

Information-theoretical analysis of index searching

Akisato Kimura^{*1,*2}, Tomohiko Uyematsu^{*2}

^{*1} Media Information Laboratory,
NTT Communication Science Laboratories,
NTT Corporation

^{*2} Department of Communications and Integrated Systems,
Graduate School of Science and Engineering,
Tokyo Institute of Technology

Abstract

- Present an information-theoretical viewpoint for similarity-based retrieval with indexes
 - This type of retrieval is formulated as a certain kind of multi-terminal source coding problem
- Clarify the optimal retrieval performance and certain relationships between retrieval parameters

Remark . *We presented the same topic at SITA2005. However, we have since found several flaws in the models and results provided in the previous report.*

Contents

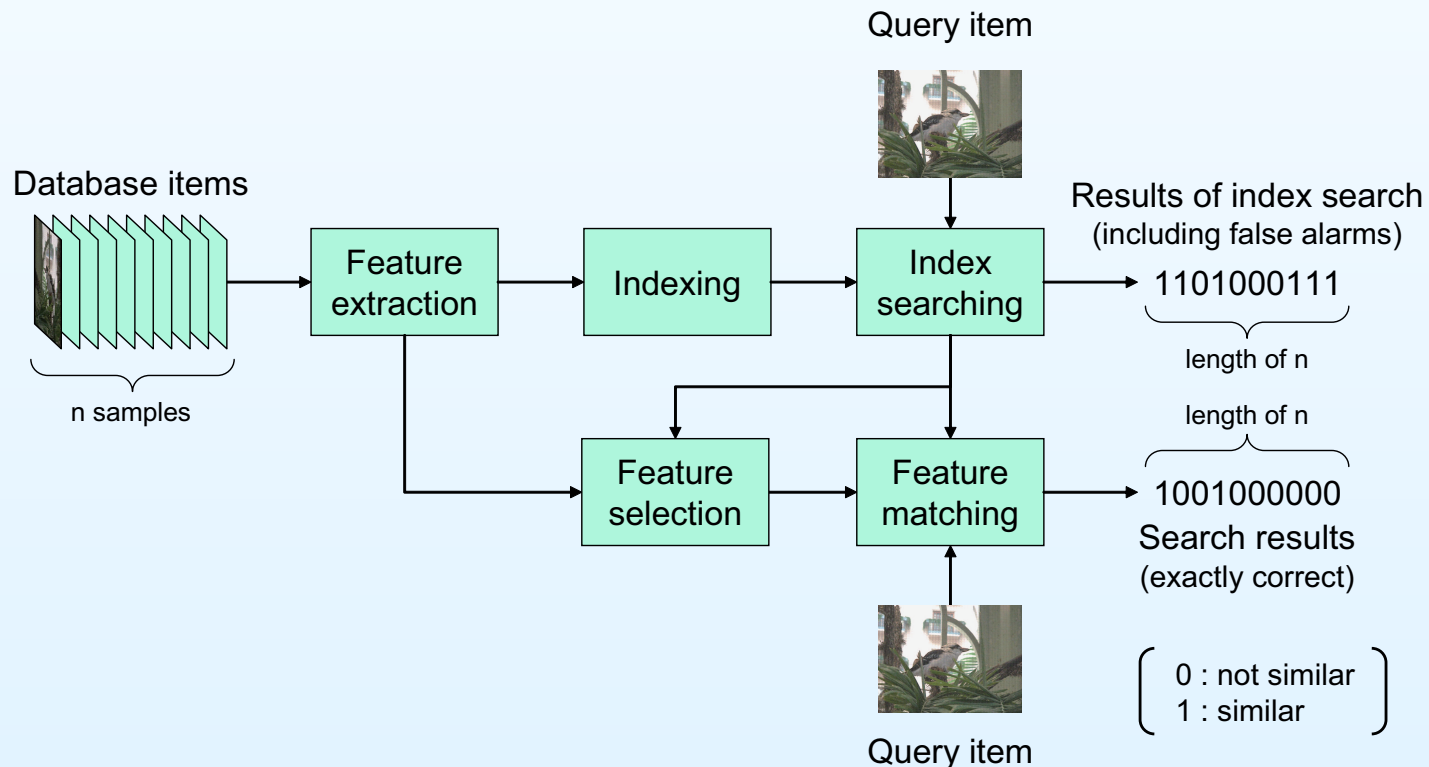
- Introduction
 - Background
 - Association with implementation
 - Related work
- Problem formulation
- Main result
 - Coding theorem
 - Interpretation of the coding theorem
- Conclusion

