Chapter 7
Structural Information Theory, I:

Mathematical Theory of Information Processing

Angsheng Li

BeiHang University

05, Nov., 2019

Outline

1. The changing CS
2. The challenges
3. Structural information theory
4. Information processing (IP)
5. Graph theory: New problems from computer science

The Changing Computing

What does a computer do?

1. The main stream of CS in the 20th century is to investigate:
   1.1 **Computable functions and computing devices**
   1.2 **Efficient algorithms**
2. The new mission of CS in the 21st century is to study:
   2.1 **Computing the real world**
   2.2 **Information processing**
   2.3 **Definition of intelligence**

What Is the Mathematical Definition of Information Processing (IP)?

The challenges are:

1. What is the mathematical definition of **information processing**?

Definition

(Information processing)**Given a system S, the information processing for S is to distinguish the laws from the noises in S.**

2. What is the **mathematical theory** that supports the current information processing?

The mathematical theory about the limit of distinguishing the laws from noises in a system.

Big Data Phenomenon

1. There are relationships among individuals of data
   - How to build the system of big data (unstructured data)?
2. (Assumption) The laws of big data exist in the relationships of data
3. Big data is an observed system in which laws are embedded in a structure of noises
4. The mission of data analysis is hence:
   To distinguish the laws from noises
   This is
   To decode the laws from the observed system of data
   Grand challenge:
   What is the relationship between information and computation?

Shannon's Information

Shannon, 1948:
Given a distribution $p = (p_1, p_2, \cdots, p_n)$, the Shannon's entropy is

$$H(p) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i. \tag{1}$$

p_i is the probability that item i is chosen, $-\log_2 p_i$ is the "self-information" of item i.

- Shannon's information measure the uncertainty of a probabilistic distribution.
- This metric and the associated notions of noises form the foundation of information theory and the information theoretic study in all areas of the current science.
- Shannon's metric provides the foundation for the current generation information technology.

Principles of Communications

Shannon's theory perfectly solves the **fundamental questions of communications**.

- The lower bound of data compression - entropy
- The transmission rate of point to point communication: channel capacity
- Shannon's theory guarantees that lossless information transmission is possible

However, Shannon's theory **fails to support the current information processing**, in which we are asked to distinguish the laws from the noises from dataspace.
Noting that, data are observed from the real world, real world data is a system evolved in nature. It includes laws and noises. Laws are embedded in a system of noises.
Urgently call for an information theory that supports the information processing, that is a mathematical theory that distinguishes the laws from noises of an observed data-space.

Understanding of Information

1. Shannon entropy: $H(p)$ is the quantification of uncertainty contained in a probability distribution or random variable (essentially, a function)
2. **Information** is defined as the amount of uncertainty that is eliminated.
   Choosing an item i according to probability distribution p, we obtained an information of amount exactly $H(p)$.
3. This suggests that
   - 3.1 Entropy is a **static metric** associated with an object (probability distribution above)
   - 3.2 Information is a **dynamic metric** determined by both an **object** and an **action**, random selection here

The Challenges

1. What is the entropy that is contained in a complex system such as graphs?
2. What is the quantification of information obtained from a complex system such as graphs?
3. How to **generate** the maximum amount of information?

Information vs Computation

1. Is information useful in computation?
2. What is the role of computation in information?
   Computing is decoding information, that is, eliminating uncertainty.
3. What is the relationship between information and intelligence?
   Information is the basis of intelligence.

The Challenges

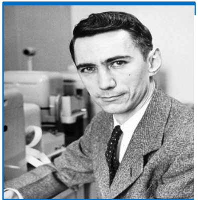
1. How to encode a graph?
2. How to define the information embedded in a graph?
3. Can a structural information decodes the essential structure of a physical system?
4. How to determine the folded structure of a DNA sequence?

