Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback

Hyun Duk Kim
Dept. of Computer Science
University of Illinois at Urbana-Champaign
hkim277@illinois.edu

Malu Castellanos
Information Analytics Lab
HP Laboratories
malu.castellanos@hp.com

Meichun Hsu
Information Analytics Lab
HP Laboratories
meichun.hsu@hp.com

ChengXiang Zhai
Dept. of Computer Science
University of Illinois at Urbana-Champaign
czhai@illinois.edu

Thomas Rietz
Dept. of Finance
The University of Iowa
thomas-rietz@uiowa.edu

Daniel Diermeier
Kellogg School of Management
Northwestern University
d-diermeier@kellogg.northwestern.edu

ABSTRACT

Many applications require analyzing textual topics in conjunction with external time series variables such as stock prices. We develop a novel general text mining framework for discovering such causal topics from text. Our framework naturally combines any given probabilistic topic model with time-series causal analysis to discover topics that are both coherent semantically and correlated with time series data. We iteratively refine topics, increasing the correlation of discovered topics with the time series. Time series data provides feedback at each iteration by imposing prior distributions on parameters. Experimental results show that the proposed framework is effective.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Causal Topic Mining, Iterative Topic Mining, Time Series

1. INTRODUCTION

Probabilistic topic models [4, 8] have proven very useful for mining text data in a range of areas including opinion analysis [11, 17], text information retrieval [19], image retrieval [9], natural language processing [6], and social network analysis [12].

Most existing topic modeling techniques focus on text alone. However, text topics often occur in conjunction with other variables through time. Such data calls for integrated analysis of text and non-text time series data. The causal relationships between the two may be of particular interest. For example, news about companies can affect stock prices, sales numbers, etc. Understanding the impact of, for example, news coverage or customer reviews, is of great practical importance.

While there are many variants of topic models [2, 3, 18], no existing model incorporates jointly text and associated "external" time series variables to identify causal topics. A semantically coherent topic is "causal" if it has strong, possibly lagged, associations with a non-textual time series variable. This allows for two-way relationships: topics may affect the time series and/or vice versa. Our method can be tailored to the specific application and help an analyst quickly identify a small set of possibly causal topics for further analysis.[1]

A basic approach to identifying causal topics is to: (1) find topics with topic modeling techniques then (2) identify causal topics using correlations and causality tests. This approach, however, ignores the possibly relevant information contained in the time series. Candidate topic sets are the same for every time series.

We instead propose a novel general text mining framework: Iterative Topic Modeling with Time Series Feedback (ITMTF). ITMTF naturally combines probabilistic topic modeling with time series causal analysis to uncover topics that are both coherent semantically and correlated with time series data. ITMTF can accommodate any topic modeling technique and any causality measure. This generality enables users to easily adapt different topic models and causality measures as needed. ITMTF iteratively refines a topic model, gradually increasing the correlation of discovered topics with the time series data through a feedback mechanism. In each iteration, the time series data informs a prior distribution of parameters that feeds back into the topic model. Thus, the discovered topics are dynamically adapted to fit the patterns of different time series data.

We evaluate ITMTF on a news data set with multiple stock price time series, including stock prices from the Iowa Electronic Markets and those of two large US companies (American Airlines and Apple). The results show that ITMTF can effectively discover causal topics from text data and the iterative process improves the quality of the discovered causal topics.

2. RELATED WORK

There are two basic topic models: Probabilistic Latent Semantic Analysis (PLSA) [8] and Latent Dirichlet Analysis (LDA) [4]. Both focus on word co-occurrences. Recent advanced techniques analyze the dynamics of topics on a time line [2, 18]. However,

[1]Our definition allows using "correlation" and "cause" interchangeably as convenient.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2505612.

they do not conduct integrated analyses of topics and external variables; the topic analysis is separate from the external time series.

There are some efforts to incorporate external knowledge in modeling. Supervised LDA [3] models topics with better prediction power than simple LDA by incorporating a reference value (e.g. movie review articles with movie ratings) in the modeling process. Labeled LDA [16] associates categorical labels and even text labels for topics. Another way of incorporating external knowledge is to use conjugate prior probabilities in the topic modeling process [13]. Topic Sentiment Mixture (TSM) models positive and negative sentiment topics using seed sentiment words such as "good" or "bad." While these methods show that topic mining can be guided by external variables, none achieves our objective of capturing the correlation structure between text topics and external time series variables. Moreover, while these models are specialized for supervision with specific external data, our general approach can flexibly combine any reasonable topic model with any causal analysis method.

