Creating Probabilistic Databases from Information Extraction Models

Rahul Gupta, Sunita Sarawagi

Presented by Guozhang Wang

DB Lunch, April 13^{rd}, 2009

Several slides are from the authors

Outline

- Problem background and challenges
- Proposed Solutions
  - Segmentation-per-row model
  - One-row model
  - Multi-row model
- Experiments and conclusion

Extracting and Managing Structured Web Data

- Information Extraction (using CRF, etc):
  - Text Segmentation (McCallum, UMASS)
  - Table Extraction (Cafarella, UW)
  - Preference Collection (Wortman, UPenn)
- Uncertainty Management:
  - RDBMS
  - Prob. RDBMS

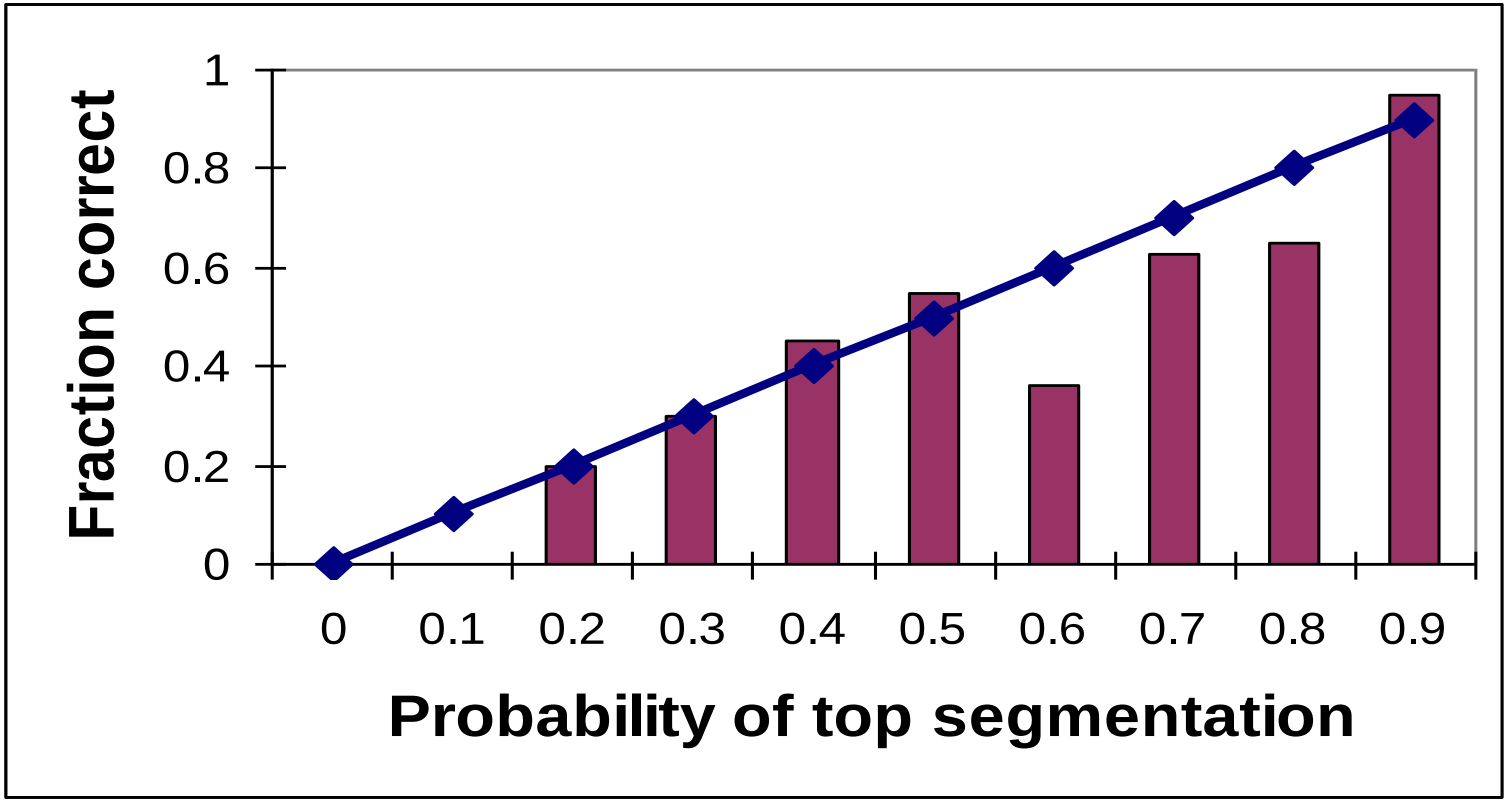
Challenges in Presenting Data

52-A Goregaon West Mumbai 400 062

House_no	Area	City	Pincode	Probability
52	Goregaon West	Mumbai	400 062	0.1
52-A	Goregaon	West Mumbai	400 062	0.2
52-A	Goregaon West	Mumbai	400 062	0.5
52	Goregaon	West Mumbai	400 062	0.2

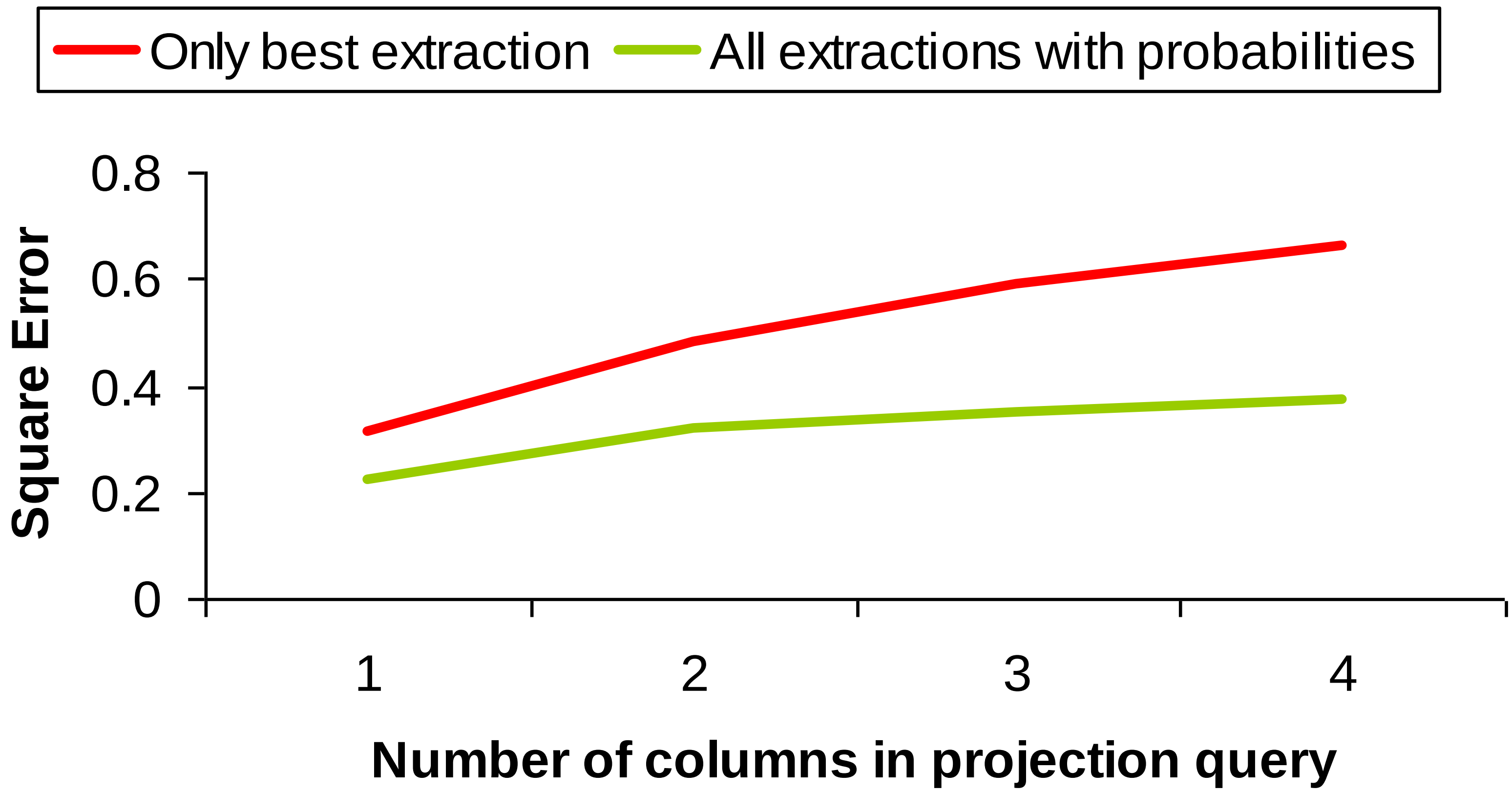
- Segmentation-per-row model
- Storage efficiency v.s. query accuracy
  - Top-1 v.s. all segmentation for each string

