A Novel TSA Based Method for Event Detection

Neelima Goyal[1], *Dr. Neetu Sharma*[2], *Dr. Rajesh Yadav*[3]
[1]*M.Tech Scholar, Department of Department of Computer Science and Engineering, GITM, Jhajjar*
[2]*Prof., Department of Department of Computer Science and Engineering, GITM, Jhajjar*
[3]*Asst. Prof., Department of Computer Science and Engineering, GITM, Jhajjar*

Abstract: With the rise of social media and online newswire, text streams are attracting more and more research interest. These streams are presented in the form of time series by nature, therefore, how to efficiently analyze these time series and extract useful information from them are of great importance. Modern time series analysis (TSA) has been applied widely in areas such as finance, physics and signal processing, however, there is not so much working exploring time series analysis in the field of text mining. While traditional time series analysis tasks are relatively well defined such as modeling and fore- casting, we now need to adapt the tasks to meet the requirement of different text mining problems. Event detection is the general task of finding any emerging events, such as significant changes in stock price, anomalies in climate data, and outbreaks of a certain disease, depending on the data we are interested in. While in text mining, event detection, which is identifying the significant new stories, is attracting more research attention given the increasing popularity of social media and digital journalism. In time series analysis, there is also a common task, change point detection, which focuses on a similar challenge. In this thesis work, we first examine the features presented by the time series of counts of terms in corpus. We then explore applying existing change point detection methods to event detection, and also propose a novel TSA based method for event detection.

Keywords: Social media, Time series, Forecasting

1. Introduction

Time series models have been the basis for any study of a behavior of process or metrics over a period of time. The applications of time series models are manifold, including sales forecasting, weather forecasting, inventory studies etc. In decisions that involve factor of uncertainty of the future, time series models have been found one of the most effective methods of forecasting. Most often, future course of actions and decisions for such processes will depend on what would be an anticipated result. The need for these anticipated results has encouraged organizations to develop forecasting techniques to be better prepared to face the seemingly uncertain future. Also, these models can be combined with other data mining techniques to help understand the behavior of the data and to be able to predict future trends and patterns in the data behavior. The evolving structure of interlinked documents, such as the World Wide Web or online citation indices, and the usage of these documents over a period of time has been of interest to both the researchers and the industry. These set of documents form a graph, with the nodes

representing the documents and the edges representing the hyperlinks or the citations. Research has been carried out in extracting information from the pure structure of such graphs and also on the usage of these documents, especially with respect to the World Wide Web. The stability of the Web structure has led to the more research related to Hyperlink Analysis and the field gained more recognition with the advent of Google [**Error! Reference source not found.**]. A survey on Hyperlink Analysis is provided in [1]. Usage aspects of such documents have also received wide attention and Srivastava et al [2] provide a good overview of Web usage mining research ideas and its applications.
Most research has thus focused more recently on mining information from structure and usage of such graphs. In this study we focus on another important dimension of mining such graphs as identified [3] - the Temporal Evolution. The Web is changing fast over time and so is the user’s interaction in the Web suggesting the need to study and develop models for the evolving Web Content, Web Structure and Web Usage. Also, of interest have been how the citation structure of research papers changes over time and how the access patterns of these papers vary over time.

The need to study the Temporal Evolution of the interlinked document has motivated us to analyze the various time series models that can be applied to them. We study the various the time series models that exist and the kind of data they are suitable to apply to. We also discuss some of the forecasting methods that are currently used. Figure 1 depicts the idea of change of such interlinked documents over time.

