Recognition Capacity versus Search Speed in Noisy Databases

July 5th, 2012

Ertem Tuncel

UNIVERSITY OF CALIFORNIA RIVERSIDE
The problem

• How large can a database be if we are to identify all of its entries reliably?
• How fast can the entries be identified?
• How large a storage space is needed?
• Ideally, we want to
  • identify as many entries as possible
  • do the search as fast as possible
  • use as little disk space as possible
• But these are competing qualities!!
Information-theoretic model

$$\Pr[W = w] = \frac{1}{M}$$

**ENROLLMENT**

**RETRIEVAL**

**DECISION**

Recognition Capacity versus Search Speed in Noisy Databases
Previous work

- O'Sullivan and Schmid (Allerton 2002) and Willems et al. (ISIT 2003):
  - Data enrolled as is (uncompressed and without structure)
  - All is retrieved to check against the query
  - To guarantee $\Pr[\hat{W} \neq W] \to 0$
    \[
    \frac{1}{n} \log_2 M \overset{\Delta}{=} R_i \leq I(Y; Z)
    \]
  - Not surprising since this is the capacity of the channel $P_{Z|Y}$ (except we don't get to choose $P_Y$).
Previous work

- Interpretation:

Typical $Y^n$  
Typical $Z^n$

$\approx 2^{nI(Y;Z)}$
decoding balls
Previous work

- Westover and O'Sullivan (IT 2008) and Tuncel (IT 2009):
  - Data enrolled after compression, but without structure.
  - All is retrieved to check against the query.
  - To guarantee $\Pr[\hat{W} \neq W] \rightarrow 0$,
    
    \[
    R_i \leq I(U; Z) \\
    R_c \geq I(U; Y)
    \]
    
    for some $U$ such that $U \leftrightarrow Y \leftrightarrow X \leftrightarrow Z$.
Previous work

- **Interpretation:**

  Typical $U^n$  
  Typical $Y^n$  
  Typical $Z^n$

  \[ \approx 2^n I(U; Y) \]  
  encoding balls

  \[ \approx 2^n I(U; Z) \]  
  decoding balls

Recognition Capacity versus Search Speed in Noisy Databases
Previous work

Recognition Capacity versus Search Speed in Noisy Databases

$I(Y; Z)$

Achievable region

$R_i$ vs $R_c$

$H(Y)$
Few other extensions

• Gündüz, Tuncel, Goldsmith, and Poor (ISIT 2009)
  - Identification over multiple databases.

• Tuncel and Gündüz (ISIT 2010)
  - Identification and lossy reconstruction in noisy databases.

• Dasarathy and Draper (ISIT 2011)
  - Error exponents versus identification and compression rates
Capacity vs search speed

- Willems (ISIT 2009)
  - A clustering-based approach
  - Query space is quantized and for each quantized query, a single cluster is searched.
  - Data is put into clusters by (indirectly) quantizing using the same code.
Typical $Y^n$ vs Typical $Z^n$ vs Typical $U^n$

$\approx 2^n [R_i - I(U;Y)]$ entries per cluster

$\approx 2^n I(U;Z)$ clusters

Recognition Capacity versus Search Speed in Noisy Databases
Summary of this work

- Different model.
  - Instead of quantizing the query and accessing a single cluster, quantize the data space and access many clusters.
  - On average, no redundancy in storage.
  - More suitable for an extension including compression.

- We especially investigate the minimum search rate subject to no loss in identification capacity.
The proposed system

$Y^n(m)$

$f_n(Y^n(m)) \subset \mathcal{K}$

$Y^n(m)$

$\cdots$

$Y^n(m)$

$C_1$

$C_2$

$D(Z^n)$

$h_n$

$\hat{W}$

Recognition Capacity versus Search Speed in Noisy Databases
Achievable rates

- If there are $K$ clusters, $M$ entries, and on average

$$L = E \left( \sum_{k \in g_n(Z^n)} |C_k| \right)$$

entries are retrieved, together with

$$\Pr[\hat{W} \neq W] \leq \epsilon$$

we say that there exists an $(M, K, L, n, \epsilon)$ data management system.
Achievable rates

• Then a rate pair \((R_i, R_s)\) is achievable if there exists an \((M, K, L, n, \epsilon)\) system such that

\[
\frac{1}{n} \log M \geq R_i - \epsilon
\]

\[
\max \left\{ \frac{1}{n} \log K, \frac{1}{n} \log L \right\} \leq R_s + \epsilon
\]

• Assumption: Retrieval cost = \(\alpha K + \beta L\)
Main result

• **Theorem**: \((R_i, R_s)\) is achievable if there exists \(T\) such that

\[ T \leftrightarrow Y \leftrightarrow X \leftrightarrow Z \]

and

\[ R_i \leq I(Y; Z) \]
\[ R_s \geq I(Y; T) \]
\[ R_i - R_s \leq I(Z; T) \]
Sketch of proof

• Randomly create $2^{n(R_s+\epsilon)} \geq K \geq 2^{nI(Y;T)}$ cluster representatives $T^n(k)$ i.i.d. $\sim P_T$

• **Enrollment:** For any $Y^n(m)$, find all $k$ such that

$$(Y^n(m), T^n(k)) \in S^n_{[Y,T]}$$

• For each $k$ where the pair is typical, include $Y^n(m)$ in $C_k$ with probability

$$q_n \approx \frac{2^{nI(Y;T)}}{K}$$
Sketch of proof

- **Cluster selection**: Given $Z^n$, retrieve all clusters for which
  \[ (Z^n, T^n(k)) \in S^n_{[ZT]} \]

- **Identification**: Among all the retrieved clusters, find the unique $m$ such that
  \[ (Z^n, Y^n(m)) \in S^n_{[ZY]} \]

- **Probability of error**: The true data will not be missed. To prevent other data to also match, $2^n(R_i - \epsilon) \leq M \leq 2^nI(Y;Z)$ suffices.
Sketch of proof

• Expected number of retrieved data vectors:
  - On average, each cluster has \( \approx \frac{2^n R_i}{K} \) entries.
  - On average, \( \approx K 2^{-nI(Z;T)} \) clusters are retrieved.
  - Since \( R_i - I(Z;T) \leq R_s \), this implies
    \[ \frac{1}{n} \log L \leq R_s + \epsilon \]
Analysis

• Compare

\[ R_i \leq I(Y; Z) \]
\[ R_s \geq I(Y; T) \]
\[ R_i - R_s \leq I(Z; T) \]

with the storage/capacity tradeoff

\[ R_c \geq I(Y; U) \]
\[ R_i \leq I(Z; U) \]
Analysis

Recognition Capacity versus Search Speed in Noisy Databases

$R_i - R_s$

$I(Y; Z)$

Achievable $(R_i - R_s, R_s)$
Analysis

Recognition Capacity versus Search Speed in Noisy Databases

$R_i$ vs $R_s$

$I(Y; Z)$

Achievable $(R_i, R_s)$
Analysis

• Example: $P_X = \{\frac{1}{2}, \frac{1}{2}\}$ and $Z = Y = X$.

Recognition Capacity versus Search Speed in Noisy Databases
Conclusions and future work

• Proposed a clustering scheme and analyzed its performance.

• Unlike in storage/capacity tradeoff, it is possible to reduce search complexity without compromising maximum identification rate

• **Future work:** Include a second layer performing compression. Clustering will also help as a first layer compressor.