Research on stock prediction using financial news content also relates to our work [14]. Such research typically identifies the most predictive words and labels news according to its effect on stock prices on a specific day using a supervised regression or a classification problem setup. In contrast, we search for general causal topics with unsupervised methods.

Granger testing [7] is popular for testing causality in economics using lead/lag relationships across multiple time series. Recent evidence shows that Granger tests can be used in an opinion mining context: predicting stock price movements with a sentiment curve [5]. However, Granger testing has not been used directly in text mining and topic analysis.

A demo system based on our framework is presented [10] with a very brief description about the system components and sample results. Here, we describe the general problem and framework in detail and evaluate the algorithms rigorously.

3. MINING CAUSAL TOPICS IN TEXT WITH SUPERVISION OF TIME SERIES DATA

Consider time series data x_1, ..., x_n, with time stamps t_1, ..., t_n, and a collection of time stamped documents from the same period, D = {(d_1,t_{d_1}), ..., (d_m, t_{d_m})}. The goal is to discover a set of causal topics T_1, ..., T_k with associated time lags L_1, ..., L_k. A causal topic T_i with time lag L_i is a topic that is semantically coherent and has a strong correlation with the time series data with time lag L_i. Note that L_i can be positive or negative, corresponding to topics that might cause, or be caused by, time series data.

4. ITERATIVE TOPIC MODELING WITH TIME SERIES FEEDBACK

We have two criteria to optimize: topic coherence and topic correlation. We want to retain the generality of the topic modeling framework while extending it to allow the time series variable to influence topic formation so we can optimize both criteria over a more flexible topic space.

4.1 Causal analysis with time series data

Potential "causal" relationships between times series are identified through contemporaneous and/or lagged correlation measures (e.g., Granger tests). The correlation lag structure suggests directional causality. If current observations in time series A correlate with later observations in B, A is said to "cause" B.

A simple and very common measure uses Pearson correlations, contemporaneously or with leads and lags. Correlations range from -1 to +1 with the sign indicating the direction of correlation and can be used as "impact" measures here. A correlation's significance depends on its value and the number of observations.

Granger tests are more structured measures of causality, measuring statistical significance at different time lags using auto regression to identify causal relationships. Let y_t and x_t be two time series. To see if x_t "Granger causes" y_t with maximum p time lag, run the following regression:

$$y_t = a_0 + a_1 y_{t-1} + ... + a_p y_{t-p} + b_1 x_{t-1} + ... + b_p x_{t-p}.$$

Then, use F-tests to evaluate the significance of the lagged x terms. The coefficients of lagged x terms estimate the impact of x on y. We average the x term coefficients, $\frac{\sum_{i=1}^{p} b_i}{|p|}$, as an impact value.

4.2 An Iterative Topic Modeling Framework with Time Series Feedback

Input: time series data X = x_1, ..., x_n with time stamp t_1, ..., t_n, and a a collection of text documents with time stamps from the same period, D = {(d_1,t_{d1}), ..., (d_m, t_{dm})}, topic modeling method M, causality measure C, a parameter tn (how many topics to model)

Output: k potentially causal topics (k≤tn): $(T_1, L_1), ..., (T_k, L_k)$

Topic modeling method M identifies topics. The causality measure C gives significance measures (e.g. p-value) and impact orientation. Figure 1 illustrates our iterative algorithm. It works as follows:

1. Apply M to D to generate tn topics $T_1, .., T_{tn}$
2. Use C to find topics with significance values $sig(C, X, T) > \gamma$ (e.g. 95%). Let CT be the set of candidate causal topics with lags: $CT = \{(T_{c1}, L_1), ..., (T_{ck}, L_k)\}$.
3. For each candidate topic in CT, apply C to find the most significant causal words among top words $w \in T$. Record the impact values of these significant words (e.g., word-level Pearson correlations with the time series variable).
4. Define a prior on the topic model parameters using significant terms and their impact values.
   (a) Separate positive impact terms and negative impact terms. If one orientation is very weak (< δ%, e.g. 10%), ignore the minor group.
   (b) Assign prior probability proportions according to significance levels.
5. Apply M to D using the prior obtained in step 4 (this injects feedback signals and guides the topic model to form topics that are more likely correlated with the time series)
6. Repeat 2-5 until satisfying stopping criteria (e.g. reach topic quality at some point, no more significant topic change, etc.). When the process stops, CT is the output causal topic list.