Background

- Possibility to capture a huge number of music clips, images and movies easily.
 - Spread of broadband network and capturing devices
 - Large amounts of low-cost storage
- **Media information retrieval** must be developed that extracts desired information through the multimedia archives quickly and accurately.
- Research issues on media information retrieval
 - **Huge amounts of information to be retrieved**
→ search speed
 - Signal degradation caused by noise and distortion
→ robustness
 - Signal fluctuation caused by viewpoint changes and music arrangement → signal models

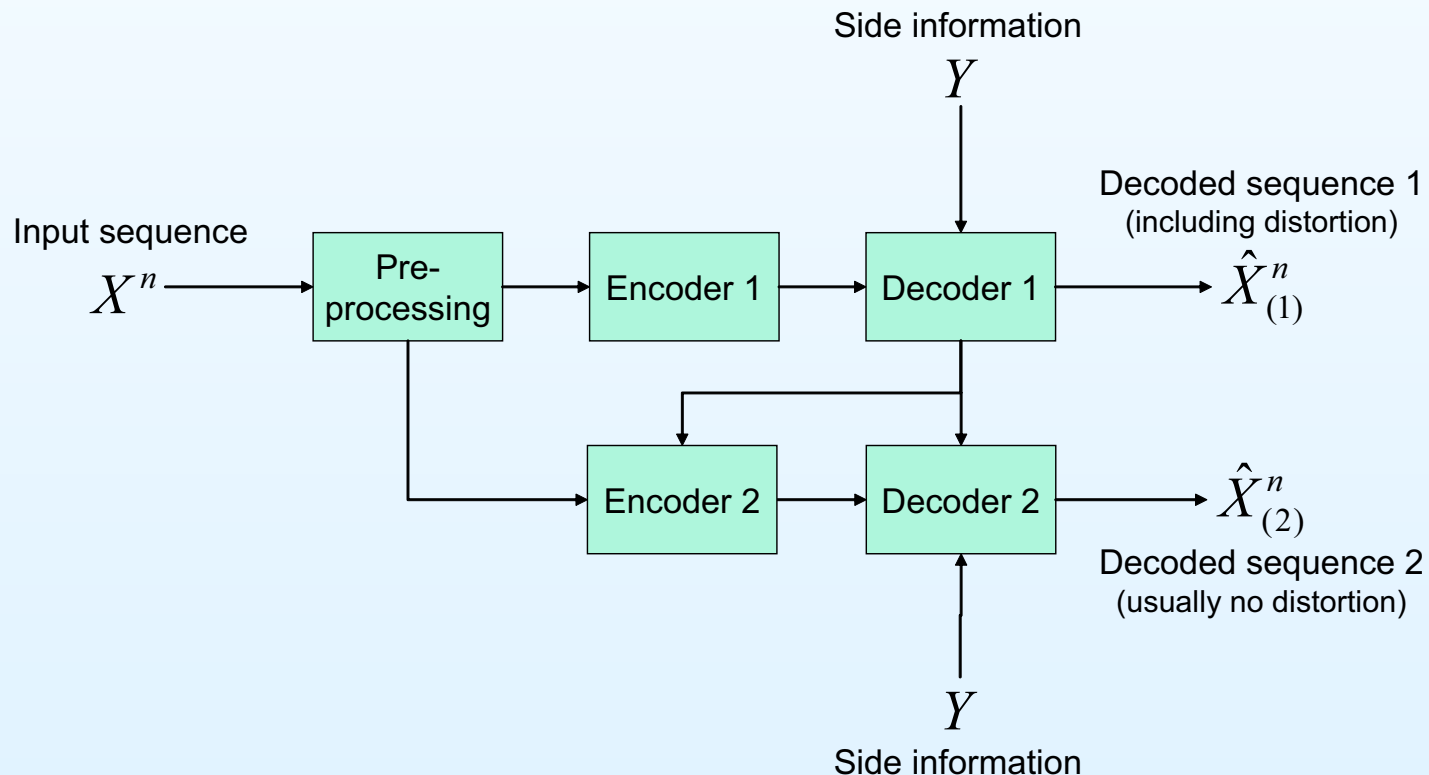
Index searching

- Commonly used as a core technique to accelerate retrieval
 - Ex. Hashing, tree structures (B-tree, R-tree [BK90])
 - Associate data items with indexes that represent the items concisely
 - First check indexes to eliminate irrelevant data items