Structural Information Theory

The **structural information theory** is a joint work with:
Yicheng Pan

Shannon

1953年， Shannon在IEEE Tranactions on Information Theory创刊号上发表论文讨论结构信息问题



Shannon

1916-2001
信息论创始人

THE LATTICE THEORY OF INFORMATION

by

C.E. Shannon

ABSTRACT

The word "information" has been given many different meanings by various writers in the general field of information theory. It is likely that at least a number of these will prove sufficiently useful in certain applications to deserve further study and permanent recognition. It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field. The present note outlines a new approach to information theory which is aimed specifically at the analysis of certain communication problems in which there exist a number of information sources simultaneously in operation. A typical example is that of a simple communication channel with a feedback path from the receiving point to the transmitting point. The problem is to make use of the feedback information for improving forward transmission, and to determine the forward channel capacity when the best possible use is made of this feedback information. Another more general problem is that of a communication system consisting of a large number of transmi ting and receiving points with some type of interconnecting network between the various points. The problem here is to formulate the best systems design whereby, in some sense, the best overall use of the available facilities is made. While the analysis sketched here has not yet proceeded to the point of a complete solution of these problems, partial answers have been found and it is believed that a complete solution may be possible.

很难说单独一个信息的概念能解决所有的应用

……

另一个更一般的问题是建立一个结构信息的度量来支持通信系统的分析。

Shannon's Question, 1953

Shannon, 1953:

1. Is there a structural theory of information that supports communication network analysis?
2. What is the optimal communication network?

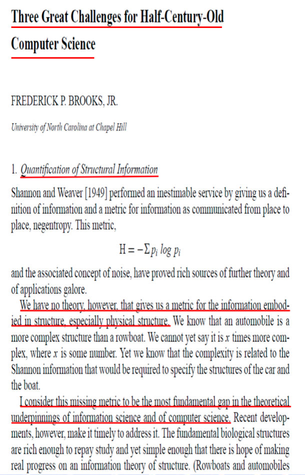
Shannon noticed that his theory fails to support communication network. The reason is as follows.
Given a communication network G,

1. (De-structuring) Let p be a distribution computed from G, degree distribution, or distance distribution, and so on. This discards the interesting properties of G.
2. Define $H(p)$ to be the information of G.
   This number $H(p)$ does not tell us anything about the interactions and communications occurred in G.

The question is hence:
How to measure the information embedded in a physical system?

Brooks Question

2003年， Brooks在Journal of the ACM发表论文讨论计算机科学挑战问题



Three Great Challenges for Half-Century-Old Computer Science

FREDERICK P. BROOKS, JR.

University of North Carolina at Chapel Hill

1. *Quantification of Structural Information*

Shannon and Weaver [1949] performed an inestimable service by giving us a definition of information and a metric for information as communicated from place to place, negentropy. This metric,

$$H = -\Sigma p_i \log p_i$$

and the associated concept of noise, have proved rich sources of further theory and of applications galore.

We have no theory, however, that gives us a metric for the information embodied in structure, especially physical structure. We know that an automobile is a more complex structure than a rowboat. We cannot yet say it is x times more complex, where x is some number. Yet we know that the complexity is related to the Shannon information that would be required to specify the structures of the car and the boat.

I consider this missing metric to be the most fundamental gap in the theoretical underpinnings of information science and of computer science. Recent developments, however, make it timely to address it. The fundamental biological structures are rich enough to repay study and yet simple enough that there is hope of making real progress on an information theory of structure. (Rowboats and automobiles

半个世纪计算机科学的三个重大挑战

1. 结构信息度量

然而，我们没有一个理论来度量嵌入在一个结构，特别物理结构中的信息。

……

我认为这个度量的缺失是信息科学和计算机科学理论基石的一个最根本的缺陷。

Brooks Question, 2003

- We have no theory that gives us a metric for the information embedded in structures
- The missing metric is the most fundamental gap in information science and computer science.
  The quantification of structural information as the first of the three great challenges for half-century old CS.

