparative Analysis

We are not aware of any methods in the literature that are aimed at tackling the prediction of probabilities of future news events. Thus, we compare the generated predictions with two baselines:

1. using prior probabilities of the occurrence of an event e given the appearance of its corresponding text in the training set, $P(e)$;

2. using an estimate of how well people do on predicting these types of events.

For the latter, we implement a method [11] that provides approximations of whether people, given two events represented by text of news stories, would agree that the first event implies the truth of the later event. This baseline evaluates the probabilities of co-occurrence in text rather than in time.

		Predicted		
		Few	Tens	Hundreds
Real	Few	40%	6%	1%
	Tens	4%	32%	1%
	Hundreds	1%	6%	9%

Table 2: Confusion matrix showing predicted versus actual number of deaths.

3.2 Prediction Results

We performed experiments evaluating the precision and recall for the general predictions and for each of the different domains. We compare our model (Full model) with the frequency-based model (Frequency), and the co-occurrence-based method (Co-occurrence). The results are presented in Table 1. We observe that in all cases the Full model outperforms the baselines.

We performed additional experiments to evaluate numerical predictions, such as forecasts of numbers of deaths. For this purpose, we searched for specific patterns in the news stories of the form "[number] died" or "[number] killed". The number matching [number] is used as the bin classification. We focus only on chains containing those patterns and evaluate our algorithms on those. In Table 2, we show a confusion matrix for the numbers of deaths. The content of each cell i, j (i is row and j is column) represents the percentage of the data that in reality belongs in bin i and is classified as belonging in bin j. For example, 4% of the events resulting in tens of deaths are predicted erroneously as events associated with only a few (less than ten) deaths. We see high performance in those types of classifications and most mistakes are observed in adjacent bins.

3.3 Algorithm Analysis

We now describe additional experiments performed to measure the performance of the procedures and the contribution of specific components of the analysis.

3.3.1 Gain from Factual Features and Generalization

We first consider the influence of adding different sources of world knowledge on the accuracy of predictions. The results are displayed in Table 3. We consider predictors based solely on lexical features (News alone), on both lexical and factual features (News + factual features), on lexical features and abstractions (News + generalization), and on using all categories of features along with the abstraction procedure (Full model). We find that adding knowledge, either when abstracting or when adding factual features, improves the performance of predictions. We see the biggest performance gain when employing both refinements.

3.3.2 Predicting Times of Forthcoming Events

Table 4 displays the average and median times between the inference-based alerts and the occurrence of reports that embody the target event for the three types of predictions we study. We consider only examples where deaths appear in the news story title and match a handcrafted template (patterns in the text of the form "[number] died" or "[number] killed") to identify certain deaths only on test chains. This procedure results in 951 death predictions. In many cases, we find that the alerts would come more than a week in advance of the target event. We illustrate this phenomenon in Figure 4, where we show predictions of the number of deaths

General Predictions		Death		Disease Outbreak		Riots	
Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.
9	21	8	41	12	273	18	30

Table 4: Median and average time between alerts based on inferred probabilities of outcome and target events in the world (days).

that come at a future time within a storyline predicted at different times between the alert and the occurrence of the deaths. A more detailed view of the timing of alerts at two and fifteen days before the event are displayed in Figure 5.

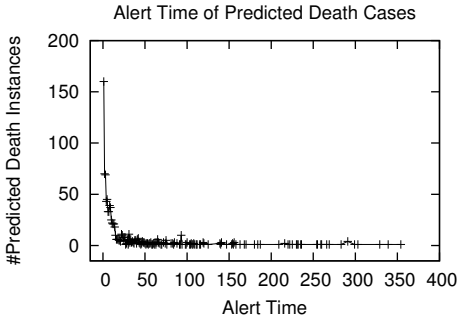

Figure 4: Number of times deaths of any number were predicted as a function of alert time (days before event).