Confidence = Probability of Correctness



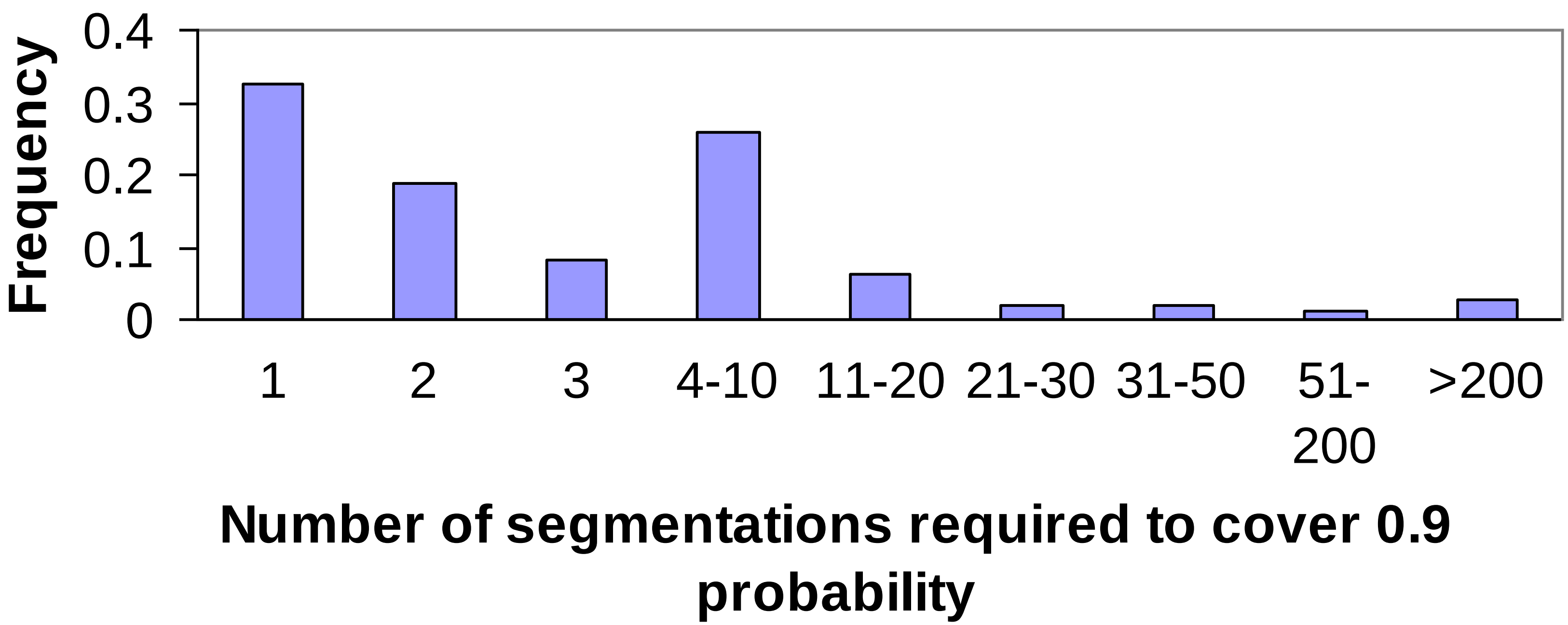
Trade-off Between Accuracy and Efficiency I