Physical Systems

Given a physical system G, the information embedded in G should determine and decode the *essential structure* of G.
For example, for a car and a boat, the essential structures of the two objects should be different, and the essential structures of a car and a boat should be determined by the information embedded in the car and the boat respectively.
Question: What is the essential structure of a physical system?

Rashevsky, 1955

Given a connected graph G of n vertices, for every vertex i, let n_i be the number of vertices in the orbit of vertex i (under automorphisms of G).
Suppose that there are k orbits with number of vertices $n_1, n_2, \cdots, n_k$. Then let

$$p = (\frac{n_1}{n}, \frac{n_2}{n}, \cdots, \frac{n_k}{n}).$$

Define the entropy of G to be the Shannon entropy of p.

Local Entropy Measures

Raychaudhury et al, 1984

Given a connected graph G, for vertices i, j, let $d(i, j)$ be the distance between vertex i and vertex j.

Define the entropy of G to be the Shannon entropy of the distributions of the distances $\{d(i, j)\}$.

Gibbs Entropy

This measures the number of bits needed to determine a graph generated from some model.

Shannon's Entropy for Graph Models

It measures the number of bits needed to describe the graph that is generated from a model.

Von Neumann Entropy

This is defined by the spectral of the Laplacian of the graph. That is the distribution of the eigenvalues of the Laplacian of the graph.

Remark: Quantum information theory is built by using this definition, however, the quantum capacity similar to Shannon cannot be defined.

Encoding An Alphabet

Given an alphabet $\Sigma = \{1, 2, \cdots, n\}$ and a probability distribution $\mathrm{p} = (p_1, p_2, \cdots, p_n)$, suppose that the probability that i occurs with probability p_i. We will encode the alphabet $\Sigma = \{1, 2, \cdots, n\}$ by binary strings $\alpha_1, \alpha_2, \cdots, \alpha_n$, respectively such that

$$L = \sum_{i=1}^{n} p_i \cdot |\alpha_i| \tag{2}$$

is minimised.

Huffman Codes

It is a binary tree T such that each item i is coded by a leaf α_i of T such that if p_i is large, then the length of α_i is short, and if p_i is small, then the length of α_i is large.
In this case, α_i is called the *codeword* of i.
This can be realised recursively by a greedy algorithm that is actually optimum.
In this case, we define

$$L^T = \sum_{i=1}^{n} p_i \cdot |\alpha_i|, \tag{3}$$

where α_i is the leaf of T that is the codeword of i.
The Huffman codes find a tree T such that L^T is minimized.

Length of binary expression

It is shown that

$$H(\mathrm{p}) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

is the tight lower bound of the length L^T, where T is the tree found by the Huffmman codes.

Number of Bits Required to Guess the Item

Therefore, the Shannon entropy of a probability distribution can be approximately understood as the minimum average length of the binary representation of the item chosen by probability distribution p.
This gives an intuitive understanding of the Shannon entropy.
In addition, the Shannon entropy can be regarded as the minimum number of guesses to determine the item chosen by the probability distribution p.
However, **probability distribution is structure free**.

The Challenges

- **How to encode a graph**?
- **What is the entropy of a graph**?

There are encoding of graphs.

Intersection Representation

Given a graph $G = (V, E)$, for every $v \in V$, we assign a subset of $T = \{1, 2, \cdots, t\}$, written S_v.
For two vertices $u, v \in V$, $u \sim v$ if and only if $S_u \cap S_v \neq \emptyset$.
For this encoding, we define the *intersection number of G* to be the least number t, the size of T such that there is an encoding as above.

Product Representation

Given a graph $G = (V, E)$ and a set $T = \{1, 2, \cdots, t\}$, a *product representation* of G assign every vertex $v \in V$ a vector σ_v of the form $(a_1, a_2, \cdots, a_k)$ such that for every pair u, v of vertices, $u \sim v$ if and only if, for $\sigma_u = (a_1, a_2, \cdots, a_k)$ and $\sigma_v = (b_1, b_2, \cdots, b_k)$, $a_j \neq b_j$ for every j from 1 to k.
The product dimension of G is the least number k such that there is an encoding as above.