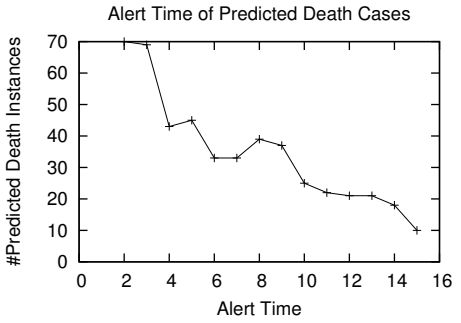

Figure 5: Number of times deaths of any number were predicted as a function of alert time (days before event).

3.4 Event Chain Extraction Evaluation

To evaluate the quality of the extracted event chains, we performed experiments on the TDT4 corpus[4], filtering only NYT articles. This corpus contains about 280,000 documents from the dates 04/01/2003–09/30/2003. Human annotation for labeling storylines was performed by the organizers of the TDT challenge. For each chain, we calculate the average precision—the percentage of articles we extracted as being in a chain that were indeed part of the storyline. We also compute the average recall, the number of articles actually in the chain that the system retrieved. We compared the event chain extractor using the entity entropy measure with

[4] http://www.nist.gov/TDT

	General Predictions		Death		Disease Outbreak		Riots	
	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall
Full model	24%	100%	83%	81%	61%	33%	91%	51%
Frequency	<1%	100%	59%	<1%	13%	3%	50%	1%
Co-occurrence	7%	100	46%	61%	40%	<1%	61%	14%

Table 1: Precision and recall of predictions for several domains.

	General Predictions		Death		Disease		Riots	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
News alone	19%	100%	80%	59%	44%	34%	88%	38%
News + factual features	19%	100%	81%	62%	52%	31%	87%	42%
News + generalization	21%	100%	81%	67%	53%	28%	88%	42%
Full model	24%	100%	83%	81%	61%	33%	91%	51%

Table 3: Precision and recall for different algorithm configurations.

the extractor working without the entropy measure. The results are summarized in Table 5. The results show that, while the recall of text clustering is very high (by 10%), the precision is significantly lower than the methods we have presented (by 30%). We therefore prefer the second method, as it provides more flexibility in training predictive models in a more precise way, with influence on the number of examples used for training the learner.

	Precision	Recall
Text Clustering	34%	**80%**
Text Clustering + Entity Entropy	**70%**	63%

Table 5: Precision and recall for chain extraction procedure.

3.5 Sample Likelihoods and Storylines

The learning and inference methodology we have described can be used to output the probabilities of key transitions of interest from sequences of observations. The system continues to refine its learning with updates of news and related data on the Web. As we mentioned, the system can provide real-time alerting from news stories on sets of specific outcomes that it is monitoring. Examples of statistics of representative learned transition probabilities are displayed in Figure 6. These transition probabilities and mean times to transition highlight the ability of the methods to provide inferences about a variety of levels of abstraction.

We now present details on several additional storylines, along with inferences and timing. Consider the example displayed graphically in Figure 1. On January 26th, 2007, the New York Times published an article about storms and floods in Africa. News of a cholera epidemic were reported four days later. In response to this stream of news, the methodology we describe yields two alerts, one when observing the drought reports in Angola at the beginning of 2006, and another one after news of the reported storms. The system learned from numerous similar incidents in its training set that the likelihood of a cholera outbreak is higher after droughts, specifically as reports on observations of drought are linked to increases in the probability of later reports of water-related disasters, which, in turn, are linked to increases in the likelihood of reports of waterborne diseases. Examples of such transitions and likelihoods include a set of Bangladesh droughts analyzed by the system. 19 significant cases of drought were reported in Bangladesh between 1960–1991 [19]. We observed that in the story lines describing those droughts, a cholera outbreak was reported later in the storyline in 84% of cases. After the 1973 drought, which was responsible for the famine in 1974, the NYT reported on October 13, 1975: "cholera epidemic hits Bangladesh; may prove worse than one that set record in '74...". On March 13 1983, a year after the 1982 drought that "caused a loss of rice production of about 53000 tons while in the same year, flood damaged 36000 tons ...", the NYT published an article entitled, "Bangladesh cholera deaths." Several months later, an article appeared entitled "cholera reportedly kills 500 in 3 outbreaks in Bangladesh". Based on these past story lines the system infers the outbreak of cholera at the end of January in 2007.