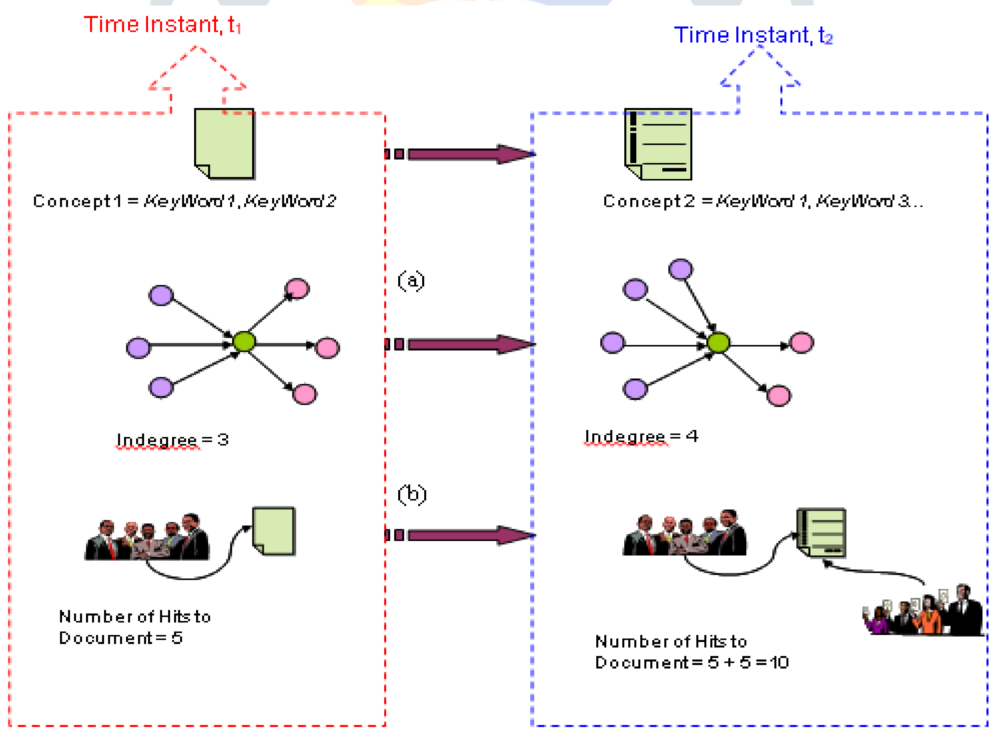


Figure 1: Temporal Evolution of a single document

(a) Change in the *Content* of a document over time

(b) Change in the *Structure* i.e. number of inlinks and outlinks; of a document over time

(c) Change in the *Structure* i.e. number of inlinks and outlinks; of a document over time

As part of our experiment, we try some of fitting in some of these models to the publicly available KDD Cup data that consists of research papers from high-energy physics. The rest of the document is organized as follows. In the section 2 we briefly describe the various time series methods that exist. Time Series Forecasting models and forecasting methods are discussed in this section 3. In section 4 we describe the data and the experimental set up. The results of the experiments are presented in section 5. Finally we provide the same conclusions and future directions.

2. Time Series Analysis Techniques

Time Series can be defined as an ordered sequence of values of a variable at equally spaced time intervals [4]. The motivation to study time series models is twofold:

- Obtain an understanding of the underlying forces and structure that produced the observed data
- Fit a model and proceed to forecasting, monitoring or even feedback and feedforward control.

Time Series Analysis can be divided into two main categories depending on the type of the model that can be fitted. The two categories are:

- Kinetic Model: The data here is fitted as xt= f(t). The measurements or observations are seen as a function of time.
- Dynamic Model: The data here is fitted as xt= f(xt-1 , xt-2 , xt-3 …).

The classical time series analysis procedures decomposes the time series function xt = f(t) into up to four components [5]:

1. Trend: a long-term monotonic change of the average level of the time series.
2. The Trade Cycle: a long wave in the time series.
3. The Seasonal Component: fluctuations in time series that recur during specific time periods.
4. The Residual component that represents all the influences on the time series that are not explained by the other three components.

The Trend and Trade Cycle correspond to the smoothing factor and the Seasonal and Residual component contribute to the cyclic factor. Often before time series models are applied, the data needs to be examined and if necessary, it has to be transformed to be able to interpret the series better. This is done to stabilize the variance. For example, if there is a trend in the series and the standard deviation is directly proportional to the mean, then a logarithmic transformation is suggested. And in order to make the seasonal affect addictive, if there is a trend in the series and the size of the seasonal effect tends to increase with the mean then it may be advisable it transform the data so as to make the seasonal effect constant from year to year. Transformation is also applied sometimes to make the data normally distributed.

The fitting of time series models can be an ambitious undertaking. There are many methods of model. These models have been well discussed in [6, 7]. The user's application and preference will decide the selection of the appropriate technique. We will now discuss some of the existing methods in time series analysis.

2.1 Smoothing Methods

Inherent in the collection of data taken over time is some form of random variation. There exist methods for reducing of canceling the effect due to random variation. An often-used technique in industry is "smoothing". This technique, when properly applied, reveals more clearly the underlying trend, seasonal and cyclic components. There are two distinct groups of smoothing methods: Averaging Methods and Smoothing Methods.

2.2 Averaging Methods

The "simple" average or mean of all past observations is only a useful estimate for forecasting when there are no trends. The average "weighs" all past observations equally. In general:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \left(\frac{1}{n}\right)x_1 + \left(\frac{1}{n}\right)x_2 + \ldots + \left(\frac{1}{n}\right)x_n$$

The (1/n) are the weights associated with each value of x. As we can see , these weights are normalized and sum upto 1.

