Temporal Motif Mining With Sequential Mining Techniques

Wang Yuxin

Tutors: Baptiste Jeudy, Rémi Emonet

28/06/2019

Outline

What is motif mining?

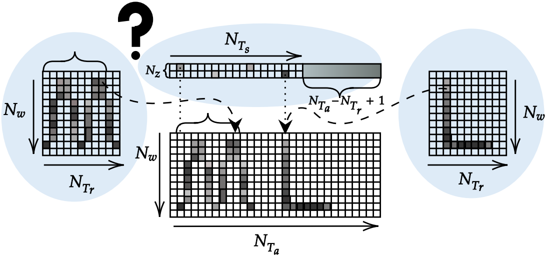


Figure 1: Generative Process

Elements

- **Temporal Document.** $N_{T_a} \times N_w$ table of counts.
- **Starting Time.** $N_z \times N_{T_s}$ table of probabilities.
- **Latent Motif.** $N_z \times N_w \times N_{T_r}$ table of probabilities.

What to solve

Posterior distribution

- Latent motif: $p(w, tr|z)$
- Starting Time: $p(z, ts|d)$

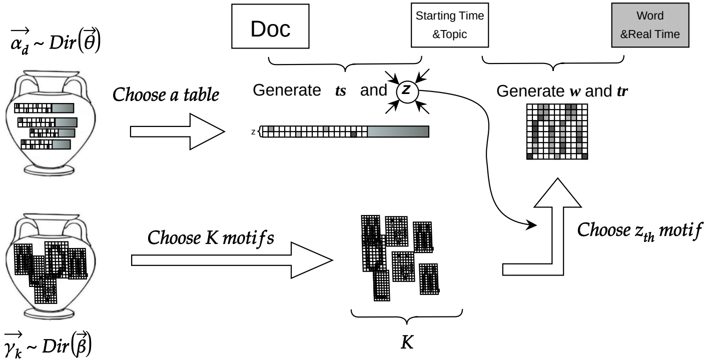


Figure 2: Document generation

Can we solve it?

Posterior distribution

- Latent motif: $p(w, tr|z) = \frac{p(w,tr,z)}{\sum_{w,tr} p(w,tr,z)}$
- Starting Time: $p(z, ts|d) = \frac{p(z,ts,d)}{\sum_{z,ts} p(z,ts,d)}$

Crux: denominator

- $p(z_i) : \sum_{k=1}^{N_w} \sum_{l=1}^{N_{T_r}} p(z_i|w = k, tr = l)p(z_i|w = k, tr = l)$
  - $p(\vec{z}) : \prod_{i=1}^{K} p(z_i) \Longrightarrow (N_w + N_{T_r})^K$
- $p(d_j) : \sum_{m=1}^{N_z} \sum_{n=1}^{N_{T_s}} p(d_j|z = m, ts = n)p(d_j|z = m, ts = n)$
  - $p(\vec{d}) = \prod_{j=1}^{N_d} p(d_j) \Longrightarrow (N_z + N_{T_s})^{N_d}$

EM Algorithm : Iterative Solution

- maximum likelihood estimation?
  - Natural idea. But no $\Longleftarrow$ *latent variables*
- Jensen inequality $\Longrightarrow$ *Lower Bound*

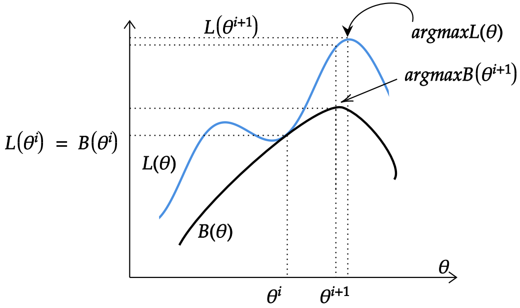


Figure 3: EM Algorithm

What we want

Limitation of EM

- Sensitive to the initialization
- Initialization affects the final result

What to improve

- Prior distribution

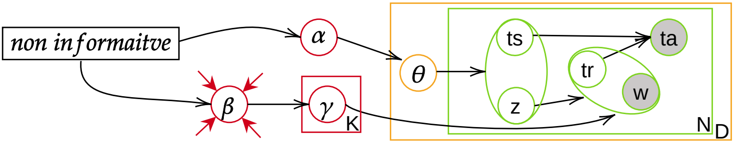


Figure 4: Graphical Model

How to do

Overview

Integrate the sequential mining technique into the probabilistic model

- Probabilistic Latent Sequential Motifs (PLSM)
- Interesting Sequence Miner (ISM)

Why ISM

- High efficiency
- Based on probabilistic model, easy to be integrated

Problems of the combination

- ISM does not consider time property
  - How to feed data into ISM?
  - How to use sequences from ISM?