The prediction method learns that not all droughts are associated with jumps in the likelihood of such outbreaks of disease. Specific sets of preconditions influence the likelihood of seeing a transition from a report of drought to a report of cholera outbreak. The method was able to recognize that the drought experienced in New York City on March 1989, published in the NYT under the title: "Emergency is declared over drought" would not be associated with a disease outbreak. The only consequence was that New York City declared water curbs, which ended on May 16th of that year. The system estimates that, for droughts to cause cholera with high probability, the drought needs to happen in dense populations (such as the refugee camps in Angola and Bangladesh) located in underdeveloped countries that are proximal to bodies of water.

As an additional example of predictions, we focus on the case of the 1991 cholera epidemic in Bangladesh. This cholera outbreak is estimated to have included 210,000 cases of cholera with more than 8,000 deaths [16]. In our experiments, we found that the running prediction system would have produced an alert four days before the beginning of the cholera outbreak, following observation of the major floods. In Figure 6, we display graphically the storyline detected. The system identifies that reports of major floods with high probability will be followed by reports of significant disease outbreak in Bangladesh. The inferences of the system are supported by a large study of cholera epidemics in Bangladesh [16], based on analyses of government figures, as well as data collected independently in 400 rural areas in Bangladesh between the years 1985N1991. The analysis shows that the number of cholera cases and deaths in 1987 and 1988 is significantly higher than in other years (300,000-1,000,000 cases vs. 50,000 cases in other years). In 1987 and 1988, severe floods occurred in Bangladesh. The study concludes that

Cause	Effect	Probability
Drought	Flood	18%
Flood	cholera	1%
Flood in Rwanda	cholera in Rwanda	67%
Flood in Lima	cholera in Lima	33%
Flood in Country with water coverage $> 5\%$	cholera in Country	14%
Flood in Country with water coverage $> 5\%$, population density > 100	cholera in Country	16%

Table 6: Probability transitions for several examples.

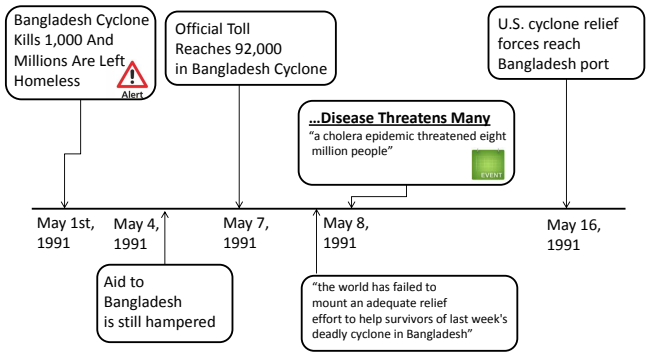

Figure 6: Example of cholera alert following storms in Bangladesh. Triangular alert icons represent inferences of significant upswings in likelihood of forthcoming cholera outbreak.

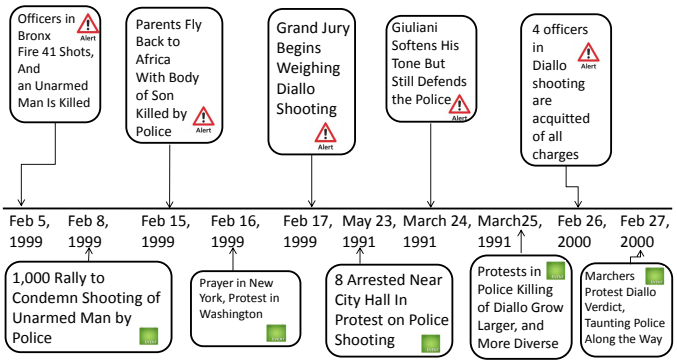

Figure 7: Example of alerts on the likelihood of forthcoming riots after shooting of unarmed minority. Triangular alert icons represent inferences of significant upswings in likelihood of a forthcoming riot.

access to medical care was one of the main reasons for high death rates in many non-rural areas. In areas where appropriate interventions were made at early stages, the death rates were significantly smaller.