Squashed-cube Embedding of Length N

Fix $S = \{0, 1, \star\}$. Define $d(0,1) = d(1,0) = 1$, and for $\{a, b\} \neq \{0, 1\}$, $d(a, b) = 0$.
For every graph $G = (V, E)$, for every $v \in V$, we define a vector $\theta_v \in S^N$, such that for every pair $u, v \in V$,

$$d_G(u, v) = \sum_{i=1}^{N} d(a_i, b_i)$$

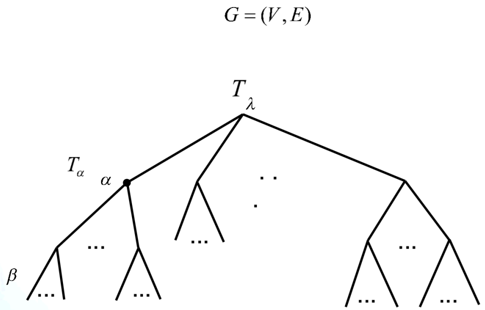
where $\theta_u = (a_1, a_2, \cdots, a_N)$, and $\theta_v = (b_1, b_2, \cdots, b_N)$.
For a graph G, what is the least number N?

Encoding a Graph

The existing encodings of graphs fail to decode the information embedded in the graph.

Encoding Tree

怎样对一个大规模图编码？



- 层谱划分
- 类似于身份证号或
- 通信地址：省、市、县、区、街道、门牌号

- 叶子节点 β，$T_\beta=\{u\}$，称 β 是图G中顶点 u 的**码字**

Figure: Encoding Tree

Encoding Tree

Definition

(Encoding tree of graphs) Let $G = (V, E)$ be an undirected and connected network. We define the *encoding tree T of G* as a tree T with the following properties:

(1) For the root node denoted λ, we define the set $T_\lambda = V$.

(2) For every node $\alpha \in T$, the immediate successors of α are $\alpha\hat{}\langle j\rangle$ for j from 1 to a natural number N ordered from left to right as j increases, where every internal node has at least two immediate successors. Therefore, $\alpha\hat{}\langle i\rangle$ is to the left of $\alpha\hat{}\langle j\rangle$ written as $\alpha\hat{}\langle i\rangle <_{\mathrm{L}} \alpha\hat{}\langle j\rangle$, if and only if $i < j$.

(3) For every $\alpha \in T$, there is a subset $T_\alpha \subset V$ that is associated with α.
For α and β, we use $\alpha \subset \beta$ to denote that α is an initial segment of β. For every node $\alpha \neq \lambda$, we use α^- to denote the longest initial segment of α, or the longest β such that $\beta \subset \alpha$.

Encoding Tree - II

(4) For every i, $\{T_\alpha \mid h(\alpha) = i\}$ is a partition of V, where $h(\alpha)$ is the height of α (note that the height of the root node λ is 0, and for every node $\alpha \neq \lambda$, $h(\alpha) = h(\alpha^-) + 1$).

(5) For every α, T_α is the union of T_β for all β's such that $\beta^- = \alpha$; thus, $T_\alpha = \cup_{\beta^- = \alpha} T_\beta$.

(6) For every leaf node α of T, T_α is a singleton; thus, T_α contains a single node of V.

Encoding Tree - III

(7) For every node $\alpha \in T$, if $T_\alpha = X$ for a set of vertices X, then we say that α is the **codeword** of X, and that X is the **marker** of α.

(8) For every vertex $v \in V$, there is a leaf node $\alpha \in T$ such that $T_\alpha = \{v\}$, that is, there is a unique **codeword** of v in T.

(9) Every leaf node in T is a codeword of a unique vertex (**marker**) in V.

Therefore, the set of the leaf nodes in T is the set of codewords of all the vertices in G.