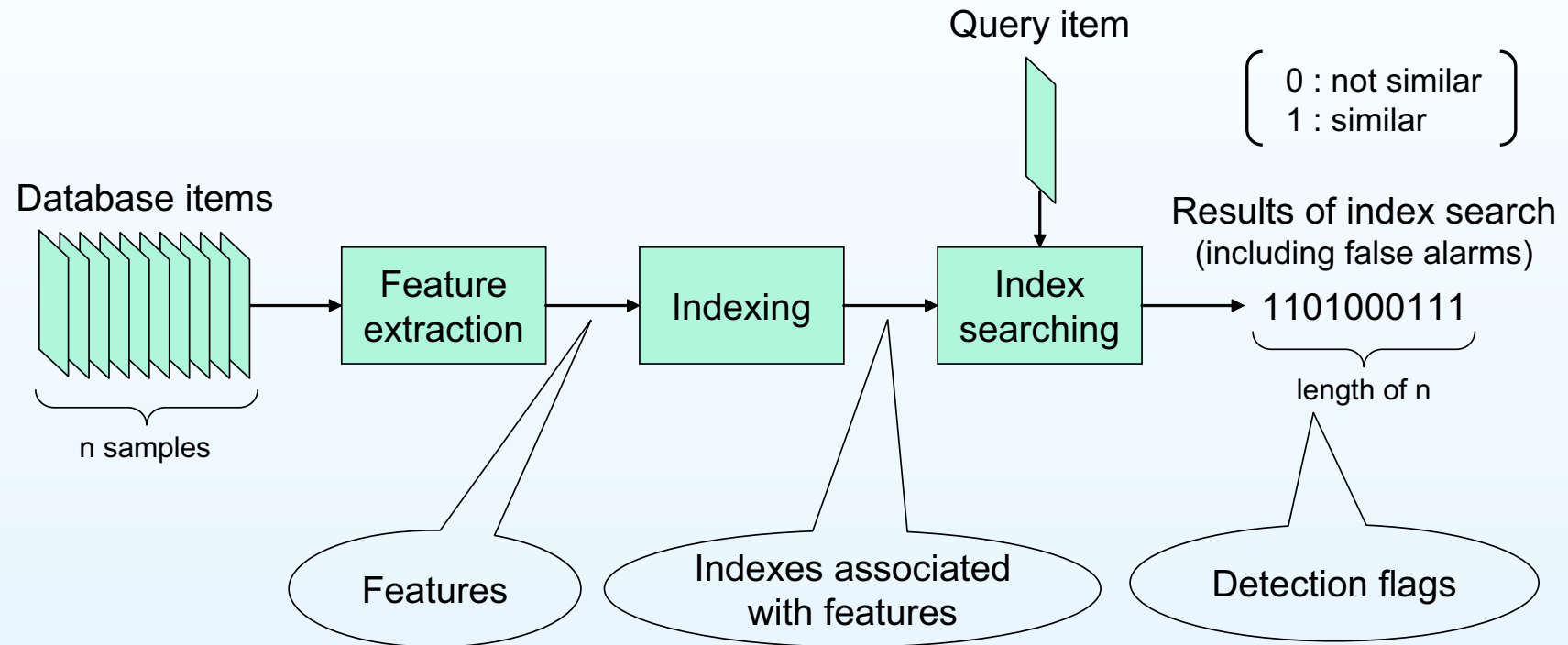
Main contributions

- Similarity-based retrieval with indexes can be formulated as a certain kind of multi-terminal source coding problem
- The achievable rate-distortion region implies
 - the optimal performance of index searching
 - relationships between index size and search time



Association with implementation (1/3)

