InCaToMi: Integrative Causal Topic Miner Between Textual and Non-textual Time Series Data

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

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

Thomas A. 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

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

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

ABSTRACT

Topic modeling is popular for text mining tasks. Recently, topic modeling has been combined with time lines when textual data is related to external non-textual time series data such as stock prices. However, no previous work has used the external non-textual time series data in the process of topic modeling. In this paper, we describe a novel text mining system, Integrative Causal Topic Miner (InCaToMi) that integrates textual and non-textual time series data. InCaToMi automatically finds causal relationships and topics using text data and external non-textual time series data using Granger Testing. Moreover, InCaToMi considers the non-textual time series data in the topic modeling process, using the time series data to iteratively improve modeling results through interactions between it and the textual data at both topic and word levels.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: [Miscellaneous]; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*

General Terms

Algorithm

Keywords

Causal Topic Mining, Integrative Topic Mining, Time series

1. INTRODUCTION

The Web 2.0 environment results in vast amounts of text data published daily. Analyzing and understanding the text data can provide useful information to individuals, industry and government. In a dynamic environment, the information content of the data can change through time. Further, topics identified in the text data may be related to external non-textual time series data. For example, news about companies can affect stock prices. Researchers may be interested in how particular topics lead to increasing or decreasing prices and use the relationships to forecast future price changes. This kind of causal analysis may be of interest in many domains. Consider election campaigns. If a survey shows declining support for a candidate, the campaign would want to understand why. Analyzing topics in the news in conjunction with poll numbers and understanding the causal relationships could inform and improve the campaign strategy.

Unfortunately, no existing text mining system can support integrative analysis of causal topics. There has been little research on how to discover causal topics from text based on external time series data. The closest system we know of is [11]. However, topic modeling systems/methods can only discover coherent topics in text. These topics are not necessarily correlated with non-textual time series.

In this paper, we present a novel text mining system called Integrative Causal Topic Miner (InCaToMi) that integrates textual and non-textual time series data. InCaToMi automatically finds causal topics from text data, compares topics to the external non-textual time series using Granger Testing [6], and allows users to explore causal topics with an interactive interface. The most important contribution of this system is a novel integrative mining approach that naturally combines topic modeling with time series causal analysis. This allows discovery of coherent topics in text that can either explain or be explained by a non-textual time series variable.

2. RELATED WORK

There are two common representative techniques in topic modeling: Probabilistic Latent Semantic Analysis (PLSA) [7] and Latent Dirichlet Analysis (LDA) [4]. Both focus on word co-occurrences. Recent advanced topic modeling techniques analyze text data on a time line [2, 10] by adding time-related variables to the model. Although previous systems may be able to show external series by overlapping them with topic modeling results, they do not conduct integrative analyses of topics and external variables; the topic analysis is separate from the external time series.

Some research incorporates external knowledge in the modeling process. Supervised Topic Modeling [3] uses topics with training data and response variables that may be used for prediction. Topic Sentiment Mixture modeling [8] incorporates background knowledge in the process using Conjugate Prior Probability. However, neither uses signals from external time series data to help model topics.

For the purpose of identifying causality in time-series data Granger testing [6] is common in economics. This technique tests for causality using the lead/lag relationships between 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.

Copyright is held by the author/owner(s).
CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
ACM 978-1-4503-1156-4/12/10.

3. ITERATIVE CAUSAL TOPIC MINING

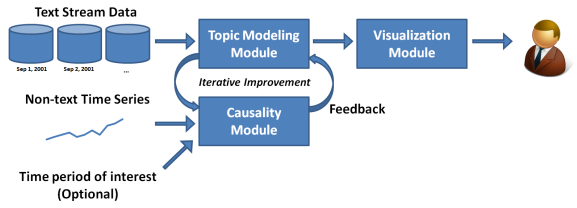

Figure 1: Overview of InCaToMi system

In contrast to other general topic modeling systems, InCaToMi considers the non-textual time series data in the text mining process and identifies topics that are correlated with the non-textual data. InCaToMi iteratively improves modeling results through interactions with the non-textual time series on both topic and word levels. First, InCaToMi finds causal topics. Then, it uses a word level analysis to check causality between word frequency time series and the external non-textual data. It improves the topic modeling process by splitting positively and negatively correlated words into different topics. Because generating and testing all the word time series is inefficient, InCaToMi focuses only on the top probability words of causal topics.

When users login to the InCaToMi system, they can upload and manage their own collection of data sets. By combining text and non-text data sets, they can trigger analysis tasks. For each analysis task started, progress and results are reported with visualization. Figure 1 shows a general overview of the InCaToMi system. The system has three main components: the topic modeling module, the causality module and the visualization module.