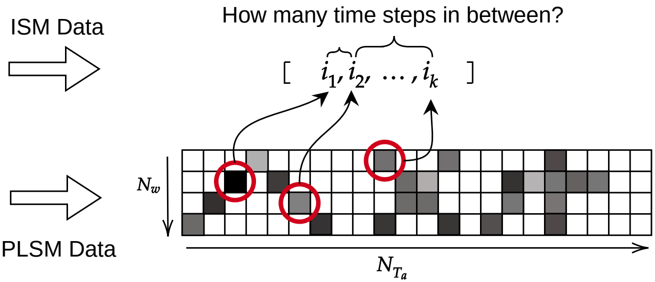
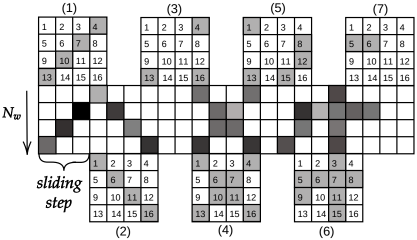


Figure 5: Problems of the combination

How to generate the sequence

- A sliding window(Same size with motifs); Each square are assigned a number $1 \sim N_w \times N_{T_r}$
- A sliding step



$\gg$

(1) [4,7,10,13]

(2) [1,6,11,16]

(3) [4,13,16]

(4) [1,4,6,7,10,11,13,16]

(5) [1,8,12,13,15]

(6) [3,5,6,7,8,9,11,15]

(7) [5,6]

Figure 6: Sequences generation

How to use sequences from ISM

Problem

- Too many sequence candidates $\Longleftarrow$ **Elimination**

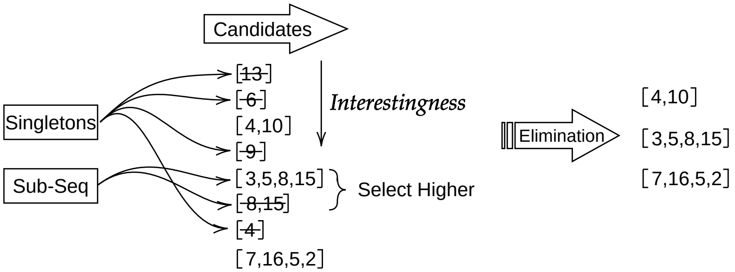


Figure 7: Eliminate sequence candidates from ISM

How to use sequences from ISM

Problem

- Unknown number of latent motifs $\Longleftarrow$ **Combination**

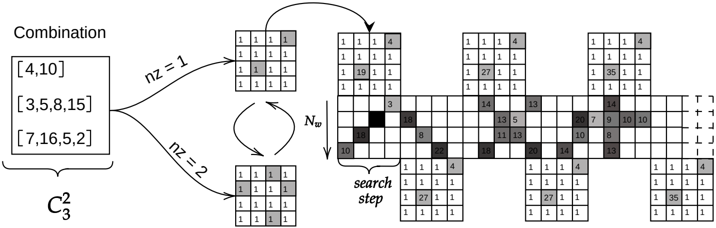


Figure 8: Initialization Process

Dir-Multi Conjugation

Why use data to update weights

- $Dir(\vec{p}) + Multi(\vec{m}|\vec{\theta}) = Dir(\vec{p} + \vec{m})$

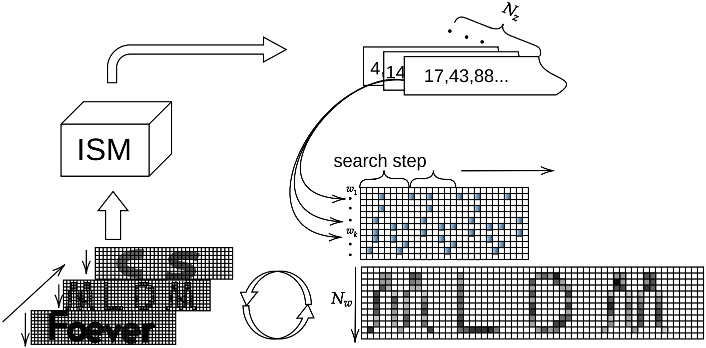


Figure 9: Combination method

Competition

Using synthetic data, we validate the performance of the combined use of PLSM and ISM.