The first stage



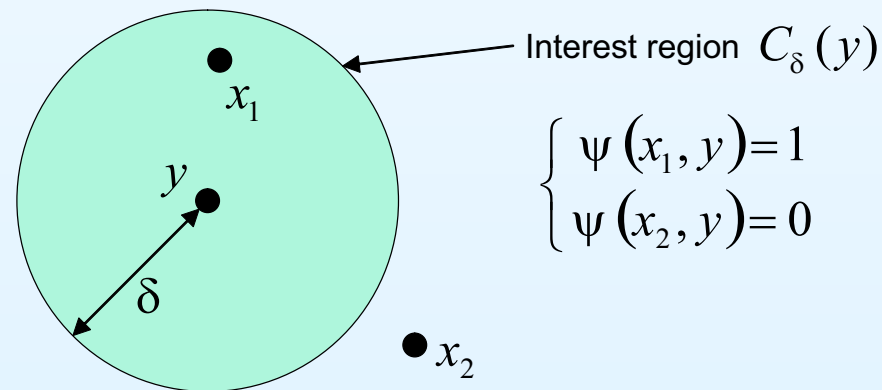
Association with implementation (2/3)

Detection flags

$$z_i \stackrel{\text{def.}}{=} \psi(x_i, y) \stackrel{\text{def.}}{=} \begin{cases} 1 & x_i \in C_\delta(y), \\ 0 & \text{Otherwise,} \end{cases} \quad (\text{actual})$$

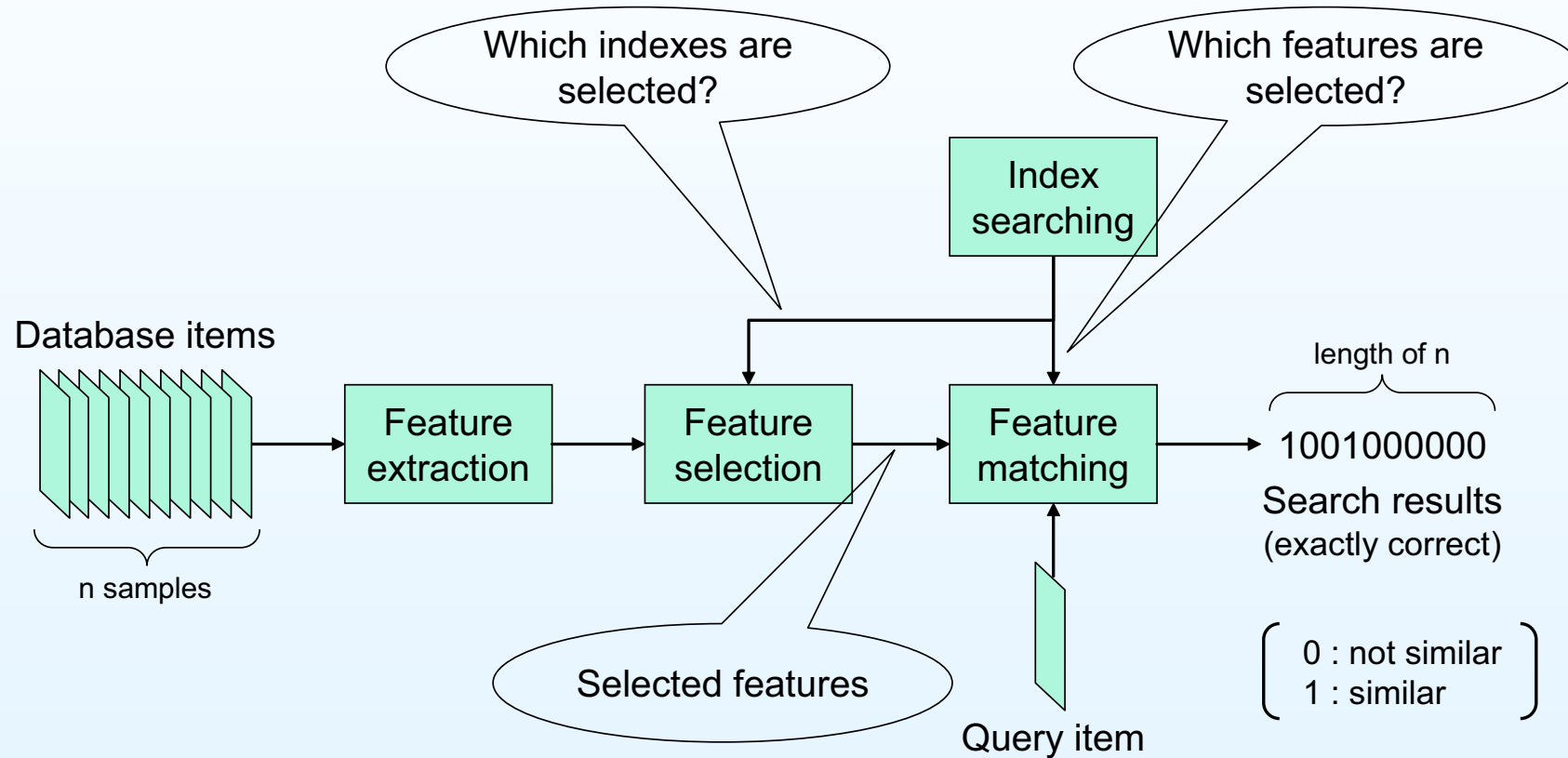
$$\hat{z}_{(j)i} \stackrel{\text{def.}}{=} \psi(\hat{x}_{(j)i}, y), \quad \forall i \in \mathcal{I}_n, j = 1, 2 \quad (\text{estimated})$$

