

Detecting Changes and Anomalies in Noisy Text Streams

Jerry Wright

Networking and Services Research Lab
AT&T Labs — Research

15 February 2010

Outline

- CoCITe
- Noise
- Mixture Distributions
- Results

Mining Text Streams for Changes

Text Stream

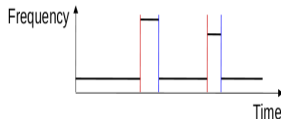
Time-stamped ascii text, usually structured into *documents* (optionally tagged with *metadata*), and containing recurrent *words*

Words may be tokenized:

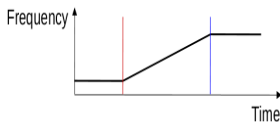
- Normalize case and punctuation
- Substitute tokens for named-entities

Frequency of words as function of time:

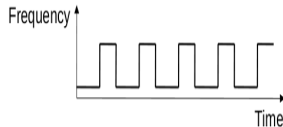
Steps and bursts



Trends



Cycles



“We’re seeing more of *this* and less of *that*, especially for *these* customers.”

Model-Based Approach

Binning

Documents binned and frequencies counted at regular intervals (typically hourly or daily)

Assumption: Documents are independent

Absolute Frequency (to track raw word-count)

Number of occurrences of word in bin at t is $\text{Poisson}(\lambda_t)$, where λ_t is piecewise-linear function of time with cyclic modulation

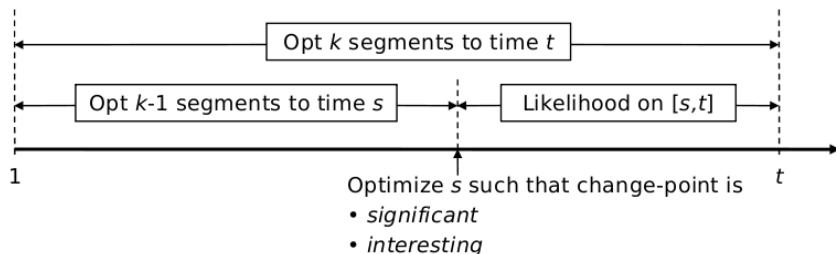
Relative Frequency (to track proportion of documents containing word)

Number of documents in bin at t containing word is $\text{Binomial}(n_t, p_t)$, where n_t is total number of documents in bin at t ,
 p_t is piecewise-linear function of time with cyclic modulation

Optimization of Model

Piecewise-Linear Segmentation

Dynamic programming algorithm to maximize likelihood

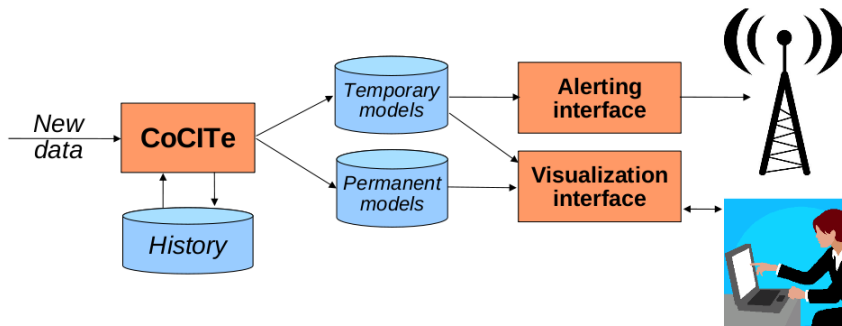


Periodic Model

Periodicity test

Number and assignment of modulation coefficients

Stream Implementation



Condensed History

Used for model re-optimization for each bin
Mostly geometrically-weighted totals

Outline

- CoCITe
- Noise
- Mixture Distributions
- Results

Noise

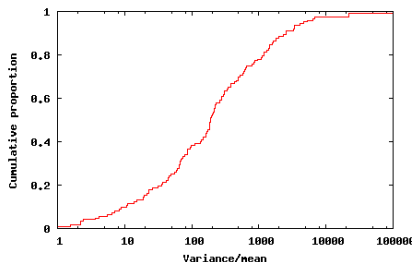
Word Occurrence Frequencies Are Noisy (*Over-Dispersed*)

Additional to steps, trends, cycles

More bin-to-bin variation than Poisson and Binomial models can account for

Absolute Frequency

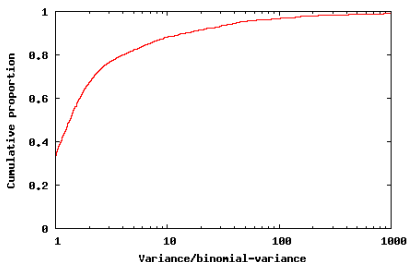
Poisson: variance = mean



(Data from a threat management system)