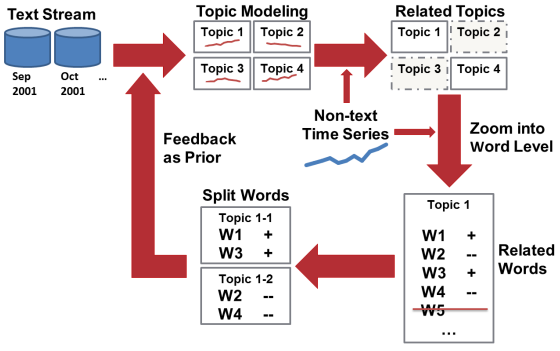


Figure 1: Overview of iterative topic modeling algorithm

ITMTF considers the non-textual time series data in the text mining process to find topics that are more highly correlated to non-textual data than general modeling systems. Moreover, ITMTF

iteratively improves modeling results by considering interactions between the text and time series data at both topic and word levels. After identifying causal topics, ITMTF shifts to word level correlations between the text and external time series data. It also improves the topic modeling process by splitting positively and negatively impacting terms into different topics. Because generating and testing all the word time series is inefficient, ITMTF focuses only on the words with the highest frequencies in the most highly correlated topics discovered in each iteration.

The ideal set of causal topics should have tight relationships with the external time series and high topic quality. Traditional topic modeling algorithms form topics based on word coherences in the text data, while causality tests can filter out non-causal topics. Focusing exclusively on one criterion sacrifices the other. Our iterative process is a greedy approximate solution to the two-item maximization problem. It takes turns optimizing each criterion. The prior formation based on causality attempts to optimize causality while a topic modeling optimizes coherence of topics.

4.2.1 Topic-level Causality Analysis

From topic modeling results, we generate a topic curve over time by computing each topic's coverage on each time unit (e.g., one day). Consider a weighted coverage count. Specifically, compute the coverage of a topic in each document, $p(topic_j|Document_d)$. This is simply the parameter $\theta_j^{(d)}$ estimated in the modeling process. Estimate the coverage of topic j at t_i, tc_i^j as the sum of $\theta_j^{(d)}$ over all the documents with t_i time stamp. Call the list of tc^j for all the time stamps the *topic stream* TS_j:

$$tc_i^j = \sum_{\forall d \ with \ t_i} \theta_j^{(d)} \ , \quad TS_j = tc_1^j, tc_2^j, ..., tc_n^j.$$

This creates a topic stream time series that, combined with the non-textual time series data, lends itself to standard time series causality measures C and testing.

Selecting lag values is important. One possibility is to use the most significant lag within a given maximum. For example, if we want to find causal topics within 5 days, we can choose the lag within 5 days with the highest significance. If we want to focus on yesterday's effect, we can choose a fixed lag of 1. The specific choice depends on the specific aims of an application.

4.2.2 Word-level Causality and Prior Generation

Based on topic causality scores, we choose a subset of promising topics with the highest causality scores and further analyze the words within each topic to provide feedback for the next iteration by generating topic priors. Specifically, for each significant topic, we check whether the top words in the topic are also significantly correlated with the external time series. For each word, we generate a word count stream WS_w by counting frequencies in the input document collection for each day:

$$wc_i = \sum_{\forall d \ with \ t_i} c(w,d) \ , \quad WS_w = wc_1, wc_2, ..., wc_n,$$

where $c(w,d)$ is the count of word w in document d. Then we measure correlations and significance between word streams and the external non-textual time series. This identifies words that are significantly correlated and their impact values.