$$\Delta(x_i, \hat{x}_{(j)i}; y) \stackrel{\text{def.}}{=} h(z_i, \hat{z}_{(j)i}) \quad (\text{Hamming distance})$$



Association with implementation (3/3)

The second stage



Related work: analysis of retrieval performance

- Geometrical approach
 - A lot of studies for nearest neighbor search (e.g. Friedman et al. [FBF77], Berchtold et al. [BBK⁺97])
 - Item distributions are supposed to be uniform.
 - Make it possible to consider volumes as “probabilities”
 - Difficult to evaluate the performance for other types of distributions
- **Information-theoretical approach** (Tuncel et al. [TKR04])
 - Formulate approximate similarity search as a kind of multi-terminal source coding problem (Wyner-Ziv coding + successive refinement)
 - Clarify some relationships between search time, search threshold, and amount of storage, based on rate-distortion theory

Preliminaries

- $\mathcal{X}, \mathcal{Y}, \hat{\mathcal{X}}$: finite alphabets, \mathcal{B} : binary alphabet, \mathcal{R} : set with real values.
- $\mathcal{X}^* = \cup_{n \geq 0} \mathcal{X}^n$: a set of all sequences with finite length over \mathcal{X} (includes null string)
- $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ ($X_i \in \mathcal{X}$): source to be encoded (i.i.d.)
- $Y \in \mathcal{Y}$: side information available only at decoder
- $\hat{\mathbf{X}}_{(j)} = \{\hat{X}_{(j)i}\}_{i=1}^{\infty}$ ($\hat{X}_{(j)i} \in \hat{\mathcal{X}}$) ($j = 1, 2$): outputs obtained from j -th decoder
- $\Delta : \mathcal{X} \times \hat{\mathcal{X}} \times \mathcal{Y} \rightarrow \mathcal{R}$: distortion function that depends on side information
- $\Delta^n : \mathcal{X}^n \times \hat{\mathcal{X}}^n \times \mathcal{Y} \rightarrow \mathcal{R}$: $\Delta^n(\mathbf{x}, \hat{\mathbf{x}}; y) = \frac{1}{n} \sum_{i=1}^n \Delta(x_i, \hat{x}_i; y)$

Problem formulation

Definition 1. (Index Searching (IS) code)

A set $(\varphi_{(01)}^n, \varphi_{(02)}^n, \varphi_{(1)}^n, \varphi_{(2)}^n, \widehat{\varphi}_{(1)}^n, \widehat{\varphi}_{(2)}^n)$ of encoders and decoders is an IS code for the source X and the side information Y if and only if

$$\varphi_{(1)}^n : \mathcal{X}^n \rightarrow \mathcal{B}^*, \quad \varphi_{(01)}^n : \mathcal{A}_n^{(1)} \times \mathcal{Y} \rightarrow \mathcal{B}^*,$$

$$\varphi_{(02)}^n : \mathcal{A}_n^{(1)} \times \mathcal{Y} \rightarrow \mathcal{B}^*, \quad \varphi_{(2)}^n : \mathcal{A}_n^{(01)} \times \mathcal{X}^n \rightarrow \mathcal{B}^*,$$

$$\widehat{\varphi}_{(1)}^n : \mathcal{A}_n^{(1)} \times \mathcal{Y} \rightarrow \widehat{\mathcal{X}}^n,$$