Structural Entropy by an Encoding Tree

Definition

(Structural entropy of a graph by an encoding tree) For an undirected and connected network $G = (V, E)$, suppose that T is an encoding tree of G.

We define the **structural entropy of G by the encoding tree T** as follows:

$$\mathcal{H}^T(G) = -\sum_{\alpha \in T, \alpha \neq \lambda} \frac{g_\alpha}{2m} \log_2 \frac{V_\alpha}{V_{\alpha^-}}, \tag{4}$$

where g_α is the number of edges from nodes in T_α to nodes outside T_α, V_β is the volume of set T_β, namely, the sum of the degrees of all the nodes in T_β.

Intuition

- The encoding tree T gives a codeword for each vertex in G
- $\mathcal{H}^T(G)$ is the amount of information required to determine the codeword of the vertex that is accessible from random walk in G.

K-dimensional Structural Entropy

Definition

(K-dimensional structural entropy) Let $G = (V, E)$ be a connected network.

- We define the **K-dimensional structural entropy of G** as follows:

$$\mathcal{H}^K(G) = \min_T \{\mathcal{H}^T(G)\}, \tag{5}$$

where T ranges over all of the encoding trees of G of height $\leq K$.

Structural Entropy

Definition

(Structural entropy) Let $G = (V, E)$ be a connected network.

- We define the **structural entropy of G** as follows:

$$\mathcal{H}(G) = \min_{T}\{\mathcal{H}^{T}(G)\}, \tag{6}$$

where T ranges over all of the encoding trees of G.

Information Processing

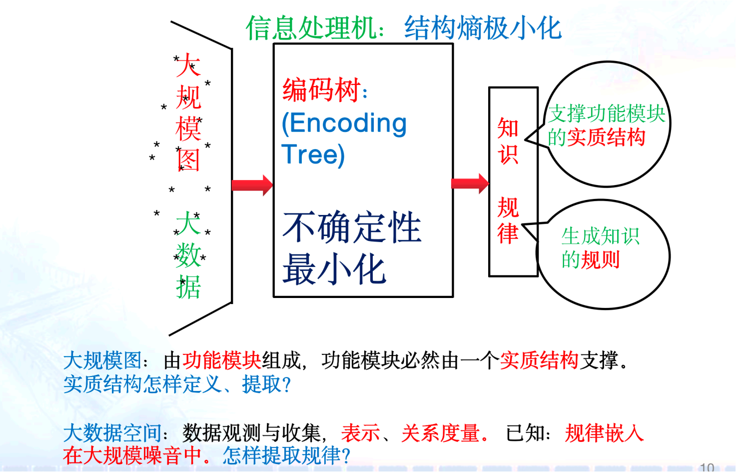


Figure: Information Processing

Losslessly Encoding Principle

Key point:

Encoding tree of a graph is a loseless encoding of the graph, which eliminates the uncertainty embedded in the graph, and does not do anything else

This differs from the existing encodings of graph defined previously.

Shannon Entropy

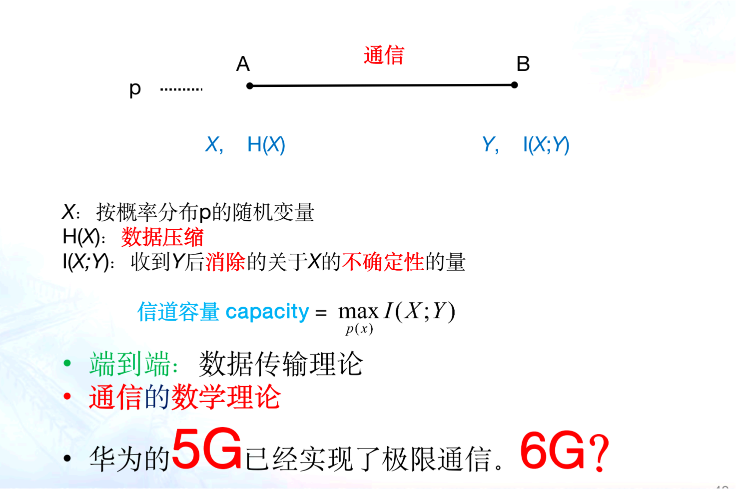


Figure: Shannon Entropy

One-dimensional Structural Entropy

Let $G = (V, E)$ be a connected graph, and p be the degree distribution of G. Then

$$\mathcal{H}^1(G) = H(p), \tag{7}$$

written $H(G)$, called **Shannon entropy of G**.