Intuitively, we want to emphasize significant topics and words in our next topic modeling iteration to focus in more promising topic space. To do this, we generate topic priors for significant words in significant topics. A topic prior is a Dirichlet distribution that favors topics assigning high probabilities to the identified significant words in significant topics. We assign prior word probabilities in proportion to the significance value of the words. This prior helps "steer" the topic modeling process to form/discover topics similar to the prior topics [13]. In the next topic modeling iteration, the discovered topics are likely to be close to the prior, which is based on the feedback from the time series variable through causality scores.

In addition to keeping significant topics and words, we also want to improve topic quality. A "good" correlated topic has a consistent impact relative to the external time series. We want relatively consistent topics, those containing words that have mostly "positive" or mostly "negative" impacts on the external time series. Therefore, if one topic has both positive and negative impact words, we separate the positive and negative impact words into two topics in the prior for the next topic modeling iteration. If one of the word impact groups is much smaller than the other (e.g. the number of positive impact words < 0.1 * the number of negative impact words), we keep only one topic and set the probability of words in the smaller group zero. This "tells" the topic model not to include such words in the refinement of the topic.

Table 1: Example of topic and word correlation analysis result (left) and prior generated (right) (Sig: significance, Prob: probability)

WORD	IMPACT	SIG (%)
social	+	99
security	+	96
gun	-	99
control	-	97
september	-	99
terrorism	-	97
...		
attack	-	96
good	+	96

⇒

WORD	PROB
social	0.8
security	0.2
gun	0.75
control	0.25
september	0.1
terrorism	0.075
...	
attack	0.05
good	0.0

Suppose, among N total topics, we identify 2 significantly correlated topics with the external time series (left of Table 1). We check correlations of the top words in these two topics. Suppose 4 and 10 words were significant for the two topics respectively. Because there are both positive and negative word groups with similar sizes, we would split the first topic into two topics and assign word probabilities based on significance values. For the second topic, only one word has a different impact orientation from the others making the negative group much smaller than the positive group. Therefore, instead of making a separate negative word group topic, we exclude it from the positive word group topic by assigning it zero weight. Right side of Table 1 shows the example prior generated.

Another challenge is selecting a cutoff for "top" words in the correlation list. The simplest solution is to set a fixed cutoff, say k, and use the top k words. However, rank alone does not determine the importance of words. Importance is determined by word probabilities as well. For example, suppose the top three words in Topic 1, A, B, and C have probabilities 0.3, 0.25 and 0.2, respectively, and the top three words in Topic 2, D, E, and F have probabilities 0.002, 0.001 and 0.0001. In this case, while A, B and C are very important in Topic 1, D, E, and F combined only represent a small part of Topic 2. Hence, Topic 2 may require more words to be considered. We address this by using a cumulative probability mass cutoff, $probM$. We use all words whose accumulated probabilities are within a cutoff.

Formally, for each topic T_j = $(w_1, \phi_{w_1}^{(j)})$, ..., $(w_{|V|}, \phi_{w_{|V|}}^{(j)})$ ($|V|$ is the number of words in the input data set vocabulary), when items are sorted by $\phi_{w_i}^{(j)}$, we can add the top ranked word to the top word list TW without violating the constraint $\sum_{w \in TW} \phi_w^{(j)} \leq probM$ where $TW = (w_1, ..., w_m)$. That is, $\sum_{w \in TW} \phi_w^{(j)} + \phi_{w_{m+1}}^{(j)} > probM$. With top word $TW = (w_1, ..., w_m)$ and significance value of each word $sig(C, X, w)$, the topic prior ϕ'_w can be computed by the following formula:

$$\phi'_w = \frac{sig(C,X,w) - \gamma}{\sum_{j=1}^{m}(sig(C,X,w_j) - \gamma)},$$

where γ is a significance cutoff (e.g. 95%).

4.2.3 Iterative Modeling with Feedback

Using the new prior, we remodel topics. New topics will be guided by priors, which depend on correlations with the external data. High probability words in the prior have a greater impact in the modeling results and words with zero probability are not included in the topic. By repeating the process of topic modeling, correlation analysis, and prior generation, the resulting topics are likely more highly correlated with the external time-series.

The strength of the prior in each iteration is set by a parameter μ in the modeling process [13]. With $\mu = 0$, modeling would not consider the prior information at all (making it the same as independent modeling). With a very high μ, words in the prior are very likely to appear in the topic modeling results. We study this parameter's influence in our experiments.