The **topic modeling module** takes text input and models topics. We use a PLSA topic model as our basic topic discovery method, here. The user specifies the number of topics. Each topic is a list of words with probabilities. Based on word probabilities, we find likelihoods of each topic in each document. For each topic and each day, we sum (across documents) the likelihoods to create an overall topic indicator (or "count") for each day. The counts give a time series data stream for each topic.

The **causality module** tests the causal relationships between each topic modeled and the non-textual time series data using Granger tests. Granger testing performs statistical significance tests for one time series "causing" the other series with different time lags using auto-regression. Let y_t and x_t be two sets of time series data, and we want to test whether x_t Granger causes y_t with a maximum p time lag. Granger testing involves estimating the 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, the test is an F-test, evaluating whether the lagged x terms are significant.

Granger tests require stationary time series. In our example application below, we take first differences ($(x_t) - (x_{t-1})$) or first differences in logged counts ($log(x_t) - log(x_{t-1})$) as input time series to make them stationary. We test causality up to 5 lags with a day as a time unit. For each topic which turns out to be causal, we also calculate the coefficient of the lagged topic stream (b_p) and reported this as an impact value. This value can give us additional information about the direction (positive/negativel impact) of the causality.

After the first execution of topic modeling and causality testing, the topics can be remodeled using **feedback** from the causality module. This refines the causal topic further resulting in higher correlations with the non-textual data. The main idea is to select some terms from a topic that are most correlated with the time series and form a new candidate topic (which improves the causal relation at the cost of potentially generating an incoherent topic). Then, try to make the topic more semantically coherent by using it as a prior for another round of topic modeling. This leads to the following general novel algorithm for iterative analysis of causal topics:

1. Model topic.
2. Find causal topics relative to non-textual series.
3. Check which of the top 100 words of each causal topic have causal relationships. For causal terms, check the 'impact' of the words.
4. Make a prior using causal terms and their impact values.
   (a) Separate positive impact terms and negative impact term groups. If one orientation is very weak (<10% of terms), ignore the minor group.
   (b) Assign prior probability proportions to the sizes of impact.
5. Re-model topics using prior.
6. Repeat 2-5.

The prior is like a small topic, a list of word and each word's probability. It provides a strong signal to make the modeling process similar to the prior topics [8]. In this way, the feedback procedure reinforces causal words be more dominant in the causal topics. Also, by separating positive and negative impact words, it drives more internally consistent topics.

We use causality between the non-textual series and both word streams and topic streams. The word level analysis gives us finer grain signals. However, generating all the word frequency time series and testing for causality is very inefficient. By focusing on causal topics, we can prune words to test. This achieves both efficiency and effectiveness. This iterative topic modeling algorithm can be generalized to use with any topic modeling techniques and causality/correlation measures.

Two statistics for each mined topic provide users with more information: causality confidence and topic purity. Causality confidence shows the level of assurance of causality. We use the p-value of Granger test between the external variable and the topic stream. Thus, it is the probability that the observed correlation is not random. We present mined topics with greater than 90% confidence. Topic purity measures the directional consistency of words within a topic. If all the causal words in a topic have same direction of impact, it would have 100% purity.

The **visualization module** provides results through an interactive interface. The Topic Summary Page shows topic trends with an interactive zoom interface and a causal topic list with their impact and measurement values (Figure 2). For further custom analysis, InCaToMi also provides a link which allows users to download the causal topic analysis results. Each topic shows the top three words. The linked Detail Page provides more topic details (Figure 3). Here, users can see the topic stream, causal words in the topic with impact factors, a list of all topic words and document samples. The Iteration Dash Board shows how topics change over iterations (Figure 4). Users can monitor modeling improvement or may choose one iteration and view the results.

4. SAMPLE RESULTS

To show the utility of the system, we test InCaToMi in two different areas. The first data set we use for this demo comes from the 2000 U.S. Presidential election campaign. We use New York Times articles from May through October of 2000. We use paragraphs that contain one or more of the key words 'Bush' or 'Gore.' Our goal is to find specific topics which changed the likelihood that one candidate or the other would win the election. For the non-textual time series, we use Democratic and Republican winner-takes-all prices from the Iowa Electronic Markets (IEM) [1].[1] Traders in this online futures market trade contracts with prices that reflect the probabilities of each candidate winning the popular vote. To insure that the prices form a valid measure, we follow the standard practice in the field and normalize daily closing (midnight) prices. Then, because

[1] `http://tippie.uiowa.edu/iem/`

the price of one candidate is always 1 minus the price of the other, a single candidate's price is a sufficient statistic in the market. We use pr(Gore):=(Gore Price)/(Gore Price + Bush Price).