An alternative way to summarize the past data is to compute the mean of successive smaller sets of numbers of past data. This smoothing process is continued by advancing one period and calculating the next average of t numbers, dropping the first number. Such type of averaging is called Single Moving Average and the general expression for the moving average is

$$\boldsymbol{M_t = [\, X_t + X_{t-1} + \ldots + X_{t-N+1}] / N}$$

There exists a variation on the MA procedure that often does a better job of handling trend. It is called Double Moving Averages for a Linear Trend Process. It calculates a second moving average from the original moving average, using the same value for M. As soon as both single and double moving averages are available, a computer routine uses these averages to compute a slope and intercept, and then forecasts one or more periods ahead.

2.1.2 Exponential Smoothing Methods

This is a very popular scheme to produce a smoothed Time Series. Whereas in Single Moving Averages the past observations are weighted equally, Exponential Smoothing assigns exponentially decreasing weights as the observation get older. In other words, recent observations are given relatively more weight in forecasting than the older observations. In the case of moving averages, the weights assigned to the observations are the same and are equal to 1/N. In exponential smoothing, however, there are one or more smoothing parameters to be determined (or estimated) and these choices determine the weights assigned to the observations. This smoothing scheme begins by setting S2 to y1, where Si stands for smoothed observation or EWMA, and y stands for the original observation. The subscripts refer to the time periods, 1, 2, ..., n. For the third period, S3 = αy2 + (1-α) S2; and so on. There is no S1; the smoothed series starts with the smoothed version of the second observation. For any time period t, the smoothed value St is found by computing

$$S_t = \alpha y_{t-1} + (1-\alpha) S_{t-1}, \quad 0 < \alpha \leq 1, \; t \geq 3$$

This is the basic equation of exponential smoothing and the constant or parameter α is called the smoothing constant. The speed at which the older responses are dampened (smoothed) is a function of the value of α. When $\alpha$is close to 1, dampening is quick and when $\alpha$is close to 0, dampening is slow. We choose the best value for α so the value which results in the smallest Mean Squared Error.

3. Time Series Models and Forecasting

Time series Models and forecasting methods have been studied by various people and detailed analysis can be found in [8, 9,11]. Time Series Models can be divided into two kinds. Univariate Models where the observations are those of single variable recorded sequentially over equal spaced time intervals. The other kind is the Multivariate, where the observations are of multiple variables. A common assumption in many time series techniques is that the data are stationary. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. Stationarity can be defined in precise mathematical terms, but for our purpose we mean a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations. There are a number of approaches to modeling time series. We outline a few of the most common approaches below.

3.1 Trend, Seasonal, and Residual Decompositions: One approach is to decompose the time series into a trend, seasonal, and residual component. Triple exponential smoothing is an example of this approach. Another example, called seasonal loess, is based on locally weighted least squares.

3.2 Frequency Based Methods: Another approach, commonly used in scientific and engineering applications, is to analyze the series in the frequency domain. An example of this approach in modeling a sinusoidal type data set is shown in the beam deflection case study. The spectral plot is the primary tool for the frequency analysis of time series. Autoregressive (AR) Models: A common approach for modeling univariate time series is the autoregressive (AR) model:

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + A_t$$

where Xt is the time series, At is white noise, and

$$\delta = (1 - \sum_{i=1}^{p} \phi_i)\mu$$

with μ denoting the process mean.

An autoregressive model is simply a linear regression of the current value of the series against one or more prior values of the series. The value of p is called the order of the AR model. AR models can be analyzed with one of various methods; including standard linear least squares techniques. They also have a straightforward interpretation.

3.3 Moving Average (MA): Models another common approach for modeling univariate time series models is the moving average (MA) model:

$$X_t = \mu + A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \ldots - \theta_q A_{t-q}$$