While we observe correlations between non-textual series and both word streams and topic streams, we do not compute correlations for all word streams. Word level analysis would give us finer grain signals. However, generating all the word frequency time series and testing correlations would be very inefficient. By narrowing down to significant topics first, we can prune the number of words to test. This increases efficiency and effectiveness.

5. EVALUATION

5.1 Experiment Design

We evaluate the proposed algorithms on the New York Times data set [2] with multiple stock time series data.

In one experiment, we examine the 2000 U.S. Presidential election campaign. The input text data comes from New York Times articles from May through October of 2000. We filter them for key words "Bush" and "Gore," and use paragraphs mentioning one or both words. The idea is to find specific topics which caused support for Bush or Gore to change and not simply election related topics. As a non-textual time series input, we use prices from the Iowa Electronic Markets (IEM)[3] 2000 Presidential Winner-Takes-All Market [1]. In this on-line futures market, prices forecast the probabilities of candidates winning[4] the election. We follow standard practice in the field and use the "normalized" price of one candidate as a forecast probability of the election outcome: (Gore price)/(Gore price + Bush price).

In another experiment, we use stock prices of American Airlines and Apple and the same New York Times text data set with a longer time period, but without keyword filtering, to examine the influence of having different time series variables for supervision.

While the framework is general, comparing different topic models is not the focus of our paper. So, we only used one topic model: PLSA implemented based on the Lemur information retrieval toolkit.[5] For correlation measures, we use both contemporaneous Pearson correlation coefficients and Granger tests. For Granger tests, we use the R statistical package[6] implementation. Granger tests require stationary time series as inputs. To make the input series stationary, we smooth with a moving average filter with window size 3 (average with adjacent values) and use first differences $(x_t) - (x_{t-1})$ of each series. We test causality with up to 5 day lags and pick the lag which shows highest significance.

5.1.1 Measures of Quality

We report two measure the quality for mined topics: causality confidence and topic purity. For causality confidence, we use the significance value (i.e. p-value) of the Granger test between the text stream and the external variable. For topic purity, we use the impact orientation distribution of significant words. If all the significant words in one topic have the same orientation, it has 100% purity. If significant words are evenly split by positive and negative impact, it has 0% purity. We calculate the entropy of significant word distributions and normalize it to the $[0, 100]$ range. Thus, the Purity of topic T is defined as:

$$pProb = \frac{the\ number\ of\ positive\ impact\ words}{the\ number\ of\ significant\ words}$$

$$nProb = \frac{the\ number\ of\ negative\ impact\ words}{the\ number\ of\ significant\ words}$$

$$Entropy(T) = pProb * log(pProb) + nProb * log(nProb)$$

$$Purity(T) = 100 + 100 * Entropy(T).$$

We report average causality confidence and purity for topics with more than 95% significance. Thus, when there are more significant topics, this measure may be penalized. However, because measuring general utility of significant topics is meaningful from a user perspective, we do not adjust this measure.

5.1.2 Baseline

The first iteration of our method is based on topic modeling without guidance from the time series and, thus, is a natural baseline. Comparing iterations and final results to this shows the benefit of iterative topic mining.

5.1.3 Parameter Tests

We test two parameters for effects on performance. The first is the number of topics modeled (tn). A large number of topics may help identify more specific and more coherent (higher purity) topics. However, topics that are too specific result in data sparseness that reduces the power of significance testing. A small number of topics gives the opposite effects: topics are likely to have higher statistical significance, but would have lower purity. Also, because many meaningful topics may be merged into a single topic, topics may be too coarse to interpret easily. The second parameter is the strength of the prior (μ). A stronger topic prior would guarantee prior information is reflected in the next topic modeling iteration. However, if the initial topic modeling (which uses random initiation without a prior) ends up at a poor local optimum, a strong prior may keep the process there, resulting in poor topics. Strong priors may also exacerbate spurious correlation resulting from noise in the first round. In contrast, weaker priors allow a less restricted iteration of topic modeling, reducing these negative effects. However, positive signals would also have weak impact. Because prior research gives no guidance for selecting these parameters, we examine how they affect the performance of our algorithm.

