

Information-theoretical analysis of index searching: Revised

Akisato Kimura

Tomohiko Uyematsu *

Abstract— We present an information-theoretical viewpoint for similarity-based retrieval along with index structures. This retrieval system has two stages: pruning data items based on the index structures, and matching surviving data items. The first stage is modeled as the Wyner-Ziv problem, while the second stage is considered as a coding problem where some of the decoding results are available as partial side information at both the encoder and decoder. We clarify upper and lower bounds of the optimal retrieval performance. This also implies certain relationships between retrieval parameters and performance.

Keywords— index searching, multiterminal source coding, rate-distortion theory, cascading, feedback

1 Introduction

The spread of the broadband network, capturing devices and the large amount of low-cost storage enables us to capture a huge number of music clips, images and movies easily. Therefore, media information retrieval must be developed that extracts desired information from multimedia archives quickly and accurately. The research issues related to media information retrieval can be divided into two main categories: the acceleration of retrieval, and the improvement of robustness against noise, distortion and signal fluctuation. This paper especially focuses on fundamental insights related to retrieval acceleration.

Multidimensional indexing methods (e.g. [1, 2]), which associate data items expressed as multidimensional vectors with indexes that represent the data items concisely, have been the basis of strategies for accelerating retrieval. Fig. 1 shows the framework of a retrieval system that employs multidimensional indexing methods. This retrieval system has two stages. The first stage prunes irrelevant data items by utilizing indexes. First, an index is provided for each feature vector extracted from a data item in the database. Then, features relevant to query items are chosen based solely on the indexes. In the second stage, surviving data items are matched. Query items and features selected in the first stage are matched, usually by calculating the distances between corresponding features.

This report describes an information-theoretical aspect of indexing and index searching. This type of retrieval can be formulated as a certain kind of multiterminal source coding problem, as shown in Fig. 2. The first stage is equivalent to the coding problem reported by Wyner and Ziv [3], where side information

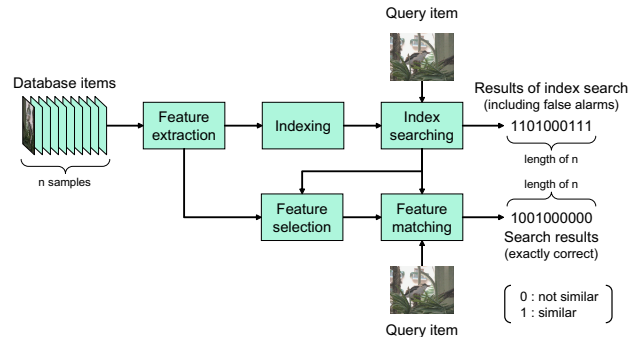


Figure 1: Signal retrieval with indexes

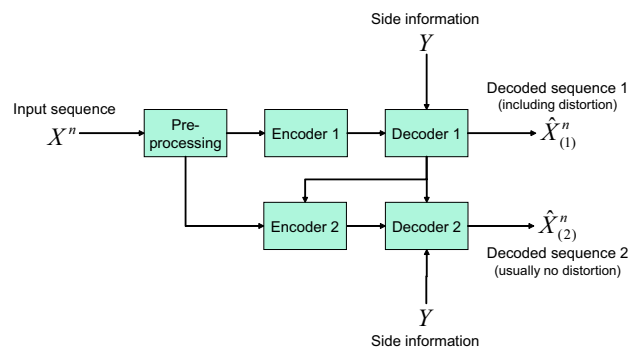


Figure 2: Model of signal retrieval with indexes

is available only at the decoder. The second stage is considered to be a coding problem such that some of the decoding results are available as partial side information at the encoder and decoder, and side information provided at the first stage is also available only at the decoder. Then, we clarify upper and lower bounds of the achievable rate-distortion region of that model. This implies the optimal retrieval performance and the relationship between search parameters, such as the index size, the time for executing the retrieval and the accuracy of the retrieval.

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.

Remark . We have investigated the coding problem shown in Fig. 2 and clarified outer and inner bounds of the achievable rate-distortion region (see [4]). However, later we consider a situation where

- side information is expressed as a single letter,
- source information to be encoded and side information are independently distributed,
- distortion functions depend on side information.

This is not the case considered in the previous report.