$$\widehat{\varphi}_{(2)}^n : \mathcal{A}_n^{(02)} \times \mathcal{A}_n^{(2)} \times \mathcal{Y} \rightarrow \widehat{\mathcal{X}}^n,$$

and images of encoders and decoders are all prefix sets, where

$$\mathcal{A}_n^{(1)} = \varphi_{(1)}^n(\mathcal{X}^n), \quad \mathcal{A}_n^{(01)} = \varphi_{(01)}^n(\mathcal{A}_n^{(1)}, \mathcal{Y}),$$

$$\mathcal{A}_n^{(02)} = \varphi_{(02)}^n(\mathcal{A}_n^{(1)}, \mathcal{Y}), \quad \mathcal{A}_n^{(2)} = \varphi_{(2)}^n(\mathcal{A}_n^{(01)}, \mathcal{X}^n).$$

Achievability

Definition 2. (IS-achievable rate quadruplet)

$(R_{01}, R_{02}, R_1, R_2)$ is an IS-achievable rate quadruplet of the source X and side information Y for a given distortion pair (D_1, D_2) if and only if there exists a sequence of IS codes $\{(\varphi_{(01)}^n, \varphi_{(02)}^n, \varphi_{(1)}^n, \varphi_{(2)}^n, \hat{\varphi}_{(1)}^n, \hat{\varphi}_{(2)}^n)\}_{n=1}^{\infty}$ for X and Y such that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} E \left[l(A_n^{(i)}) \right] \leq R_i, \quad (i = 01, 02, 1, 2)$$

$$\limsup_{n \rightarrow \infty} E \left[\Delta^n(X^n, \hat{X}_{(i)}^n; Y) \right] \leq D_j, \quad (j = 1, 2)$$

where

$$A_n^{(1)} = \varphi_{(1)}^n(X^n), \quad A_n^{(01)} = \varphi_{(01)}^n(A_n^{(1)}, Y),$$

$$A_n^{(02)} = \varphi_{(02)}^n(A_n^{(1)}, Y), \quad A_n^{(2)} = \varphi_{(2)}^n(A_n^{(01)}, X^n).$$

Achievable rate-distortion region

Definition 3. (IS-achievable rate region)

$$\mathcal{R}_{IS}(X, Y | D_1, D_2) = \{(R_{01}, R_{02}, R_1, R_2) : \\ (R_{01}, R_{02}, R_1, R_2) \text{ is an IS-achievable rate quadruplet of } (X, Y) \\ \text{for } (D_1, D_2)\}.$$

Main Theorem (1/2)

Theorem 1.

$$\mathcal{R}_{IS}(X, Y | D_1, D_2) \subseteq \{(R_{01}, R_{02}, R_1, R_2) : \text{ (outer bound)}$$

$$R_1 \geq I(X; UV), R_{01} \geq 0, R_{02} \geq I(X; V), R_2 \geq I(X; W|V)\}$$

where random variables $U \in \mathcal{U}$, $V \in \mathcal{V}$ and $W \in \mathcal{W}$ are selected s.t.

- The alphabet sizes are bounded as

$$|\mathcal{U}| \leq |\mathcal{X}| + 1, |\mathcal{V}| \leq |\mathcal{U} \times \mathcal{X} \times \mathcal{Y}| + 4, |\mathcal{W}| \leq |\mathcal{U} \times \mathcal{V} \times \mathcal{X}| + 1,$$

- The Markov chain $U \rightarrow X \rightarrow Y$ is satisfied,
- There exist functions $\phi_{(1)}$ and $\phi_{(2)}$, which satisfy