Relative Frequency

Binomial: variance < mean

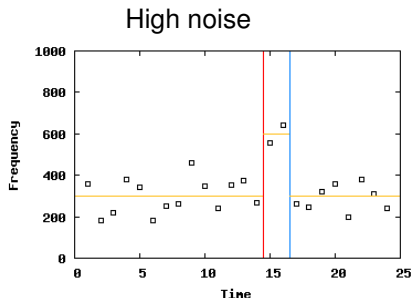
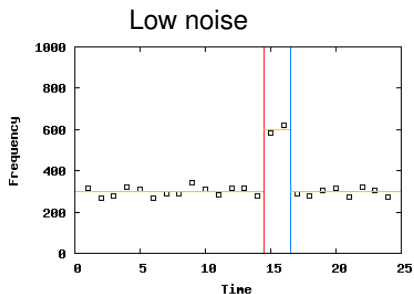


(Data from a CHI Scan customer care app)

Impact On Change-Detection

Noise Weakens Significance

Significance P-value governs: number of segments discovered,
ranking of alerts



Approaches to Noise

Filter and Attenuate



- 😊 Cheap
- 😞 Attenuates signal as well as noise

Adapt and Mitigate



- 😞 Expensive
- 😊 Clearer perception of desired signal

Outline

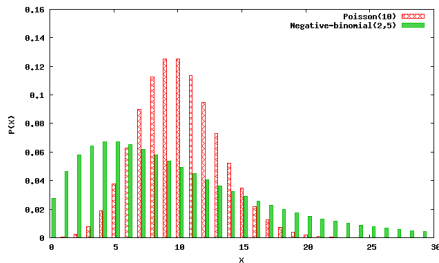
- CoCITe
- Noise
- Mixture Distributions
- Results

Gamma-Poisson Mixture (Negative Binomial)

Absolute Frequency (to track raw word-count)

Number of occurrences of word in bin at t is $\text{Poisson}(\Lambda_t)$,
 where $\Lambda_t \sim \gamma(\mu_t/\theta_t, \theta_t)$,

where μ_t is piecewise-linear function of t with cyclic modulation
 θ_t controls dispersion (slowly varying)



$$P(X = x) = \frac{\Gamma(\mu/\theta + x)\theta^x}{x!\Gamma(\mu/\theta)(1 + \theta)^{\mu/\theta + x}}$$

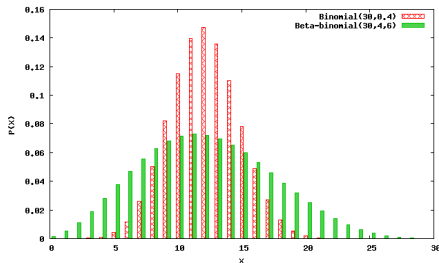
$$P(X \leq x) = I_{1/(1+\theta)}(\mu/\theta, x + 1)$$

(regularized incomplete beta function)

Beta-Binomial Mixture

Relative Frequency (to track proportion of documents containing word)

Number of documents in bin at t containing word is binomial(n_t, P_t),
 where $P_t \sim \beta(p_t/\theta_t, (1 - p_t)/\theta_t)$,
 where p_t is piecewise-linear function of t with cyclic modulation
 θ_t controls dispersion (slowly varying)

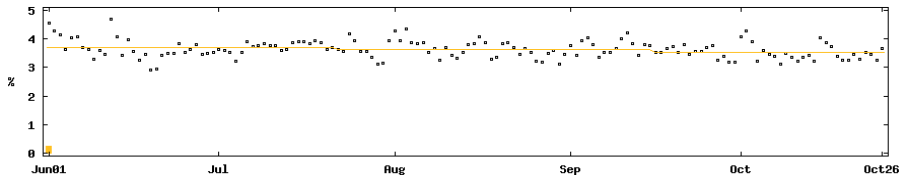


$$P(x) = \binom{n}{x} \frac{B(p/\theta + x, (1 - p)/\theta + n - x)}{B(p/\theta, (1 - p)/\theta)}$$