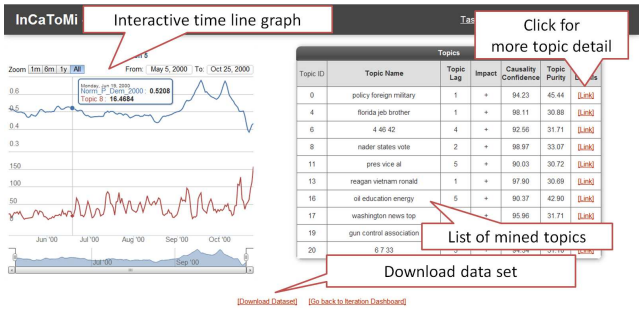

Figure 2: InCaToMi screen shot 1: Topic overview

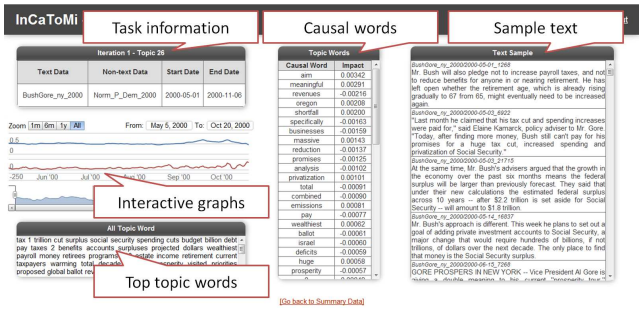

Figure 3: InCaToMi screen shot 2: Topic details

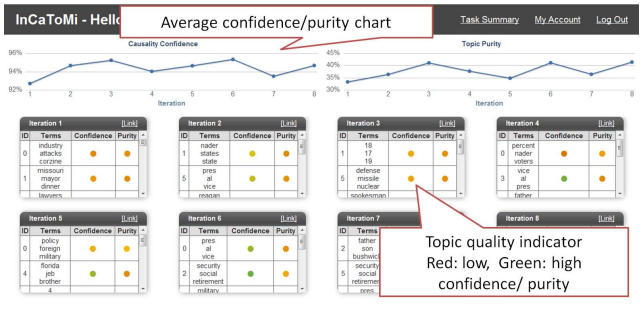

Figure 4: InCaToMi screen shot 3: Topic iteration dashboard

The mining results of iteration 1 (Figure 2) reveal several important issues that form causal topics in the 2000 U.S. Presidential elections, e.g., foreign policy, oil energy, education and gun control. Such topics are also cited in the political science literature [9] and Wikipedia.[2] The dash board (Figure 4) shows that more iterations increase causal confidence and topic purity.

The second data set shows usefulness of integrative topic mining more clearly. The task case is 'September 11 attack and stock prices'. The September 11 attack had a major impact on the US stock market resulting in large losses. We check whether InCaToMi can recover the event from text data and stock prices. We input August to October articles of New York Times news [3] and Dow Jones Industrial Average (DJI) closing prices for the same period. This task is more difficult than the previous case. For the presidential election case, because we used prefiltered text with candidate names, all the topics likely related to election campaign. However, for the September 11 case, we used all news articles over time period which cover a wide range of topics. Thus, it is more challenging to model the specific topic and select it as a causal topic.

Without feedback, simple topic modeling does not retrieve any September 11 related topics. However, after iterations of integrative modeling, InCaToMi includes 'israel, palestinian, bin laden, terrorism' and 'united state airline' as causal topics. This example clearly shows the benefit of integrative topic modeling.

In the demo, we present output of the two data sets with an interactive interface. For each, we first present a topic summary, and users can explore topic details by clicking detail view links. In the dashboard tab, user can check how topics change over iterations.

5. SUMMARY

We present a novel text mining system: Integrative Causal Topic Miner (InCaToMi) using textual and non-textual time series data. We develop a novel algorithm used in InCaToMi for finding causal topics with iterative causal topic quality improvement through interaction with external time series. We show the usefulness of the system in two cases. The demo system is publicly available.[4]

Although the basic usefulness is shown in this demo, we will be able to study causal topic mining algorithms more systematically and with variations in future work. In addition to identifying causal relationships post hoc, InCaToMi in real time might be useful for forecasting. Out of sample testing could be used to evaluate the prediction effectiveness of the model.

6. ACKNOWLEDGMENTS

Thanks to Riddhiman Ghosh (HP Labs) for helpful comments. This work is supported in part by an HP Innovation Research Award.

7. ADDITIONAL AUTHORS

Additional author: Carlos Ceja (HP Labs, email: carlos.ceja@hp.com).

8. 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] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference*, pages 153–162, New York, NY, USA, 2010. ACM.

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

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

[4] http://sifaka.cs.uiuc.edu/~hkim277/InCaToMi