$$D_1 \geq E [\Delta (X, \phi_{(1)}(U, Y); Y)],$$

$$D_2 \geq E [\Delta (X, \phi_{(2)}(W, V, Y); Y)].$$

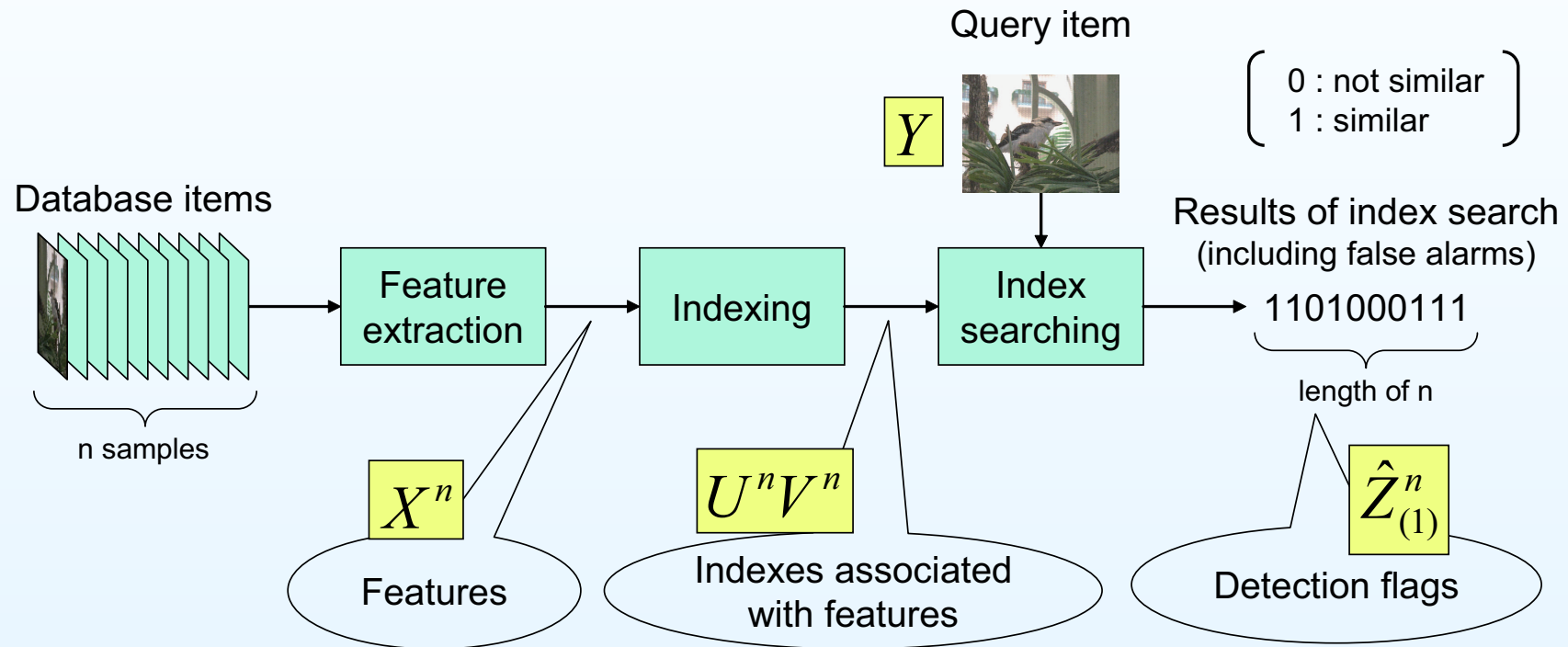
Main Theorem (2/2)

Theorem 1. *(Contd.) An inner bound is obtained in the same functional forms, while the Markov chain is replaced as*

$$\begin{aligned}UV &\rightarrow X \rightarrow Y, \\W &\rightarrow VXY \rightarrow U.\end{aligned}$$

Interpretation of the main theorem (1/2)