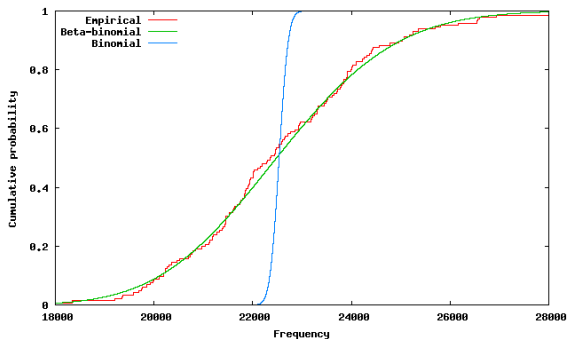
where $B()$ is the complete beta function

$P(X \leq x)$ is ugly

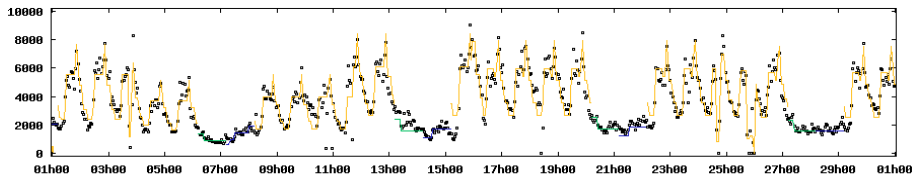
Goodness of Fit of Beta-Binomial



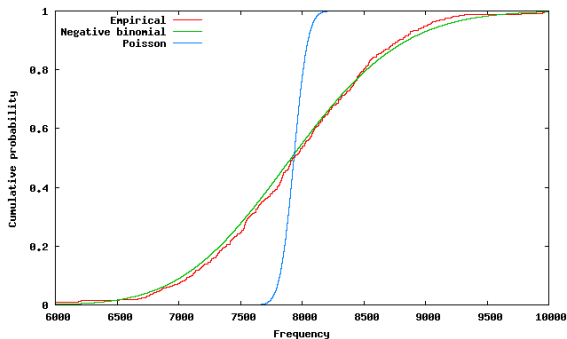
Data from a CHI Scan
customer care app,
 χ^2 not significant



Goodness of Fit of Negative Binomial



Data from a threat management system, scaled to "iid" sequence using periodic model, χ^2 not significant



Implementation

- Test for over-dispersion
 - Poisson — Dean-Lawson statistic (1989)
 - Binomial — Tarone statistic (1979)

Implementation

- Test for over-dispersion
 - Poisson — Dean-Lawson statistic (1989)
 - Binomial — Tarone statistic (1979)
- Likelihood — use probability mass function

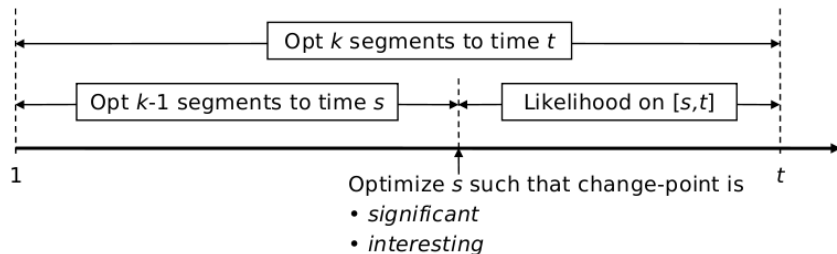
Implementation

- Test for over-dispersion
 - Poisson — Dean-Lawson statistic (1989)
 - Binomial — Tarone statistic (1979)
- Likelihood — use probability mass function
- Estimation of over-dispersion parameter θ_t
 - Moments estimates using geometrically-weighted sums over data
 - Suitable for stream implementation

Implementation

- Test for over-dispersion
 - Poisson — Dean-Lawson statistic (1989)
 - Binomial — Tarone statistic (1979)
- Likelihood — use probability mass function
- Estimation of over-dispersion parameter θ_t
 - Moments estimates using geometrically-weighted sums over data
 - Suitable for stream implementation
- Significance test
 - No standard tests and little prior art
 - Must be efficient ($\sim \mu s$)

Implementation



- Significance test
 - No standard tests and little prior art
 - Must be efficient ($\sim \mu s$)

Implementation

```
For each bin
  For each metavalue
    For each word
      For each t
        For each number of segments
          For each s
            Is it significant?
```

- Significance test
 - No standard tests and little prior art
 - Must be efficient ($\sim \mu s$)

Implementation

- Test for over-dispersion
 - Poisson — Dean-Lawson statistic (1989)
 - Binomial — Tarone statistic (1979)
- Likelihood — use probability mass function
- Estimation of over-dispersion parameter θ_t
 - Moments estimates using geometrically-weighted sums over data
 - Suitable for stream implementation
- Significance test
 - No standard tests and little prior art
 - Must be efficient ($\sim \mu s$)
 - CDFs used to obtain upper and lower bounds on P-value (allowing for variance of nuisance parameter), then weighted geometric mean