* The authors are with the Department of Communications and Integrated Systems, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan. Akisato Kimura is also with NTT Communication Science Laboratories, NTT Corporation, 3-1 Morinosato Wakamiya, Atsugi-shi, Kanagawa, 243-0198 Japan.

2 Related work

Of the many studies that have focused on the theoretical analysis of retrieval performance, especially for nearest neighbor searches with multidimensional indexes, most have been based on geometrical approaches (e.g. [5, 6]). With these approaches, the probability distributions of data items are assumed to be uniform, which makes it possible to consider volumes in multidimensional spaces as probabilities. Therefore, it is difficult to evaluate the performance for other types of distributions with their approaches.

In recent years, some work based on information-theoretical approaches has been reported [7, 8], and related applications have also been considered in recent reports [9, 10]. With this approach, we can take account of various kinds of probability distributions. For example, Tuncel et al. [7] formulated an approximate similarity search as a kind of multiterminal source coding problem similar to Wyner-Ziv coding with successive refinement [11], and clarified certain relationships between search time, search threshold, and amount of storage.

On the other hand, our work provides an information-theoretical model of signal retrieval with indexes, and clarifies the optimal performance of the retrieval by focusing particularly on relationships between index size, search accuracy and search time.

3 Notations and definitions

3.1 Preliminaries

Let \mathcal{X} be a finite set, \mathcal{B} a binary set, and \mathcal{B}^* a set of all finite sequences in the alphabet \mathcal{B} . Let $|\mathcal{X}|$ be the cardinality of \mathcal{X} and $\mathcal{I}_M = \{1, 2, \dots, M\}$. A member of \mathcal{X}^n is written as $x^n = (x_1, x_2, \dots, x_n)$, and substrings of x^n are written as $x_i^j = (x_i, x_{i+1}, \dots, x_j)$ for $i \leq j$. When the dimension is clear from the content, vectors will be denoted by boldface letters, i.e., $\mathbf{x} \in \mathcal{X}^n$. $\mathcal{M}(\mathcal{X})$ denotes the set of all probability distributions on \mathcal{X} . Also for a finite set \mathcal{Z} , $\mathcal{M}(\mathcal{X}|P_Z)$ denotes the set of all probability distributions on \mathcal{X} given a distribution $P_Z \in \mathcal{M}(\mathcal{Z})$, namely each member of $\mathcal{M}(\mathcal{X}|P_Z)$ is characterized by $P_{XZ} \in \mathcal{M}(\mathcal{X} \times \mathcal{Z})$ as $P_{XZ} = P_{X|Z}P_Z$. A discrete memoryless source is denoted by an infinite sequence $\{X_i\}_{i=1}^\infty$ of independent and identically distributed copies of a random variable X taking values in \mathcal{X} with a generic distribution $P_X \in \mathcal{M}(\mathcal{X})$. We will denote a source by referring to its random variable X . We will use $H(\cdot)$ and $I(\cdot; \cdot)$ to denote the entropy and mutual information of a set of random variables. A similar convention is used for other random variables and vectors. In the following, all bases of exponentials and logarithms are set at 2.

In the following, we denote a source to be encoded as X . Here, let us introduce a random variable Y having values in a finite set \mathcal{Y} with a distribution P_Y . The random variables X and Y are assumed to be independently distributed, i.e. the joint distribution P_{XY} is written as $P_{XY} = P_X P_Y$. Also, let U , V and W be auxiliary random variables whose joint distribution

P_{UVW} is unspecified, and which have values in sets \mathcal{U} , \mathcal{V} and \mathcal{W} , respectively.

Let $\hat{\mathcal{X}}$ be a reproduction alphabet, and $\Delta : \mathcal{X} \times \hat{\mathcal{X}} \times \mathcal{Y} \rightarrow [0, \infty)$ be a single-letter distortion function that depends on side information Y , denoted as $\Delta(x, \hat{x}; y)$. The vector distortion function is defined as

$$\Delta^n(\mathbf{x}, \hat{\mathbf{x}}; \mathbf{y}) = \frac{1}{n} \sum_{k=1}^n \Delta(x_k, \hat{x}_k; y_k). \quad (1)$$

4 Problem formulation

Definition 1. (Index Searching (IS) code)