Lower Bound of Structural Entropy

Theorem

Let $G = (V, E)$ be an undirected, connected graph. Then:

$$\mathcal{H}(G) \geq \Phi(G) \cdot (H(G) - 1),$$

where $\Phi(G)$ is the conductance of G and $H(G)$ is the Shannon entropy of G, or the one-dimensional structural entropy of G.

General Lower Bound

Let $G = (V, E)$ be a connected graph of n vertices. Then

$$\mathcal{H}^2(G) = \Omega(\log \log n). \qquad (8)$$

Structural Entropy of a Probability Distribution

Let p be a probability distribution, and T be an encoding tree of p. Then

$$\mathcal{H}^T(p) = H(p). \tag{9}$$

Phase Transition

Let G be a graph generated from the small world model. Then:

1. If $r < 2$, then

$$\mathcal{H}^2(G) = \Omega(\log n). \tag{10}$$

2. If $r \geq 2$, then

$$\mathcal{H}^2(G) = O(\log \log n). \tag{11}$$

Two-Dimensional Structural Entropy

Theorem

(With Pan Peng, 2019) For every real number a with $0 < a < 1$, there exists a family of graphs $\{G_n\}$ such that for $G = G_n$,

$$\mathcal{H}^2(G) = \Theta((\log_2 n)^a). \tag{12}$$

Intuition

Given a connected graph G, the structural entropy $\mathcal{H}(G)$ is the minimum number of bits required to determine the codeword of an encoding tree of the graph for the vertex that is accessible from random walk with stationary distribution in the graph.

Resistance

Let $G=(V,E)$ be a connected graph, and $\mathcal{P}=\{X_1,X_2,\cdots X_N\}$ be a partition of the vertices V.
Define resistance of G by $\mathcal{P}$ as

$$\mathcal{R}^{\mathcal{P}}(G)=-\sum_{j=1}^{N}\frac{V_j-g_j}{2m}\log_2\frac{V_j}{2m}. \tag{13}$$

Define the resistance of G by

$$\mathcal{R}(G)=\max_{\mathcal{P}}\{\mathcal{R}^{\mathcal{P}}(G)\}. \tag{14}$$

Theorem

$$\mathcal{R}(G)=\mathcal{H}^1(G)-\mathcal{H}^2(G). \tag{15}$$

If $\mathcal{R}(G)$ is large, then G is secure against cascading failure of virus attacks.

Resistance of Regular Graphs

Theorem

(With Y. Pan) For any $d > 2$, any d-regular graph G,

$$\mathcal{R}(G) \geq (\frac{2}{d} - o(1)) \log_2 n \tag{16}$$

Theorem

(With P. Peng) For any $d > 2$, for a random d-regular graph G, with high probability,

$$\mathcal{R}(G) \leq (\frac{2}{d} + o(1)) \log_2 n. \tag{17}$$

Information Compression and Enrichment

Let $G = (V, E)$ be a connected graph, and $\mathcal{P} = \{X_1, X_2, \cdots X_N\}$ be a partition of the vertices V.