We also study examples of prior deaths and riots. In Figure 7, we present a partial storyline and the alerts inferred for the Dialo case of 1999. One of the storylines is presented in detail in Table 7 (Top). The system identified in an automated manner that for locations with large immigrant populations (e.g., Ohio and New York), the shooting of an unarmed person by the police can cause protests. Additional events in the news, such as reports of the funeral of the people who have been killed in similar way, of the beginning of the trial of the policemen who performed the shooting, of support for the policemen, and of the end of the trial are all associated with increases in the likelihood of later reports of protests. A sample storyline at the basis of the inferred probabilities is presented in Table 7 (Bottom).

4. RELATED WORK

Prior related research includes efforts in political science on forecasting forthcoming international political crises from coded event data, including event data extracted from news stories [23]. Research in this realm includes the application of HMMs to identify similarities among attributes that appear to be linked to the development of international crises [22]. Relevant research has also explored predicting riots [13] and the sales of movie tickets [5, 12, 18] from signals derived from social media such as Twitter. Other investigations have leveraged information in text of news and book corpora to qualitatively estimate how multiple aspects of human culture evolve [25, 25, 17]. Other relevant work in search and retrieval has focused correlating logs of queries input to search engines with future events in both traditional media [20] and blogs [1]. Ginsberg et al. [10] used queries for predicting H1N1 influenza outbreaks. Other research has sought to predict how Web content changes. Kleinberg [14, 15] developed general techniques for summarizing the temporal dynamics of textual content and for identifying bursts of terms within content. Similarly, other works [4] build time-series models over publication dates of documents relevant to a query in order to predict future bursts. In other related work, Radinsky et al. [21] extracted generalized templates in the form of "x causes y" from past news. The templates were applied on a present news title, generating a plausible future news title.

In this work, we take a probabilistic approach and perform more general-purpose predictions without relying on templates. We also combine heterogeneous online sources, leveraging world knowledge mined from more than 90 sources on the Web, to enrich and generalize historical events for the purpose of predicting future news.

5. CONCLUSIONS

We presented methods for mining chains of events from 22 years of news archives to provide a methodology that provides real-time predictions about the likelihoods of future world events of interest. The system harnesses multiple Web resources to generalize the events that it learns about and predicts. We discussed how we can learn patterns from large amounts of data, monitor large quantities of information sources, and continue to learn new probabilistic asso-

Date	Title
Jan 16, 1992	Jury in Shooting by Officer Hears Conflicting Accounts
Feb 11, 1992	Closing Arguments Conflict on Killing by Teaneck Officer
Feb 12, 1992	**[Past Event] Officer Acquitted in Teaneck Killing**
Feb 13, 1992	Acquitted Officer Expresses Only Relief, Not Joy
Feb 16, 1992	**[Past Riot] 250 March in Rain to Protest Teaneck Verdict**

Feb 24, 2000	Diallo Jurors Begin Deliberating In Murder Trial of Four Officers
Feb 26, 2000	**[Riot Alert] 4 officers in Diallo shooting are acquitted of all charges**
Feb 26, 2000	Rage Boils Over, and Some Shout 'Murderers' at Police
Feb 26, 2000	Civil Rights Prosecution Is Considered
Feb 27, 2000	**[Riot Event] Marchers Protest Diallo Verdict...**
Feb 27, 2000	2 jurors defend Diallo acquittal

Table 7: Top table: Partial sample of a historical storyline used to infer probabilities. Bottom table: Partial storyline with an alert.