- Compare with the Non-informative Prior

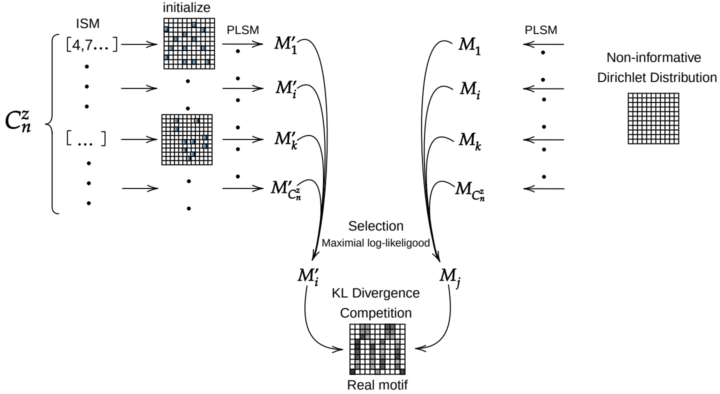


Figure 10: comparison with the Non-informative prior

Experiments Results

We made the competition on data with three overlapping levels.

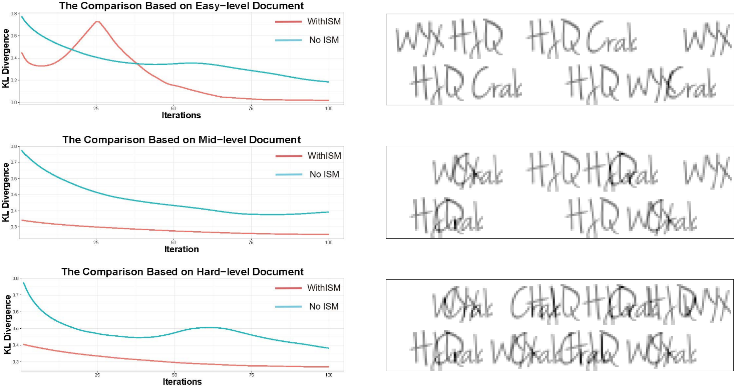


Figure 11: Competition results

Drawbacks

- The growth rate of the size of the full combination set
- The limitation length of the single sequence ISM can handle $\Longleftarrow$ *Around* 10^4 *items*

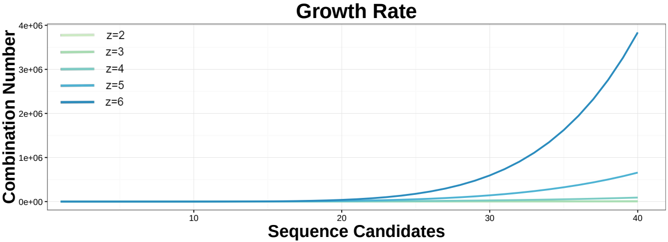


Figure 12: Growth Rate

Future Perspective

- Heuristic algorithm on sequence selection
- Select sequence based on log-likelihood
- Optimization in PLSM

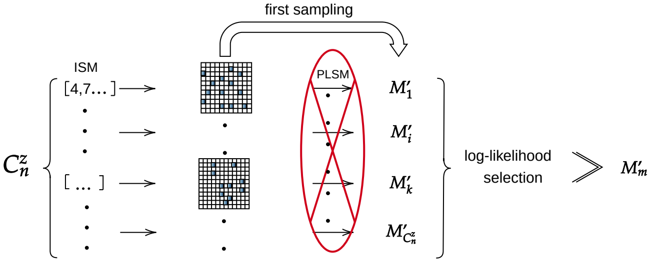


Figure 13: Improvement on sequence selection

Thank You for the listening