Define **compression information of G by $\mathcal{P}$** as

$$\mathcal{C}^{\mathcal{P}}(G) = -\sum_{j=1}^{N} \frac{V_j - g_j}{V_j} \frac{V_j}{2m} \log_2 \frac{V_j}{2m}, \tag{18}$$

in which

- $\frac{V_j - g_j}{V_j}$ is the fraction of information of G compressed in X_j,
- $-\frac{V_j}{2m} \log_2 \frac{V_j}{2m}$ is the information of G contained in X_j,
- we call $-\frac{V_j - g_j}{V_j} \frac{V_j}{2m} \log_2 \frac{V_j}{2m}$ the **information enrichment** of X_j in G, written $\gamma_G(X_j)$.

We define the **information enrichment of G** by

$$\gamma(G) = \max_{X \subset V, V_X \leq m} \{\gamma_G(X)\}. \tag{19}$$

Information Enrichment vs Intelligence?

Project:

- Information enrichment of graphs
- Information distribution of graphs

Compressing Information by Encoding Tree

Definition

(Compressing information of a graph by an encoding tree) For an undirected and connected network $G = (V, E)$, suppose that T is an encoding tree of G.

We define the **compressing information of G by the encoding tree T** as follows:

$$\mathcal{C}^T(G) = -\sum_{\alpha \in T, \alpha \neq \lambda} \frac{V_\alpha - g_\alpha}{2m} \log_2 \frac{V_\alpha}{V_{\alpha^-}}, \tag{20}$$

where g_α is the number of edges from nodes in T_α to nodes outside T_α, V_β is the volume of set T_β, namely, the sum of the degrees of all the nodes in T_β.

Decoding Information by Encoding Tree

Definition

(Decoding information of a graph by an encoding tree) For an undirected and connected network $G = (V, E)$, suppose that T is an encoding tree of G.

We define the **decoding information of G by the encoding tree T** as follows:

$$\mathcal{D}^T(G) = \mathcal{H}^1(G) - \mathcal{H}^T(G). \tag{21}$$

Compressing Information

Definition

(Structural information) Let $G = (V, E)$ be a connected network.

- We define the **compressing information of G** as follows:

$$\mathcal{C}(G) = \max_{T}\{\mathcal{C}^{T}(G)\}, \tag{22}$$

where T ranges over all of the encoding trees of G.

Compressible Graphs

Definition

Given G, k and ρ, we say that G is (n, k, ρ)-compressible, if:

$$\rho^k(G) \geq \rho,$$

where

$$\rho^k(G) = \frac{\mathcal{C}^k(G)}{\mathcal{H}^1(G)}$$

Information Compressing

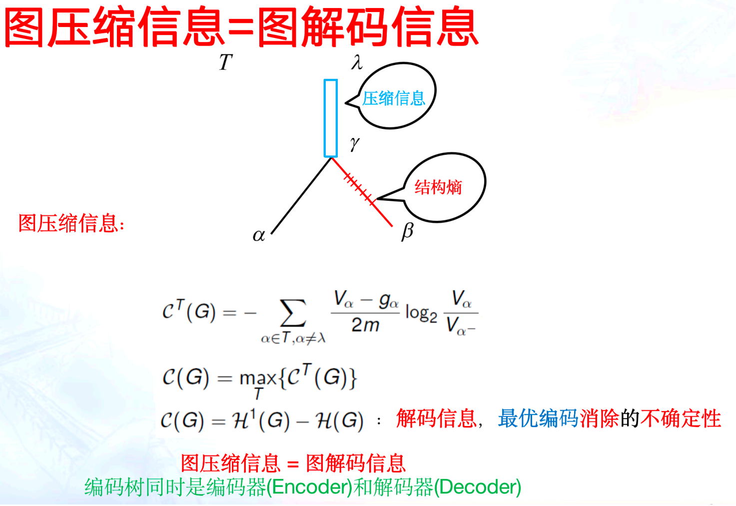


Figure: Information Compressing

Algebraic Properties

1. There are graphs with $\rho^2(G) = 1 - \epsilon$, so highly compressible.
2. There are graphs cannot be significantly compressed.