ciations. To demonstrate the approach, we presented the results of several evaluations and representative examples of sequences of events and proactive alerts. We considered as sample inferences predictions about disease outbreaks, riots, and deaths. We believe that the methods highlight directions in building real-time alerting services that predict significant increases in global events of interest. Beyond knowledge that is easily discovered in studies or available from experts, new relationships and context-sensitive probabilities of outcomes can be discovered with such automated analyses. Systems employing the methods would have fast and comprehensive access to news stories, including stories that might seem insignificant but that can provide valuable evidence about the evolution of larger, more important stories. We hope that this work will stimulate additional research on leveraging past experiences and human knowledge to provide valuable predictions about future events and interventions of importance.

6. REFERENCES

[1] E. Adar, D. S. Weld, B. N. Bershad, and S. D. Gribble. Why we search: visualizing and predicting user behavior. In *WWW*, 2007.

[2] A. Ahmed, Q. Ho, J. Eisenstein, E. Xing, A. J. Smola, and C. H. Teo. Unified analysis of streaming news. In *Proc. of WWW*, 2011.

[3] J. Allan, editor. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[4] G. Amodeo, R. Blanco, and U. Brefeld. Hybrid models for future event prediction. In *CIKM*, 2011.

[5] S. Asur and B. A. Huberman. Predicting the future with social media, 2010.

[6] C. Bizer, T. Heath, and T. Berners-Lee. Linked data – the story so far. *IJSWIS*, 2009.

[7] J. Carbonell, Y. Yang, J. Lafferty, R. D. Brown, T. Pierce, and X. Liu. Cmu report on tdt-2: segmentation, detection and tracking, 2000.

[8] C. Cieri, D. Graff, M. Libermann, N. Martey, and S. Strassel. Large, multilingual, broadcast news corpora for cooperative research in topic detection and tracking: The tdt-2 and tdt-3 corpus efforts. In *LREC*, 2000.

[9] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, 2005.

[10] J. Ginsberg, M. Mohebbi, R. Patel, Brammer, M. L., Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009.

[11] O. Glickman, I. Dagan, and M. Koppel. A probabilistic classification approach for lexical textual entailment. In *Proc. of AAAI*, 2005.

[12] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In *In Proc. of NAACL-HLT*, 2010.

[13] Kalev. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 15(9), 2011.

[14] J. Kleinberg. Bursty and hierarchical structure in streams. In *KDD*, 2002.

[15] J. Kleinberg. Temporal dynamics of on-line information systems. *Data Stream Management: Processing High-Speed Data Streams. Springer*, 2006.

[16] J. Michel, Y. Shen, A. Aiden, A. Veres, M. Gray, Google Books Team, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. Nowak, and E. Aiden. Cholera epidemics in bangladesh: 1985-1991. *Journal of Diarrhoeal Diseases Research (JDDR)*, 10(2):79–86, 1992.

[17] J. Michel, Y. Shen, A. Aiden, A. Veres, M. Gray, Google Books Team, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. Nowak, and E. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331:176–182, 2011.

[18] G. Mishne. Predicting movie sales from blogger sentiment. In *In AAAI Spring Symposium*, 2006.

[19] R. Nagarajan. *Drought Assessment*. Springer, 2009.

[20] K. Radinsky, S. Davidovich, and S. Markovitch. Predicting the news of tomorrow using patterns in web search queries. In *WI*, 2008.

[21] K. Radinsky, S. Davidovich, and S. Markovitch. Learning causality for news events prediction. In *Proceedings of WWW*, 2012.

[22] D. Richards, editor. *Political Complexity: Nonlinear Models of Politics*. Ann Arbor: University of Michigan Press, Norwell, MA, USA, 2000.

[23] R. J. Stoll and D. Subramanian. Hubs, authorities, and networks: Predicting conflict using events data, 2006.

[24] F. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proc. of WWW*, 2007.

[25] C. Yeung and A. Jatowt. Studying how the past is remembered: Towards computational history through large scale text mining. In *Proc. of CIKM*, 2011.