5.2 Experiment Results

5.2.1 Sample Results

2000 Presidential Election: Table 2 shows sample results from the 2000 U.S. Presidential election. It shows the top three words of significant causal topics mined (Pearson correlation, tn=30, μ=50, 5th iteration). The result reveals several important issues from the campaigns, e.g. tax cuts, abortion, gun control and energy. Such topics are also cited in political science literature [15] and Wikipedia[7] as important election issues. This shows that our iterative topic mining process can converge to issues expected to affect the election.

Stock Time Series, AAMRQ vs. AAPL: To study how different time series affect the topics discovered, we compare the topics discovered from the *same* text data set using two different time series.

[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19

[3] http://tippie.uiowa.edu/iem/

[4] "Winning" as defined by the IEM is taking the majority of the two-party popular vote

[5] http://www.lemurproject.org/

[6] http://www.r-project.org/

[7] http://en.wikipedia.org/wiki/United_States_presidential_election,_2000#General_election_campaign

Table 2: Significant topic list of 2000 Presidential Election (Each line is a topic with top three probability words.)

TOP THREE WORDS IN SIGNIFICANT TOPICS
tax cut 1
screen pataki giuliani
enthusiasm door symbolic
oil energy prices
pres al vice
love tucker presented
partial **abortion** privatization
court supreme **abortion**
gun control nra
news w top

We use New York Times articles from July 2000 through December 2001 as the text input. We use American Airlines (AAMRQ) and Apple (APPL) stock prices as external time series. American Airlines' stock (travel industry) dropped significantly in September 2001 because of the 9/11 terrorist attack, while Apple stock (IT industry) was less affected. We start with the same modeled topic list at the first iteration. Thus, any differences in modeled topics are from feedback of the external time series.

Table 3: Significant topic list of two different external time series: AAMRQ and AAPL (Each line is a topic. Top three probability words are displayed.)

AAMRQ	AAPL
russia russian putin	paid notice st
europe european germany	russia russian europe
bush gore presidential	olympic games olympics
police court judge	she her ms
airlines airport air	oil ford prices
united trade terrorism	black fashion blacks
food foods cheese	**computer technology software**
nets scott basketball	**internet com web**
tennis williams open	football giants jets
awards gay boy	japan japanese plane
moss minnesota chechnya	...

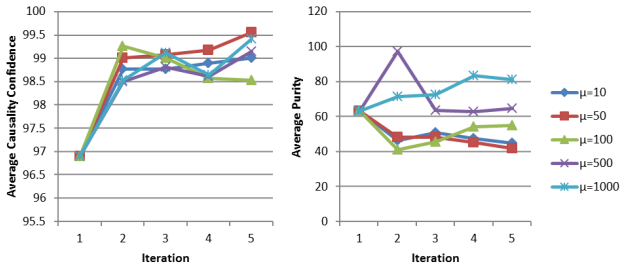
Table 3 shows the top three words of significant topics mined using the two different external time series after three rounds (tn=30 and μ=1000). Topics associated with American airlines include clearly relevant words such as "airlines airport air" and "united trade terrorism." One topic is clearly about the terrorist attack. Topics associated with Apple differ dramatically. Relevant topics, "computer technology software" and "internet com web", reference Apple's IT industry.

This example also shows a limitation of our algorithm. In addition to clearly relevant topics, there appear other general topics (e.g., sports). This task presents more challenges than the 2000 U.S. Presidential election example because of the diversity in text data and long time period. While we use text articles which are related to candidates for the Presidential election case, we use all articles in the time period for this example. Greater topic diversity can lead to more spurious correlation independent of real causality. Moreover, our analysis is over 18 months and the algorithm measures correlation over the entire time period. Therefore, if an event is only locally correlated, it may not be selected in the final output topic list. How to measure and deliver local correlation feedback remains for future work.

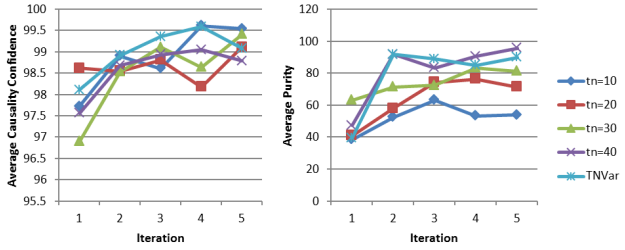
Despite these difficulties, our algorithm shows how different time series inputs select different topics relevant to themselves using the same text data and initially modeled topics. Thus, our algorithm can effectively guide topic modeling. Pre-filtering relevant articles and shorter time periods may yield better results. Moreover, while some topics seem unrelated at first blush, they may reveal unexpected, but meaningful, relationships.