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

$$\begin{aligned} \varphi_{(1)}^n &: \mathcal{X}^n \rightarrow \mathcal{B}^*, & \varphi_{(01)}^n &: \mathcal{A}_n^{(1)} \times \mathcal{Y} \rightarrow \mathcal{B}^*, \\ \varphi_{(02)}^n &: \mathcal{A}_n^{(1)} \times \mathcal{Y} \rightarrow \mathcal{B}^*, & \varphi_{(2)}^n &: \mathcal{A}_n^{(01)} \times \mathcal{X}^n \rightarrow \mathcal{B}^*, \\ \hat{\varphi}_{(1)}^n &: \mathcal{A}_n^{(1)} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}^n, \\ \hat{\varphi}_{(2)}^n &: \mathcal{A}_n^{(02)} \times \mathcal{A}_n^{(2)} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}^n, \end{aligned}$$

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

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

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 the source X and side information Y such that

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{1}{n} E \left[l(\mathcal{A}_n^{(i)}) \right] &\leq R_i, \quad (i = 01, 02, 1, 2) \\ \limsup_{n \rightarrow \infty} E \left[\Delta^n(X^n, \hat{X}_{(i)}^n; Y) \right] &\leq D_j, \quad (j = 1, 2) \end{aligned}$$

where

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

Definition 3. (IS-achievable rate region)

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

5 Association with implementation

Fig. 3 shows an example of image retrieval implementation with indexes. Each sample x_i of a sequence x^n emitted from the source X corresponds to a feature vector extracted from a stored image, where n is the number of stored images. A letter y emitted from a side information source Y corresponds to a query provided

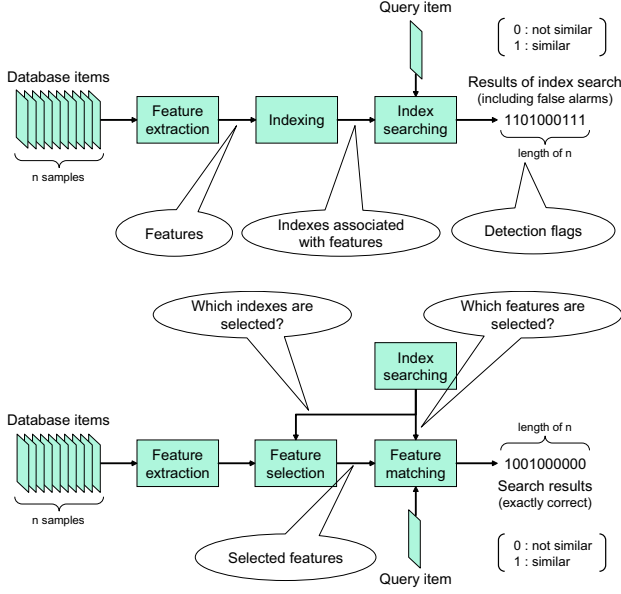


Figure 3: Implementation example of signal retrieval with indexes: above: first stage, below: second stage

by users. Although several types of queries are possible, here y is assumed to be a set of feature vectors extracted from image examples provided by users. It is natural for the number of image examples to be much smaller than the number of images in the database. Therefore, y is more suitable for representing queries than y^n .

Each encoder and decoder is composed of sample-wise encoders and decoders, all of which have the same functional form, e.g.,

$$\varphi_{(1)}^n(\mathbf{x}) = \varphi_{(1)}(x_1) * \cdots * \varphi_{(1)}(x_n),$$

where $*$ stands for an operator that performs string concatenation. The encoder $\varphi_{(1)}$ provides an index for each feature. Examples of indexing operations involve the vector quantization of feature vectors [12] and multidimensional indexing methods [1, 2].

To provide an example of operations for the decoder $\hat{\varphi}_{(1)}$, let us introduce detection flags denoted as