- Query Accuracy



Trade-off Between Accuracy and Efficiency II

- Storage Efficiency



Goal of This Paper

- Design data models to achieve good trade-offs between storage efficiency and query accuracy
  - To achieve query accuracy
    - Approximate the extracted segmentation distribution as similar as possible
    - Similarity metric: KL-Divergence

$$KL(P||Q) = \sum_s P(s) \log (P(s)/Q(s))$$

Outline

- Problem background and challenges
- Proposed Solutions
  - Segmentation-per-row model
  - One-row model
  - Multi-row model
- Experiments and conclusion

Proposed Data Models

- Segmentation-per-row model (Exact)
- One-row model (Column Independence)
- Multi-row model (Mixture of the two)

Segmentation-per-row Model

HNO	AREA	CITY	PINCODE	PROB
52	Bandra West	Bombay	400 062	0.1
52-A	Bandra	West Bombay	400 062	0.2
52-A	Bandra West	Bombay	400 062	0.5
52	Bandra	West Bombay	400 062	0.2

- Exact but impractical. We can have too many segmentations!

One-row Model

HNO	AREA	CITY	PINCODE
52 (0.3)	Bandra West (0.6)	Bombay (0.6)	400 062 (1.0)
52-A (0.7)	Bandra (0.4)	West Bombay (0.4)	

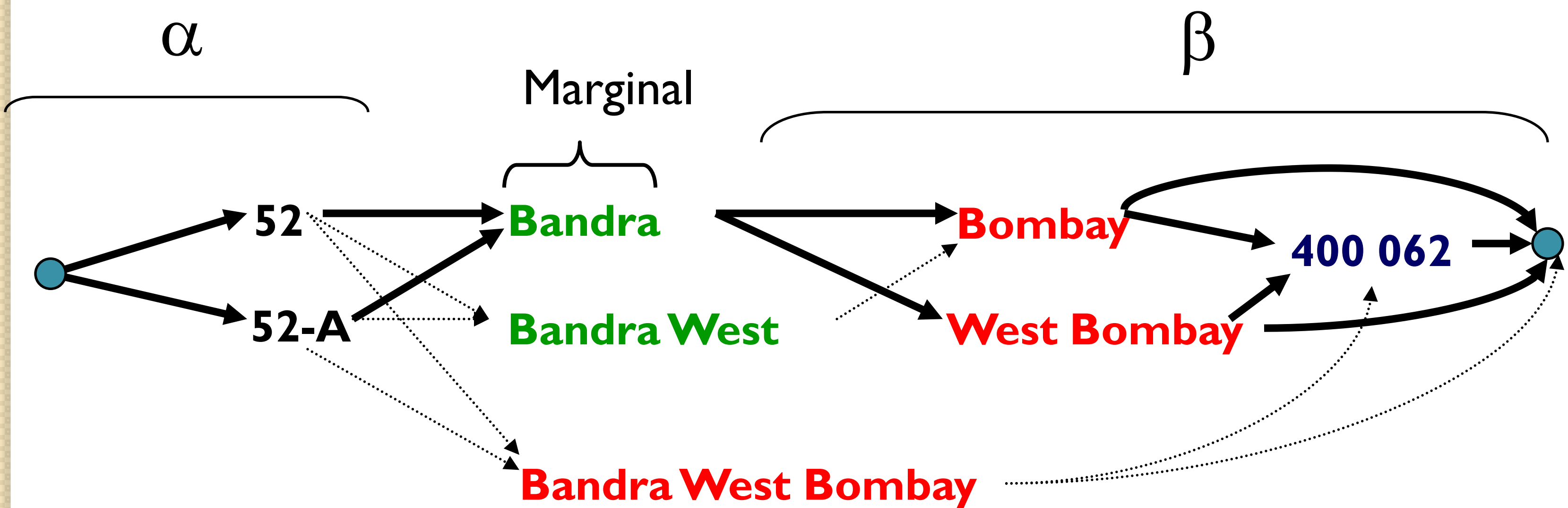
- Each column has an independent multinomial distribution "$Q_y(t,u)$"
  - E.g. P(52-A, Bandra West, Bombay, 400 062) = 0.7 x 0.6 x 0.6 x 1.0 = 0.252
- Simple model, but computed confidences are approximated (even wrong)

Populating One-row Model

$$\text{Min } KL(P||Q) = \text{Min } KL(P|| \textstyle\prod_y Q_y)$$
$$= \text{Min } \textstyle\sum_y KL(P_y||Q_y)$$