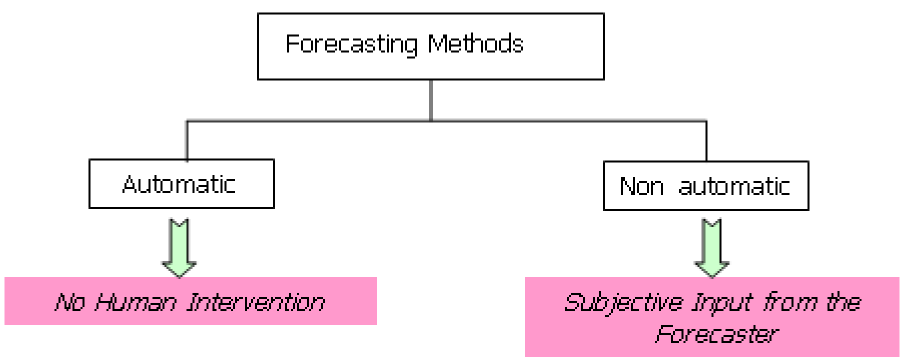
where Xt is the time series, μ is the mean of the series, At-i are white noise, and 1, ... , q are the parameters of the model. The value of q is called the order of the MA model. That is, a moving average model is conceptually a linear regression of the current value of the series against the white noise or random shocks of one or more prior values of the series. The random shocks at each point are assumed to come from the same distribution, typically a normal distribution, with location at zero and constant scale. The distinction in this model is that these random shocks are propagated to future values of the time series. Fitting the MA estimates is more complicated than with AR models because the error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares. MA models also have a less obvious interpretation than AR models. Note, however, that the error terms after the model is fit should be independent and follow the standard assumptions for a univariate process.

$$\begin{aligned} \mathrm{X}_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \\ A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \ldots - \theta_q A_{t-q} \end{aligned}$$

-Jenkins Approach: The Box-Jenkins ARMA model is a combination of the AR and MA models where the terms in the equation have the same meaning as given for the AR and MA model [8]. The Box-Jenkins model assumes that the time series is stationary. Box and Jenkins recommend differencing non-stationary series one or more times to achieve stationarity. Doing so produces an ARIMA model, with the "I" standing for "Integrated". Some formulations transform the series by subtracting the mean of the series from each data point. This yields a series with a mean of zero. Whether you need to do this or not is dependent on the software you use to estimate the model. Box-Jenkins models can be extended to include seasonal autoregressive and seasonal moving average terms. Although this complicates the notation and mathematics of the model, the underlying concepts for seasonal autoregressive and seasonal moving average terms are similar to the non-seasonal autoregressive and moving average terms. The most general Box-Jenkins model includes difference operators, autoregressive terms, moving average terms, seasonal difference operators, seasonal autoregressive terms, and seasonal moving average terms. As with modeling in general, however, only necessary terms should be included in the model.

4. Forecasting Methods

The main objective of forecasting for a given series x1, x2, x3,….. xN ; to estimate future values such as xN+k , where the integer k is called the lead time [6]. The forecast of xN+k made at a time N for k steps ahead is denoted by $\hat{x}(N,k)$.



Error! Reference source not found.**: Depicts a classification of Forecasting Methods based on the kind of approach**

A straight line model is used to relate the time series, Yt, to time, t, and the least squares line is used to forecast he future values of Yt.

$$E(Yt) = \beta 0 + \beta 1t.$$

It is risky to use a least squares regression model outside the experimental region, especially for prediction purposes. Cyclical or Trade effects like the effects of an inflation or recession are not included.

Among the other forecasting techniques that are based on time series models and techniques discussed earlier we present only the one based on exponentially smoothing as it fits in most cases and we have also implemented the model, in our experiment.

The exponential smoothed forecast for Yt+1 is the smoothed value at time t.

$$F_{t+1} = E_{t},$$

where F_{t+1} is the forecast of $Y_{t+1.}$

$$F_{t+1} = E_t = wY_t + (1-w)\, E_{t-1.}$$

$$= wY_t + (1-w)\, F_{t.}$$

$$= F_t + w(Y_t - F_t).$$

4.1 Exponential smoothed forecast are appropriate only when trend and seasonal components are relatively insignificant. Smoothed values will tend to lag behind when a long-term trend exists. Averaging tends to smooth any seasonal component.

The most popular forecasting technique is the Holt-Winters Forecasting Technique [11]. It consists of both exponential component (Et) and a trend component (Tt).

The calculation time begins at t=2, because the first two observations are needed to obtain the first estimate of trend T2.

$$E_2 = Y_{2.}$$

$$T_2 = Y_2 - Y_{1.}$$

$$E_t = wY_t + (1-w)\,(E_{t-1} + T_{t-1}),\ 0<w<1.$$

$$T_t = v(E_t - E_{t-1}) + (1 - v)T_{t-1.},\ 0<v<1.$$

Note: 'v' closer to zero suggests more weight to past estimates of trend, and 'v' value closer to one suggests more weight to current change in level.