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

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

where $\mathcal{C}_\delta(y)$ is an *interest region* of the query y (See Figure 4). Usually, the interest region of the query y is defined as a region where the distance from y falls below a predefined threshold value δ . We consider that x_i is similar to y when $z_i = 1$. One of the main strategies for determining each element $\hat{x}_{(1)i}$ of reproduction sequences $\hat{\mathbf{x}}_{(1)}$ is to select the item most similar to the query from the set of items to which the given index $\varphi_{(1)}(x_i) = k_{(1)}(i)$ ($i \in \mathcal{I}_n$) is assigned, e.g.,

$$\hat{x}_{(1)i} \stackrel{\text{def.}}{=} \hat{\varphi}_{(1)}(k_{(1)}(i), y)$$

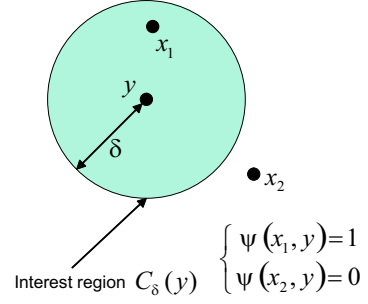


Figure 4: Interest regions and detection flags

$$\stackrel{\text{def.}}{=} \arg \min_{\tilde{x} \in \mathcal{X}, \varphi_{(1)}(\tilde{x}) = k_{(1)}(i)} d(\tilde{x}, y),$$

where $d: \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$ is a distance function such as Euclidean distance. Note that $\mathcal{X} = \hat{\mathcal{X}} = \mathcal{Y}$ is assumed. An example of single-letter distortion functions Δ is the Hamming distance $d_h: \mathcal{B} \times \mathcal{B} \rightarrow \mathcal{B}$ between detection flags, denoted as

$$\begin{aligned} \Delta(x_i, \hat{x}_{(1)i}; y) &= d_h(\psi(x_i, y), \psi(\hat{x}_{(1)i}, y)) \\ &= d_h(z_i, \hat{z}_{(1)i}). \end{aligned}$$

$\Delta = 1$ means that there is either a false dismissal (it is determined that $\hat{x}_{(1)i}$ is not similar to y , whereas x_i is similar to y) or a false detection (it is determined that $\hat{x}_{(1)i}$ is similar to y , whereas x_i is not similar to y).

The second feedback encoder $\varphi_{(02)}^n$ sends a sequence $\hat{\mathbf{z}}_{(1)}$ of detection flags directly to the second decoder $\hat{\varphi}_{(2)}^n$, namely informs which data items survive.

$$\varphi_{(02)}(k_{(1)}(i), y) = \hat{z}_{(1)i} \quad (i \in \mathcal{I}_n). \quad (2)$$

On the other hand, the first feedback encoder $\varphi_{(01)}$ informs only which indexes survive. Since the number of indexes is usually set much smaller than the number of data items (otherwise indexing has little meaning), the coding rate of $\varphi_{(01)}^n$ can be reduced to 0.

The encoder $\varphi_{(2)}$ of the second stage knows which indexes survive. Therefore, it suffices to send features corresponding solely to the surviving data items. When the false detection rate is low in the first stage, the coding rate of the second encoder can be reduced. On the other hand, if the false detection rate is high, the coding rate must increase. This implies that our coding model would involve a trade-off between the coding rate R_1 of the first stage and that of the second stage R_2 .

The decoder $\hat{\varphi}_{(2)}$ of the second stage knows which data items survive. Since it has already been determined that items removed during the first stage are not similar to the query item y , such data items do not need to be reproduced. Therefore, the reproduction function for the second stage can be easily determined as follows:

$$\begin{aligned} \hat{x}_{(2)i} &\stackrel{\text{def.}}{=} \hat{\varphi}_{(2)}(\varphi_{(2)}(i), z_{(2)i}, y) \\ &\stackrel{\text{def.}}{=} \begin{cases} x_i & \text{(if } z_{(2)i} = 1) \\ \text{arbitrary} & \text{(if } z_{(2)i} = 0) \end{cases}, \end{aligned}$$

$$\varphi_{(2)}(i) \stackrel{\text{def.}}{=} (\varphi_{(2)}(x_i, \varphi_{(01)}(k_{(1)}(i), y)).$$

The distortion function used in the first stage also employed in the second stage.

$$\begin{aligned} \Delta(x_i, \hat{x}_{(2)i}; y) &= d_h(\psi(x_i, y), \psi(\hat{x}_{(2)i}, y)) \\ &= d_h(z_i, \hat{z}_{(2)i}). \end{aligned}$$

6 Statement of results

Here, we state the main theorem.