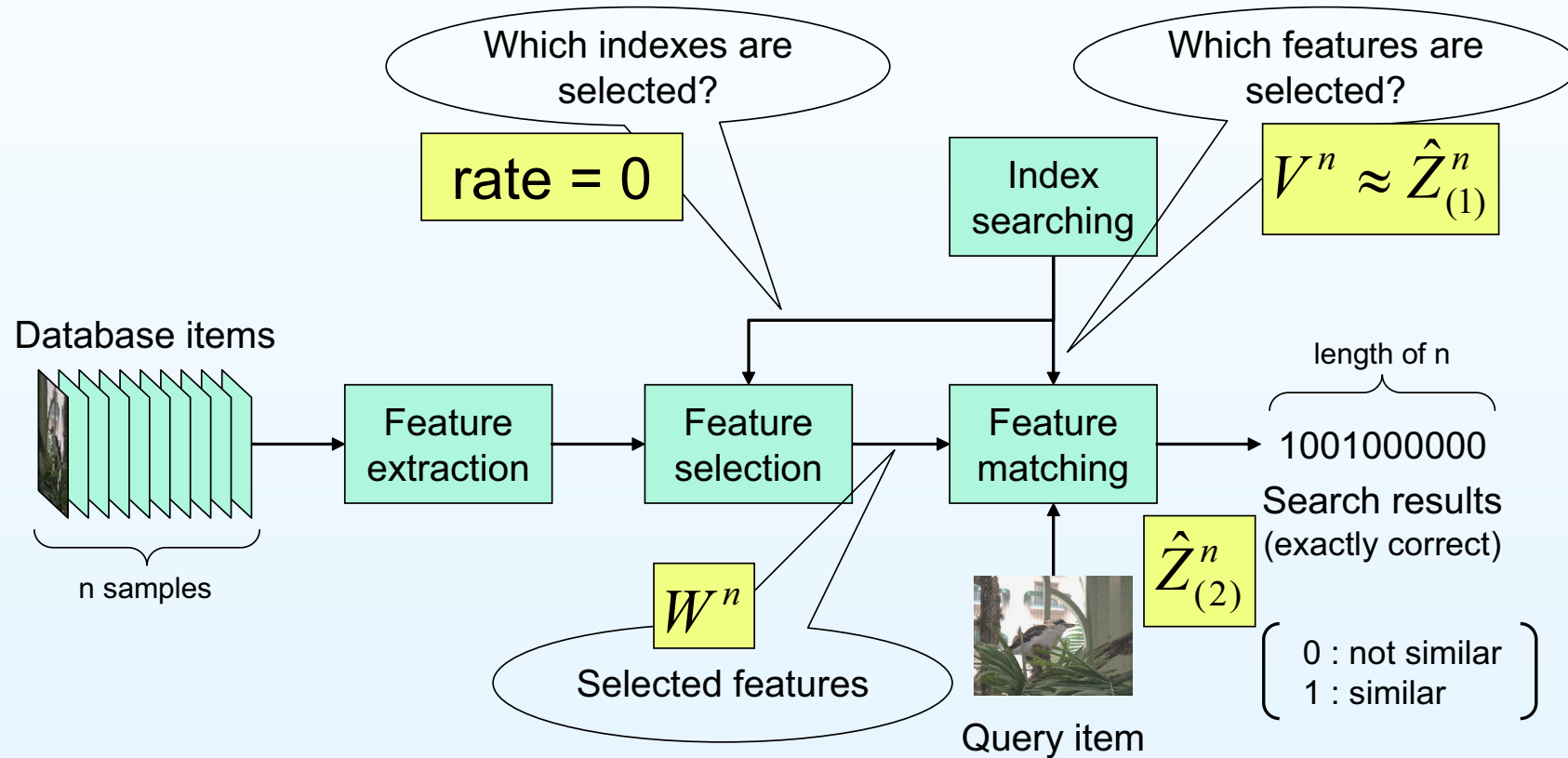
The first stage



- Outer bound: $U \rightarrow X \rightarrow Y$
- Inner bound: $UV \rightarrow X \rightarrow Y$

Interpretation of the main theorem (2/2)

The second stage



Conclusions

- Presented an information-theoretical model of similarity-based retrieval with indexes
- Clarified the optimal performance of the retrieval and some relationships between retrieval parameters from an information-theoretical viewpoint

Future work

- Extend the results to other classes of information sources
- Evaluate existing indexing methods from an information-theoretic point of view
- Model and evaluate other types of retrieval

Thank you

References

- [BBK⁺97] S. Berchtold, C. Boehm, D. Keim, F. Frebs, and H. P. Kriegel. A cost model for nearest neighbor search in high dimensional data spaces. In *Proc. ACM SIGMOD International Conference on Management of Data (ACM SIGMOD)*, pages 78–86, May 1997.
- [BK90] N. Beckman and H. P. Kriegel. The R*-tree : an efficient and robust access method for points and rectangles. In *Proc. ACM SIGMOD International Conference on Management of Data (ACM SIGMOD)*, pages 322–331, June 1990.
- [FBF77] J. Friedman, J. Bentley, and R. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software*, 3(3):209–226, September 1977.
- [TKR04] E. Tuncel, P. Koulgi, and K. Rose. Rate-distortion approach to databases: storage and content-based retrieval. *IEEE Trans. Inform. Theory*, 50(6):953–967, June 2004.

Some materials will be available at <http://www.brl.ntt.co.jp/people/akisato/>