- Has a closed form solution $Q_y(t,u) = P(t,u,y)$ where $P(t,u,y)$ is marginal dist'n.
- Marginal $P(t,u,y)$ can be computed using forward-backward message passing algorithm:

Forward-Backward Algorithm



- $P(t,u,y) = c\beta_u(y)\sum_{y'}\alpha_{t-1}(y')\,\mathrm{Score}(t,u,y,y')$

Multi-row Model

HNO	AREA	CITY	PINCODE	Prob
52 (0.167) 52-A (0.833)	Bandra West (1.0)	Bombay (1.0)	400 062 (1.0)	0.6
52 (0.5) 52-A (0.5)	Bandra (1.0)	West Bombay (1.0)	400 062 (1.0)	0.4

- Rows with same ID are mutually exclusive with row probability "π_k"
- Columns in same row are independent
  - E.g. P(52-A, Bandra West, Bombay, 400 062) = 0.833 x 1.0 x 1.0 x 1.0 x 0.6 + 0.5 x 0.0 x 0.0 x 1.0 x 0.4 = 0.50

Populating Multi-row Model (fix k)

Min KL(P||Q) = Max $\sum_s$ KL(P_s|| $\sum_k \pi_k {Q^k}_s$)

- We cannot obtain the optimal parameter values in closed form because of the summation within the log
- However, we can reduce this to a well-known mixture model parameter estimation problem, and solve it using EM algorithm.

Enumeration-based EM Approach

- Initially guess the parameter values π_k and $Q^k_y(t,u)$
- E Step: soft assign each segmentation s_d to segmentation k
- M Step: update the parameters with ML values using the above soft assignment

Note the E step need to enumerate all segmentations s_d

Enumeration-less Approach

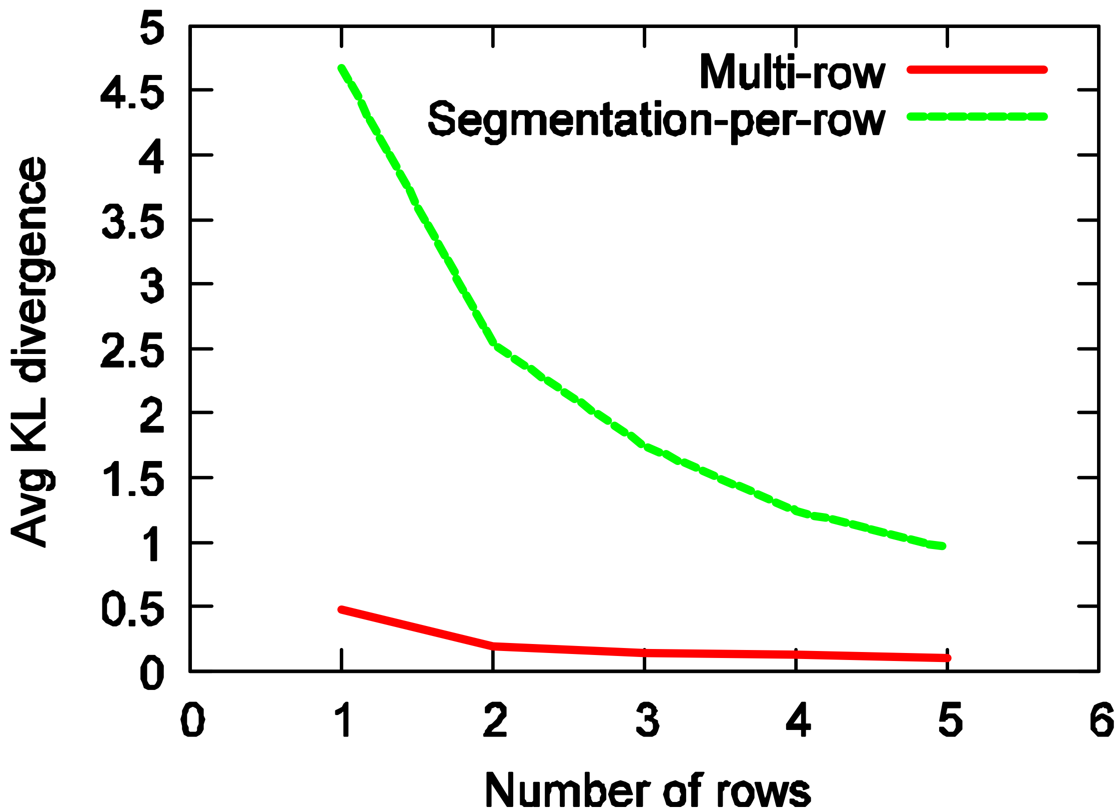
- Observation:
  - We need to enumerate segmentations at E step since we use soft assignment.
- Idea:
  - Use hard assignment instead, so that each s_d belongs to exactly one component.
    - We use a decision tree to make the hard assignment (use information gain to split node)
    - Then we can have a closed form solution to the optimization problem
    - Merge mechanism to remove the disjointness limit