Firstly, the exponentially smoothed and trend components, Et and Tt, for each observed value of Yt(t>=2) are calculated. The one-step-ahead forecasting is determined using.

$$F_{t+1} = E_t + T_{t.}$$

And the k-step-ahead forecast using:

$$F_{t+1} = E_t + kT_t$$

5. Problem Formulation

With the rise of social media and online newswire, text streams are attracting more and more research interest. These streams are presented in the form of time series by nature, therefore, how to efficiently analyze these time series and extract useful information from them are of great importance. Modern time series analysis (TSA) has been applied widely in areas such as finance, physics and signal processing, however, there is not so much working exploring time series analysis in the field of text mining. While traditional time series analysis tasks are relatively well defined such as modeling and fore- casting, we now need to adapt the tasks to meet the requirement of different text mining problems.

Event detection is the general task of finding any emerging events, such as significant changes in stock price, anomalies in climate data, and outbreaks of a certain disease, depending on the data we are interested in. While in text mining, event detection, which is identifying the significant new stories, is attracting more research attention given the increasing popularity of social media and digital journalism. In time series analysis, there is also a common task, change point detection, which focuses on a similar challenge

6. Proposed Solution

In this work, first examine the features presented by the time series of counts of terms in corpus. We then explore applying existing change point detection methods to event detection, and also propose a novel TSA based method for event detection.

Framework

We now present our proposed method. Instead of using a square loss like conventional fused lasso, we use a negative log likelihood of generating the documents as our loss function to minimize [92]. The negative log likelihood together with the regularization term forms our objective function. The fused lasso is applied to the scenario where sparsity lies in the difference of coefficients. This enhances the local consistency in coefficients.

Batch Setting

In batch setting, we want to detect the _rst stories in the corpora with all data available.

Event Detection Score

In TDT or FSD tasks, we typically need to assign a score for the detected documents. The higher the score is, the more confidence we put in it as a first story. For example in [79], this score of a document is called novelty score, which is the document's distance from the closest previous document. The higher this score is, the more different the document is from any previous document. Therefore, this distance is a valid score.

Dataset

We first describe the dataset we use in our experiments. We have three time-stamped data sets, including CNN TV transcripts and standard TDT5 dataset. To generate time series, we use day as our time unit and first put all the stories of the same day together, as described previously. Then we do stemming and generate term counts for each day with normalization. We will introduce the datasets respectively.

7. Conclusion

In this thesis work, we explore the application of time series analysis in the problem of event detection in text mining. We compare the similarities and differences between traditional TSA and event detection, and propose a novel method in batch setting. Experiments on various datasets show that our method can generate performance which is competitive with other baseline methods.

There are also many aspects where our method can be improved and problems that are worthy of study in the future. First of all, we can extend our batch setting to online setting to better detect events in a real time fashion, modifying the updating process in gradient descent. Another potential direction is applying our method to other time series data. Since change point detection is an important problem in many fields of research, it is meaningful that we adapt our algorithm to take new challenges from other disciplines.

References

1. P. Desikan, J. Srivastava, V. Kumar, P.-N. Tan, “Hyperlink Analysis – Techniques & Applications”, Army High Performance Computing Center Technical Report, 2014.
2. J. Srivastava, R. Cooley, M. Deshpande and P-N. Tan. “Web Usage Mining: Discovery and Applications of usage patterns from Web Data”, SIGKDD Explorations, Vol 1, Issue 2, 2010.
3. J. Srivastava, P. Desikan and V. Kumar, “Web Mining – Accomplishments and Future Directions”, Invited paper in National Science Foundation Workshop on Next Generation Data Mining, Baltimore, MD, Nov. 1-3, 2002.
4. *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/
5. J.T. McClave, P.G. Benson and T. Sincich, “Statistics for Business and Economics”, Prentice Hall, 2011
6. C. Chatfield, “The Analysis of Time Series – An Introduction”, Chapman and Hall, 1996.
7. P.J. Brockwell and R.A. Davis. (1987). Time Series: Theory and Methods, Springer-Verlang
8. G. E. P Box, G. M. Jenkins, and G. C. Reinsel. (1994).Time Series Analysis, Forecasting and Control, 3rd ed. Prentice Hall, Englewood Clifs, NJ.
9. P.J. Brockwell and R.A. Davis, (2002). Introduction to Time Series and Forecasting, 2nd. ed., Springer-Verlang.
10. KDD Cup 2003, http://www.cs.cornell.edu/projects/kddcup/index.html.
11. R.Yaffee “A. Introduction to time series analysis and forecasting with applications of SAS and SPSS /.” San Diego: Academic Press, c2015.