Topics over Time:

A Non-Markov Continuous-Time Model of Topical Trends

Paper by Xuerui Wang and Andrew McCallum

Presented by Linda Buisman

Overview

- Motivation
- Approach
- Results
- Analysis
- Conclusion

Motivation

- Information retrieval & text mining
- Text is highly-dimensional
- Topic models
  - Discover summaries of documents
  - Reduce dimensions
  - Model co-occurrences of words
    - mouse, cat, Tweety → cartoons
    - mouse, keyboard → computer supplies
- Topics over time
  - Co-occurrences are dynamic
  - Additional modality – time
    - united, states, war @ 1850 → Mexican-American War
    - united, states, war @ 2006 → War in Iraq

Modelling time

- Earlier approaches
  - Discretize
    - Fixed interval size does not fit all topics
  - Markov model
    - State at time t+1 depends on t, but not earlier
- Solution
  - Treat time as a continuous variable
  - Time is a parameter in a Bayesian network

Bayesian network

- Generative model
  - vs discriminative (SVM, NN, …)
- Bayes' rule:

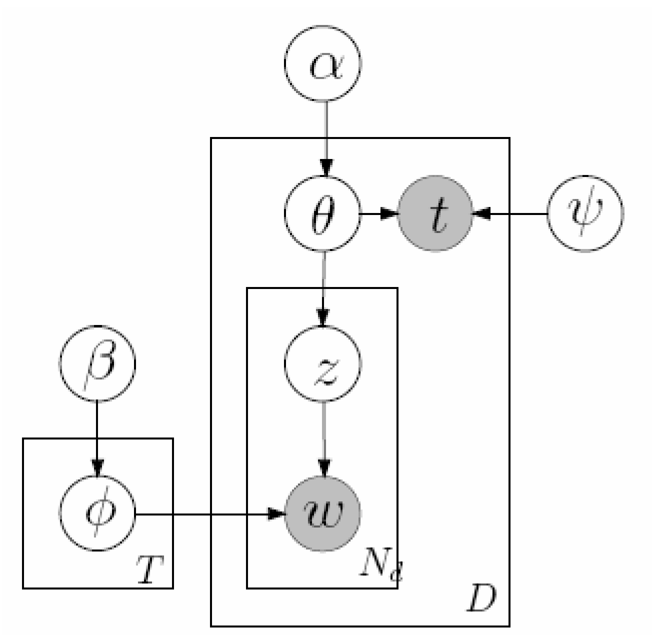
$$P(H \mid X) = \frac{P(X \mid H) \times P(H)}{P(X)}$$

- Bayesian network
  - Directed graph of parameters
- *A* connected to *B*:
  - Probability of *B* conditionally depends on *A*
- Generation step
  - Estimate conditional probabilities for all (hidden) parameters
- Goal
  - Predict probability of hypothesis *H* being true for observation *X*

Topics-over-time model

- Based on an earlier topic model LDA
- “Bag-of-words” approach
  - Word count in a document is significant
  - Position and order are not significant
- Timestamp of document becomes another parameter
- Generate Bayesian network from existing documents
  - Exact inference computationally infeasible
  - Use approximate inference
- Goal
  - Predict the probability of a document belonging to topic T

Model



Known parameters:

w	word
t	timestamp

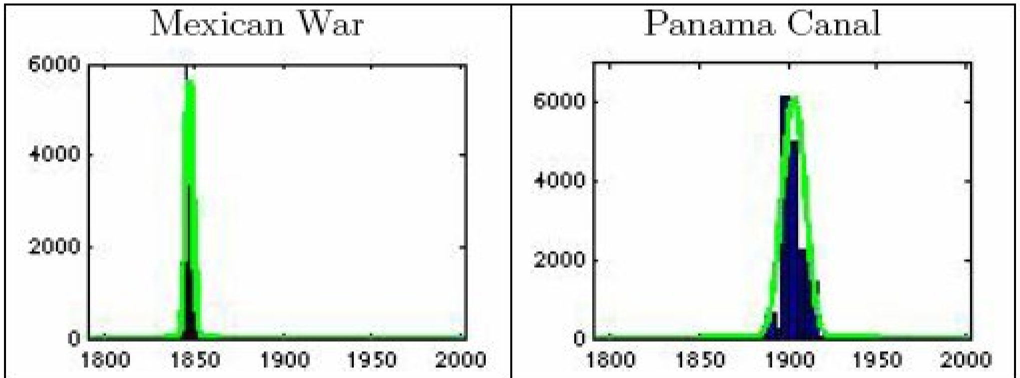
Hidden parameters:

z	topic associated with word
φ	distribution of words for topic
ψ	distribution of time for topic
θ	distribution of topics for document

Diagram from Wang & McCallum

Results

Distribution of topic over time

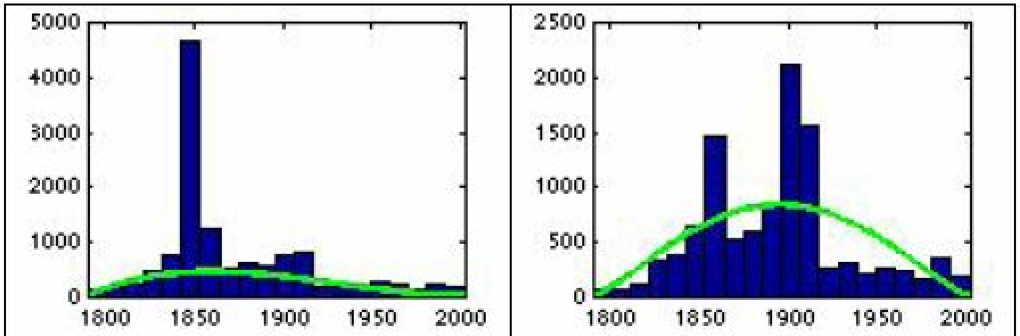


states	0.02032	government	0.02928
mexico	0.01832	united	0.02132
government	0.01670	states	0.02067
united	0.01521	islands	0.01167
war	0.01059	canal	0.01014
congress	0.00951	american	0.00872
country	0.00906	cuba	0.00834
texas	0.00852	made	0.00747
made	0.00727	general	0.00731
great	0.00611	war	0.00660

Words associated with a topic

Diagram from Wang & McCallum

Comparison with basic LDA



mexico	0.06697	government	0.05618
government	0.02254	american	0.02696
mexican	0.02141	central	0.02518
texas	0.02109	canal	0.02283
territory	0.01739	republic	0.02198
part	0.01610	america	0.02170
republic	0.01344	pacific	0.01832
military	0.01111	panama	0.01776
state	0.00974	nicaragua	0.01381
make	0.00942	isthmus	0.01137

Confuses Mexican War with WWI

Confuses Panama Canal with other activities in Central America

Diagram from Wang & McCallum

Analysis

- Generative vs discriminative methods
  - Discriminative usually faster
  - Accuracy depends on application
  - Generative model offers more information
    - E.g. not just topic(s) of a document, but also:
      - Predict time-stamp, given a document
      - Distribution of topics over time

Analysis (cont)

- Limitations and simplifications
  - “Bag-of-words” instead of word sequences or phrases
    - Computer science vs computer, science
  - No account of position within document
    - Title, introduction, body, footnote

Analysis (cont)

- General and flexible approach
- Possible extensions
  - Add time to Group-Topic and Author-Recipient models
  - Capture changes in group formation over time

Conclusion

- TOT = LDA + time modality
- Improves the detection of topics
- Adds other features
- Extensible
- No ground-breaking innovation
  - Rather, a useful addition to an existing method