Outline

- Problem background and challenges
- Proposed Solutions
  - Segmentation-per-row model
  - One-row model
  - Multi-row model
- Experiments and conclusion

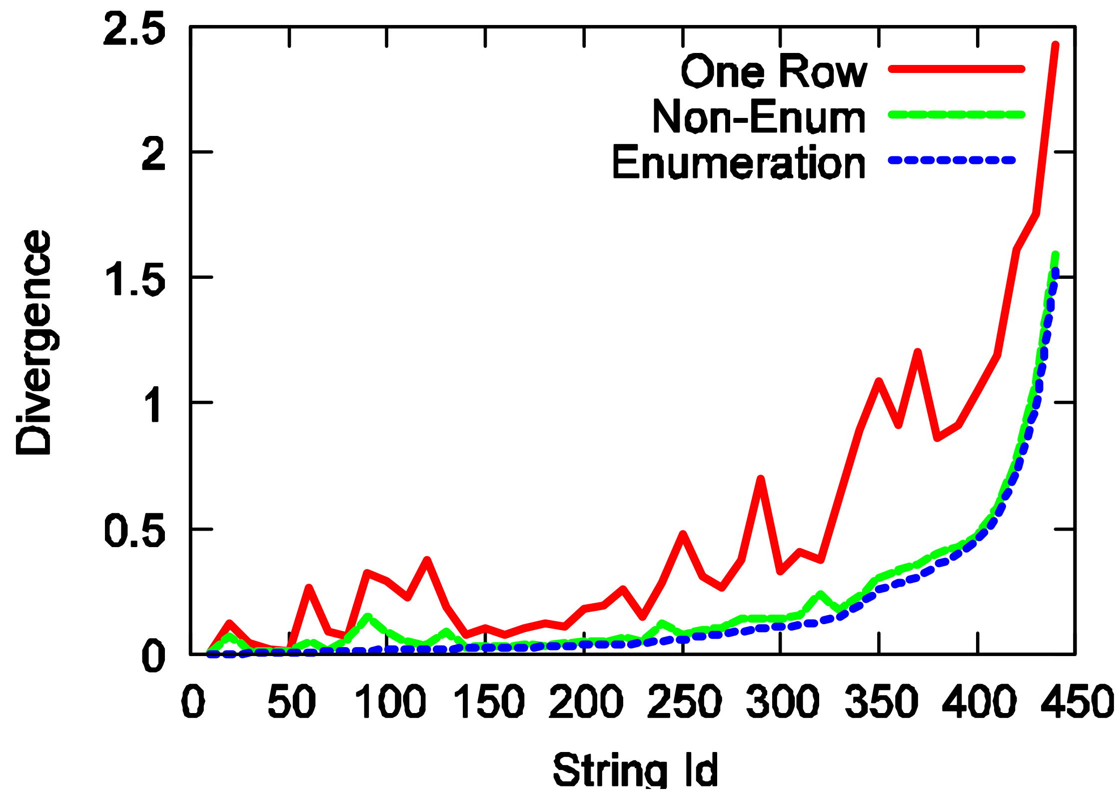
Experiment I

- Comparing multi-row with SPR



Experiment II

- Comparing multi-row with one-row



Lessons Learned ?

- Column Independence might not be suitable in some cases (8% v.s. 25%)
- Multi-row model has a good illustration of the correlations between columns

- (but) How to implement this probabilistic model?
  - One single row in Multi-row model will take more space
- Are accuracy and space efficiency equally important in this application scenario?

Questions?