Theorem 1. (Coding theorem of IS code)

$$\begin{aligned} \mathcal{R}_{IS}(X, Y|D_1, D_2) \subseteq \{ &(R_{01}, R_{02}, R_1, R_2) : \\ &R_1 \geq I(X; UV), \\ &R_{02} \geq I(X; V), \\ &R_2 \geq I(X; W|V) \} \quad (\text{outer bound}) \end{aligned}$$

where random variables U , V and W whose alphabets are \mathcal{U} , \mathcal{V} and \mathcal{W} , respectively, are selected such that

- The alphabet sizes are bounded as

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

- The Markov chain $U \rightarrow X \rightarrow Y$ is satisfied,
- There exist functions $\phi_{(1)} : \mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$ and $\phi_{(2)} : \mathcal{W} \times \mathcal{V}^{(2)} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$, which satisfy

$$\begin{aligned} D_1 &\geq E[\Delta(X, \phi_{(1)}(U, Y); Y)], \\ D_2 &\geq E[\Delta(X, \phi_{(2)}(W, V, Y); Y)]. \end{aligned}$$

An inner bound is obtained in the same functional forms, while the Markov chain is replaced as

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

The achievable rate region indicated in Theorem 1 is very similar to that of the cascading refinement system [4]. This indicates that information emitted from the first feedback encoder $\varphi_{(01)}^n$ has little meaning in terms of improving the asymptotic performance of the system.

Acknowledgements

The authors first wish to thank Prof. Te Sun Han, Prof. Mamoru Hoshi and Prof. Hiroyoshi Morita of University of Electro-Communications, and Prof. Ryutaro Matsumoto of Tokyo Institute of Technology for their valuable discussions and helpful comments. The authors also thank Dr. M. Brandon Westover of Washington University at Saint Louis and Prof. Yasutada Oohama of Kyushu University for providing useful research documents. [9, 10, 13].

References

- [1] N. Beckman and H.P. Kriegel, "The R*-tree : an efficient and robust access method for points and rectangles," Proc. ACM SIGMOD International Conference on Management of Data (ACM SIGMOD), pp.322–331, June 1990.
- [2] N. Katayama and S. Satoh, "The SR-tree : an index structure for high-dimensional nearest neighbor queries," Proc. ACM SIGMOD International Conference on Management of Data (ACM SIGMOD), pp.369–380, May 1997.
- [3] A.D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," IEEE Trans. Inform. Theory, Vol.22, No.1, pp.1–10, January 1976.
- [4] A. Kimura and T. Uyematsu, "Multiterminal source coding for cascading and feedback refinement systems," Proc. Shannon Theory Workshop (STW), pp.25–31, September 2006.
- [5] J. Friedman, J. Bentley, and R. Finkel, "An algorithm for finding best matches in logarithmic expected time," ACM Trans. Math. Software, Vol.3, No.3, pp.209–226, September 1977.
- [6] S. Berchtold, C. Boehm, D. Keim, F. Frebs, and H.P. Kriegel, "A cost model for nearest neighbor search in high dimensional data spaces," Proc. ACM SIGMOD International Conference on Management of Data (ACM SIGMOD), pp.78–86, May 1997.
- [7] E. Tuncel, P. Koulgi, and K. Rose, "Rate-distortion approach to databases: storage and content-based retrieval," IEEE Trans. Inform. Theory, Vol.50, No.6, pp.953–967, June 2004.
- [8] J. Nayak, S. Ramaswamy, and K. Rose, "Correlated source coding for fusion storage and selective retrieval," Proc. IEEE International Symposium on Information Theory (ISIT), pp.92–96, September 2005.
- [9] M.B. Westover and J.A. O'Sullivan, "Achievable rates for pattern recognition: binary and Gaussian cases," Proc. IEEE International Symposium on Information Theory (ISIT), pp.28–32, September 2005.
- [10] M.B. Westover and J.A. O'Sullivan, "Achievable rates for pattern recognition," IEEE Trans. Inform. Theory, 2005. submitted.
- [11] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner-Ziv problem," IEEE Trans. Inform. Theory, Vol.50, No.8, pp.1636–1654, August 2004.
- [12] A. Kimura, K. Kashino, T. Kurozumi, and H. Murase, "A quick search method for multimedia signals using global pruning," IEICE Trans. Informations and Systems, Vol.J85-D-II, No.10, pp.1552–1562, October 2002. (in Japanese).
- [13] A.H. Kaspi, Rate-distortion for correlated sources with partially separated encoders, Ph.D. thesis, Cornell University, January 1979.