Implementation

- Test for over-dispersion
 - Poisson — Dean-Lawson statistic (1989)
 - Binomial — Tarone statistic (1979)
- Likelihood — use probability mass function
- Estimation of over-dispersion parameter θ_t
 - Moments estimates using geometrically-weighted sums over data
 - Suitable for stream implementation
- Significance test
 - No standard tests and little prior art
 - Must be efficient ($\sim \mu s$)
 - CDFs used to obtain upper and lower bounds on P-value (allowing for variance of nuisance parameter), then weighted geometric mean
- Measure of interest

Significance Test for Two Beta-Binomials

Comparing Two Binomials — Fisher's Exact Test

Using unknown common $P(A = 1) = p$,

$$P(\text{table}) = \binom{n_{01}}{n_{11}} \binom{n_{02}}{n_{12}} p^{n_{10}} (1 - p)^{n_{20}}$$

Conditioning on row totals, nuisance parameter p disappears:

$$P(\text{table} | n_{10}, n_{20}) = \binom{n_{01}}{n_{11}} \binom{n_{02}}{n_{12}} / \binom{n_{00}}{n_{10}}$$

Sum over tables with same row totals and no more likely than actual one $\rightarrow$ P-value

2×2 contingency table

		B		
		1	2	
A	1	n_{11}	n_{12}	n_{10}
	2	n_{21}	n_{22}	n_{20}
		n_{01}	n_{02}	

Comparing Two Beta-Binomials

Table probability is product of two beta-binomials, same nuisance parameter p .

Conditioning on row totals does not eliminate p .

Could use Barnard's test instead: For each p , sum over all tables no more likely than actual one, then maximize over $p \rightarrow$ P-value

Very slow!

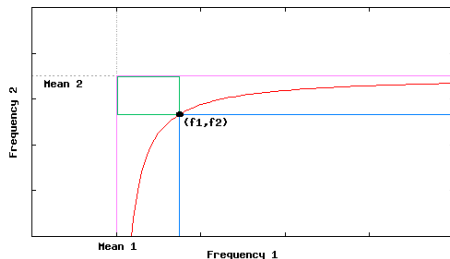
Fast Significance Test (Both Distributions)

Estimate common mean from data.

Allow for variance of this estimate: if r.v. Y is a function of r.v. X then

$\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}[E(Y|X)]$, and assume same family.

One observation must then be larger its expected mean and one smaller.



Critical region below red contour (probability equal to that for observed (f_1, f_2)).

Total mass of rectangular regions can be obtained quickly from product of CDFs.

Lower bound on P-value from blue rectangle.

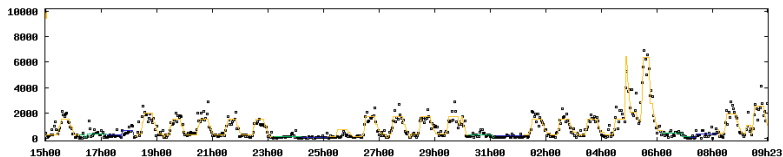
Upper bound from difference between purple and green rectangles.

Weighted geometric mean of upper and lower (tighter) bound.

Outline

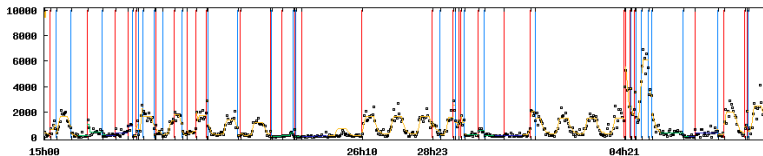
- CoCITe
- Noise
- Mixture Distributions
- **Results**

Example from Threat Management System Data



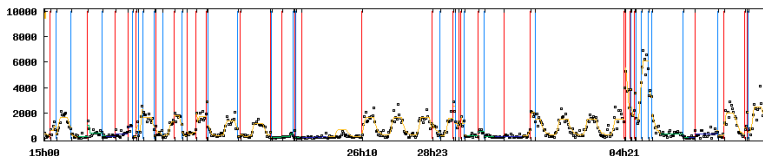
Example from Threat Management System Data