5.2.2 *Quantative evaluation results*

We ask two questions: 1) Is our joint mining algorithm more effective than simply combining an existing topic model with a time series analysis method in a straightforward way? 2) Is the feedback mechanism in the proposed framework beneficial? To answer both, we study how results change between the baseline method (with no feedback) and ITMTF through multiple iterations.



(a) With different μ



(b) With different tn

Figure 2: Causality confidence (left) and purity (right) with different parameter settings over iteration (Presidential election data, Granger tests)

Figure 2(a) shows performance evaluation results with different μs using Granger tests. Average causality confidence increases over iterations regardless of the strength of feedback, μ. The performance improvement is particularly notable between the first iteration (baseline with no feedback) and the second iteration (with one feedback round). Clearly, feedback is beneficial. After the second iteration, performance shows slower improvement. Later rounds appear to fine tune topics resulting from the first round.

The average purity graph shows mixed results. For small μ values (μ=10, 50, 100), iterations do not always improve purity. With higher μ values (μ=500, 1000), average purity improved from the first iteration to the second. Furthermore, for the highest μ value (μ=1000), it showed a steady increase. Weak μs would allow topic modeling more room to explore variations regardless of prior. Therefore, the purity improvement may not be guaranteed in each iteration. Thus, high μ values lead to higher increases in purity. Further, reported purity is the averaged purity value of all the significant topics found. The number of significant topics increases dramatically between the first and second iteration. Thus, the drop in average purity may not be a negative sign. Still, higher μs ensure purity improvement.

Figure 2(b) shows performance evaluation results with different topic numbers (tn) using Granger tests. Initially, small topic numbers (tn=10, 20) had higher confidence levels than large topic numbers (tn=30, 40). Intuitively, the statistical signal is stronger with small topic numbers, while sparseness associated with large topic numbers reduces statistical strength. However, with more iterations, the relative order changes. Thus, the feedback process helps overcome sparseness problems. Significantly correlated topics and words are kept by iterations of priors and topic modeling iterations add coherence to them. Thus, in the end, the number of topics has little effect on average confidence.

In general, modeled topics with larger tn show higher purity than with small tn. As expected, each topic is specific, and the chance of

combining multiple real topics in one is smaller. Therefore, topics likely have better purity.[8]

To show that the improvement is not an artifact of noise in topic modeling, we test significance of the performance improvement between the first and second iteration. We execute 20 separate trials with random initiation and applied ITMTF (tn=30, μ=1000). Paired t-test between the first and second iterations showed >99% significance for each measure (t-value: 3.87 for average confidence, 14.34 for average purity). Thus, feedback significantly improves average correlation and topic quality over simple topic modeling. Beyond the first feedback round, causality improvements are relatively small. Thus, in practice, one round may be sufficient.

Overall high μ values clearly improve topic quality. Results on tn values are less clear. Next, we describe an approach and experiment results for finding the optimal tn.

5.2.3 Optimal Number of Topics

In practice, selecting the appropriate number of topics (tn) presents challenges. We propose a "variable topic number" approach. Our algorithm naturally splits topics by word impacts in each iteration. Therefore, we can start with an small tn, let the algorithm split topics in each prior and use the increased number of topics in the next iteration. We can also add some buffer topics in some or all rounds to give topic modeling room to explore more topics. For example, 7 out of 10 initial topics may be deemed causal in a round and 5 out of 7 may be split. The next prior will include 12 topics. Adding 5 more buffer topics would result in 17 topics for the next iteration.

Topics tend to have high causality with small tn. Likely, many will be retained as causal from the beginning. With iteration, topics are split. While the number of topics rises, the proportion of causal topic will likely fall. We suggest stopping when the number of causal topics starts to fall relative to the previous iteration, which means topic splitting hurts more than iterative modeling improves causal topic identification. When we actually apply this starting with $tn = 10$, the number of significant causal topics starts to decrease after 30 total topics, so we set tn=30.

