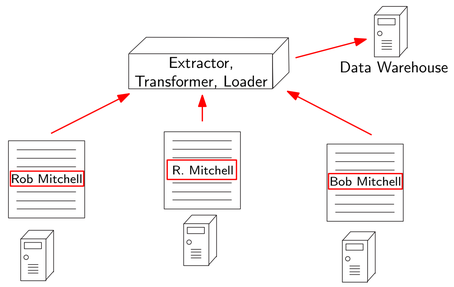
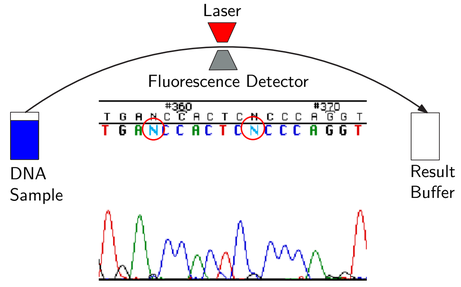
Probabilistic String Similarity Joins

Jeffrey Jestes, Feifei Li, Zhepeng Yan, Ke Yi

Data Integration



DNA Sequencing



String-Level Model

S(1) = (σ_1, p_1) = ("Bob Mitchell", 0.75)
S(2) = (σ_2, p_2) = ("Rob Mitchell", 0.25)

$S = \{(\sigma_1, p_1), (\sigma_2, p_2), \ldots, (\sigma_m, p_m)\}$
where $\sum_{i=1}^{m} p_i = 1$ and $\sigma_i \in \Sigma^*$

Character-Level Model

S = TAT $\begin{matrix} \text{A} & 0.20 \\ \text{T} & 0.80 \end{matrix}$ TCG

$S = S[1] \ldots S[n]$
$S[i] = \{(c_{i,1}, p_{i,1}), \ldots, (c_{i,\eta_i}, p_{i,\eta_i})\}$
$c_{i,j} \in \Sigma$, $p_{i,j} \in (0, 1]$ and $\sum_{j=1}^{\eta_i} p_{i,j} = 1$.

Problem Formulation

R	
id	S
1	{ (Microsoft, 0.90), (Microsoft Inc., 0.10) }
2	{ (Yahoo, 0.80), (Yahoo!, 0.20) }
⋮	⋮

$\bowtie$

T	
id	S
1	(Google, 1)
2	{ (AT&T, 0.70), (ATT, 0.30) }
⋮	⋮

A join on R and T on probabilistic string attribute S returns all pairs of records (r_i, t_j) s.t. $r_i \in R$, $t_j \in T$ and $\widehat{d}(r_i.S, t_j.S) \leq \tau$ where,

$$EED = \widehat{d}(S_1, S_2) = \sum_{s \in \Omega} w(s) \cdot d(s)$$

Possible Worlds and Expected Edit Distance

S_1	
$\sigma_{1,i}$	$p_{1,i}$
cat	0.50
kitty	0.50

S_2	
$\sigma_{2,i}$	$p_{2,i}$
dog	0.75
doggy	0.10
puppy	0.15

Ω					
$\sigma_{1,i}$	$p_{1,i}$	$\sigma_{2,j}$	$p_{2,j}$	w(s)	$d(\sigma_{1,i}, \sigma_{2,j})$
cat	0.50	dog	0.75	0.375	3
cat	0.50	doggy	0.10	0.05	5
cat	0.50	puppy	0.15	0.075	5
kitty	0.50	dog	0.75	0.375	5
kitty	0.50	doggy	0.10	0.05	4
kitty	0.50	puppy	0.15	0.075	4

$\widehat{d}(S_1, S_2) = 0.375 \cdot 3 + 0.05 \cdot 5 + 0.075 \cdot 5$
$+0.375 \cdot 5 + 0.05 \cdot 4 + 0.075 \cdot 4 = 4.125$

String-Level Probabilistic qGrams

A **string-level probabilistic** q**-gram** is a quadruple (i, p, ℓ, g) where,
i is the choice index (cid) ℓ is the start position
p is the choice probability g is the q-gram starting at ℓ

Character-Level Probabilistic qGrams

A **character-level probabilistic** q**-gram** is a pair $(\ell, S[\ell..\ell + q - 1])$ where,
ℓ is the beginning position of the q-gram
$S[\ell..\ell + q - 1]$ is the *probabilistic substring*
$S[\ell] \ldots S[\ell + q - 1]$