Absolute frequency using Poisson

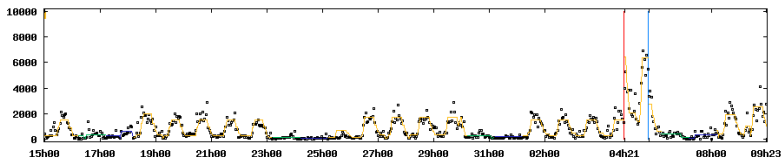


Example from Threat Management System Data

Absolute frequency using Poisson



Absolute frequency using Gamma-Poisson (Negative Binomial)

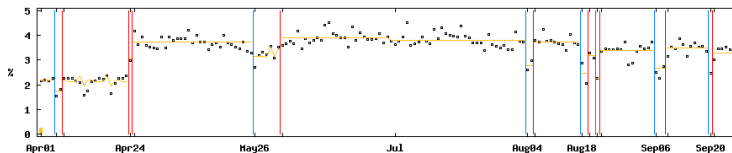


Change-point at 2009060421 significant at $\sim 10^{-20}$

Variance $\approx 144 \times \text{mean}$

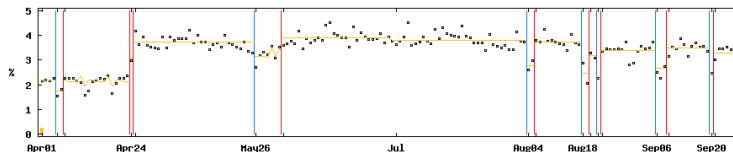
Example from Customer Care Data

Relative frequency using Binomial

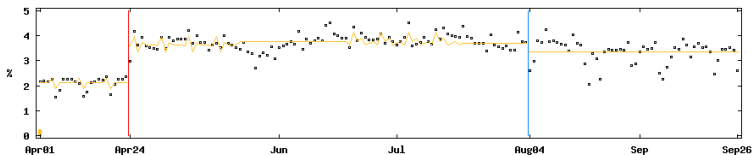


Example from Customer Care Data

Relative frequency using Binomial



Relative frequency using Beta-Binomial



Change-point at 20090424 significant at $\sim 10^{-19}$

Variance $\approx 233 \times$ binomial-variance

Examples from LCD Aquaint Newswire Corpus

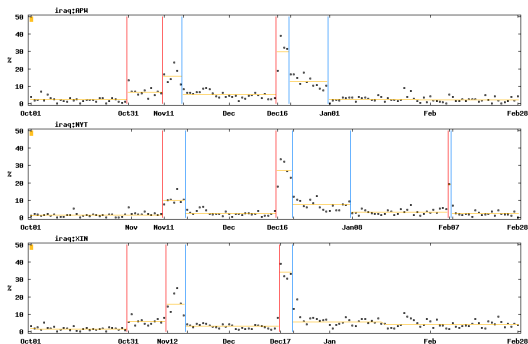
~800k news articles over 28 months (June 1998 — September 2000) from

- Associated Press Worldstream (APW)
- New York Times News Service (NYT)
- Xinhua News Service (XIN)

Bursts for Iraq
1998-1999:

Nov11 U.N. evacuation

Dec17 Start of military
action

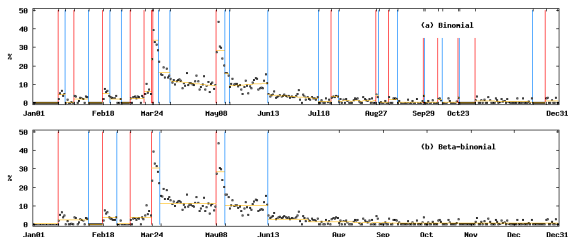


Examples from LCD Aquaint Newswire Corpus

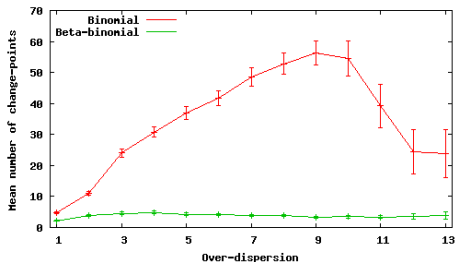
Yugoslavia 1999

(a) Binomial

(b) Beta-binomial



Number of change-points
during 1999 vs
over-dispersion for 9000
words



Summary

- CoCITe is looking for step changes, trends and bursts in word frequencies within text streams
- Cycles are an important source of inherent variation and must be allowed for
- Noise is another important source of inherent variation
- Mixture distributions model and mitigate this
- Clearer perception of desired signal

“We’re seeing more of *this* and less of *that*, especially for *these* customers.”

- Thanks to
 - Dave Kapilow, Alicia Abella, Patrick Haffner
 - Chaim Spielman, Dan Sheleheda, Dave Gross