We test this variable topic number algorithm against fixed topic numbers in our topic number analysis. *TNVar* in Figure 2(b) shows average confidence and purity compared to the fixed tn methods. In both average confidence and purity, the variable topic number approach performs well, proving its efficacy.

6. CONCLUSIONS

Here, we present a novel general text mining framework: Iterative Topic Modeling with Time Series Feedback (ITMTF) for causal topic mining. ITMTF uses text data and a non-text external time series as inputs and iteratively models topics related to changes in the external time series. Experimental results show that ITMTF finds topics that are both more pure and more highly correlated with the external time series than typical topic modeling, especially with a strong feedback loop.

The general problem of mining causal topics opens new directions for future research, both theoretical and applied. ITMTF can be generalized using any topic modeling techniques and causality/correlation measures desired. Our examples illustrate one of many ways to implement the framework, which can certainly be implemented in other ways. In future work, we hope to extend ITMTF's ability to identify locally correlated events. In addition, the proposed alternating optimization strategy between coherence and causality is only a heuristic without theoretical guarantees of convergence. While, empirically, this strategy works well on tested data sets, an obvious and interesting extension would be to integrate topic models with time series data more tightly using a single unified objective function for optimization.

[8]We performed the same series of tests using Pearson correlation coefficients. In general, the results are similar, but not reported here because of space limitations.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under Grant Number CNS-1027965 and by an HP Innovation Research Award.

7. REFERENCES

[1] J. Berg, R. Forsythe, F. Nelson, and T. Rietz. *Results from a Dozen Years of Election Futures Markets Research*, volume 1 of *Handbook of Experimental Economics Results*, chapter 80, pages 742–751. Elsevier, 2008.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, New York, NY, USA, 2006. ACM.

[3] D. M. Blei and J. D. Mcauliffe. Supervised topic models. 2007.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.

[6] J. Boyd-Graber, D. M. Blei, and X. Zhu. A topic model for word sense disambiguation. In *EMNLP '07: Proceedings of the 2007 conference on Empirical Methods in Natural Language Processing*, 2007.

[7] C. W. J. Granger. Essays in econometrics. chapter Investigating causal relations by econometric models and cross-spectral methods, pages 31–47. Harvard University Press, Cambridge, MA, USA, 2001.

[8] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 1999 international ACM SIGIR conference on research and development in Information Retrieval*, pages 50–57, New York, NY, USA, 1999. ACM.

[9] E. Hörster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *CIVR '07: Proceedings of the 2007 ACM international conference on Image and video retrieval*, pages 17–24, New York, NY, USA, 2007. ACM.

[10] H. D. Kim, C. Zhai, T. A. Rietz, D. Diermeier, M. Hsu, M. Castellanos, and C. Ceja. Incatomi: Integrative causal topic miner between textual and non-textual time series data. In *CIKM '12: Proceedings of the 2012 ACM international Conference on Information and Knowledge Management*, pages 2689–2691, New York, NY, USA, 2012. ACM.

[11] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *CIKM '09: Proceedings of the 2009 ACM international Conference on Information and Knowledge Management*, pages 375–384, New York, NY, USA, 2009. ACM.

[12] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: joint models of topic and author community. In *ICML '09: Proceedings of the 2009 annual International Conference on Machine Learning*, pages 665–672, New York, NY, USA, 2009. ACM.

[13] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180, New York, NY, USA, 2007. ACM.

[14] G. Mitra and L. Mitra. *The handbook of news analytics in finance /*. Wiley ;, Hoboken, N.J. :, 2011.

[15] G. Pomper. *The election of 2000: reports and interpretations*. ELECTION OF. Chatham House Publishers, 2001.

[16] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[17] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL '08: Proceedings of the 2008 annual meeting on Association for Computational Linguistics*, pages 308–316, Columbus, Ohio, 2008. Association for Computational Linguistics.

[18] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference*, pages 424–433, New York, NY, USA, 2006. ACM.

[19] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR '06: Proceedings of the 2006 international ACM SIGIR conference on research and development in Information Retrieval*, pages 178–185, New York, NY, USA, 2006. ACM.