String-Level Relational Representation

probabilistic strings

id	S
1	{(add, 0.8), (plus, 0.2)}
2	{(up, 0.9), (op, 0.1)}

relational representation

id	cid	A	p	len
1	1	add	0.8	3.2
1	2	plus	0.2	3.2
2	1	up	0.9	2
2	2	op	0.1	2

q-grams

id	cid	ℓ	g
1	1	1	#a
1	1	2	ad
1	1	3	dd
1	1	4	d$
1	2	1	#p
1	2	2	pl
1	2	3	lu
⋮	⋮	⋮	⋮

String-Level Probabilistic qGram Lower Bound

Theorem *For any string-level probabilistic strings S_1 and S_2:*

$$\widehat{d}(S_1, S_2) \geq 1 + \frac{\max(\mathbf{E}[|S_1|], \mathbf{E}[|S_2|])}{q} - \frac{\sum_{(\rho_1, \rho_2) \in G_{S_1} \cap G_{S_2}} p(\rho_1) p(\rho_2) + 1}{q}$$

```
1 SELECT R.id AS rid, T.id AS tid
2 FROM R,T,Rq,Tq
3 WHERE Rq.g=Tq.g AND Rq.id=R.id
4 AND Tq.id=T.id AND Rq.cid=R.cid
5 AND Tq.cid=T.cid
6 GROUP BY R.id, T.id, R.len, T.len
7 HAVING 1+(max(R.len,T.len)-
8 SUM(R.p*T.p)-1)/q≤ τ
```

- We also derive SQL lower-bounds which utilize positional and length information.

Character-Level Relational Representation

probabilistic strings

id	S
1	{(A, 0.8), (C,0.2)}, {(G, 0.7), (T, 0.3)}
2	{(A, 1)}, {(G, 0.6), (T,0.4)}
3	{(C, 1)}, {(A, 1)}, {(G, 1)}

relational representation

id	A	len
1	⋆A0.8|C0.2⋆⋆G0.7|T0.3⋆	2
2	A⋆G0.6|T0.4⋆	2
3	CAG	3

q-grams

id	p	ℓ	g
1	0.80	1	#A
1	0.20	1	#C
1	0.56	2	AG
1	0.24	2	AT
1	0.14	2	CG
1	0.06	2	CT
1	0.70	3	G$
1	0.30	3	T$
⋮	⋮	⋮	⋮

Character-Level Probabilistic qGram Lower Bound

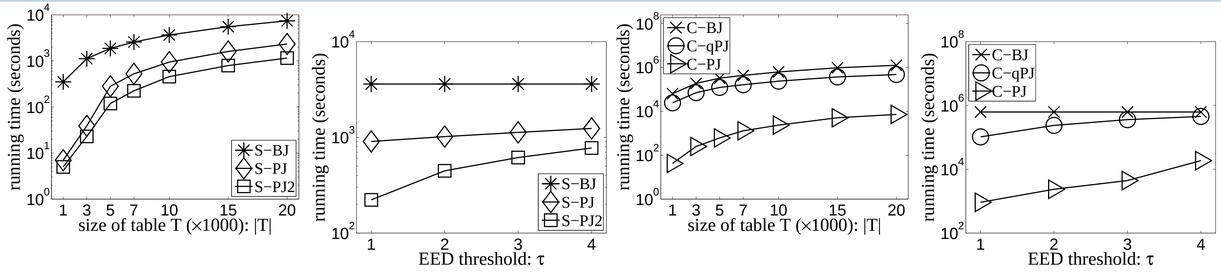
Theorem *For any character-level probabilistic strings S_1, S_2:*

$$\widehat{d}(S_1, S_2) \geq 1 + \frac{\max(|S_1|, |S_2|)}{q} - \frac{\sum_{\substack{\gamma_1 \in G_{S_1} \\ \gamma_2 \in G_{S_2}}} \Pr(\gamma_1 = \gamma_2) + 1}{q}$$

```
1 SELECT R.id AS rid,T.id AS tid
2 FROM R,T,Rq,Tq
3 WHERE Rq.g=Tq.g AND R.id=Rq.id
4 AND T.id=Tq.id
5 GROUP BY R.id, T.id, R.len, T.len
6 HAVING 1 + max(R.len,T.len)/q -
7 SUM( Tq.p*Rq.p)/q - 1/q ≤ τ
```

- We also derive SQL lower and upper-bounds which utilize positional and length information. DP-based lower and upper-bounds are also derived.

Experiments



String-Level *Author1* Character-Level *Genome*