Theorem

If G is highly compressible, then there is a large $k = \frac{n}{\mathrm{poly} \log n}$ *such that* $\lambda_k = o(1)$*, where* λ_k *is the k-th largest eigenvalue of the Laplacian of G.*

Decoding Information

Definition

(Decoding information) Let $G = (V, E)$ be a connected network.

- We define the **decoding information of G** as follows:

$$\mathcal{D}(G) = \max_{T}\{\mathcal{D}^{T}(G)\}, \tag{23}$$

where T ranges over all of the encoding trees of G.

Shannon Entropy

Theorem

(Shannon entropy) Let $G = (V, E)$ be a connected network. Let p be the degree distribution of G. Then:

$$\mathcal{H}^1(G) = H(p) \tag{24}$$

Given a probability distribution p, we can define the coding tree of p and the structural entropy of p given by T. In this case, we have:
For any encoding tree T,

$$\mathcal{H}^T(p) = H(p). \tag{25}$$

Compressing and Decoding Principle

Theorem

Let $G = (V, E)$ be a connected network. Then:

$$\mathcal{C}(G) = \mathcal{H}^1(G) - \mathcal{H}(G) = \mathcal{D}(G). \tag{26}$$

Therefore, **any information lost in the compression of data can be losslessly decoded by an encoding tree, the decoder**.

This means that

For either unstructured or structured data, data compression will never loss any information

Shannon Entropy and Structural Entropy Are Complement

The compressing and decoding theorem:

$$\mathcal{C}(G) = \mathcal{H}^1(G) - \mathcal{H}(G) = \mathcal{D}(G). \tag{27}$$

indicates that the Shannon entropy minus the structural entropy is exactly the compressing information and the decoding information, so that

Shannon entropy and the structural entropy combining together characterise the information lost in data compression and the information recovered from a structural decoder.

Shannon Entropy vs Structural Entropy

- Metric
  Shannon: measure as a number, Structural Entropy: Measure as a number with an accompanying encoding tree that determines an encoding minimising uncertainty
- Objects
  Shannon: unstructured objects, Structural entropy: both structured and unstructured objects
- Dimensionality
  Shannon: One- dimensional, Structural entropy: High dimensional
- Measuring methods
  Shannon: Global, SE: Both local and global
- Role
  S: measuring uncertainty, SE: Simultaneously measuring uncertainty and decoding the essential structure

Shannon Entropy vs Structural Entropy- Big Picture

- Shannon Information Theory
  A mathematical theory of communication - point-to-point transmission
- Structural Information Theory
  A mathematical theory of information processing (IP)

Personalised Searching

A easy application is a personalised searching, remarkably better than the PageRank.

Theory

- Structural Information Theory
- Information Theoretical Approach to Graphs
- Information Optimization Theory
- Network Theory
- Coding Theory of Big Data

A New Theory for Information Processing

Provable Theory for Data Analysis

Principles of Structural Information Theory

- **Structural entropy minimisation** is the principle for data analysis
- **Encoding tree** optimizes data structure of massive data
- **Encoding** eliminates the uncertainty that is embedded in a complex system
- **Structural information decoding** is a general model for information processing
- **Information is generated by structures**, providing a new understanding of the notion of information

Open Questions

To develop the methods of the application of structural information theory in other information processing such as:

- Image analysis
- Natural language processing
- Object identification
- Dimensionality reduction
- Representation identification
- Learning algorithms

Structural Dimension

Definition

Let $G = (V, E)$ be an undirected and connected graph.

1. We define the structural dimension of G to be the least k such that

$$\mathcal{H}^k(G) = \mathcal{H}(G)$$

   We use $D(G)$ to denote the structural dimensional of G.

2. We define the upper structural dimension of G to be the greatest k such that

$$\mathcal{H}^k(G) = \mathcal{H}(G)$$

   We use $D^U(G)$ to the upper structural dimension of G.

Structural Dimension Problem

Does for every undirected and connected graph G, the following hold:

$$D(G) = D^U(G)?$$

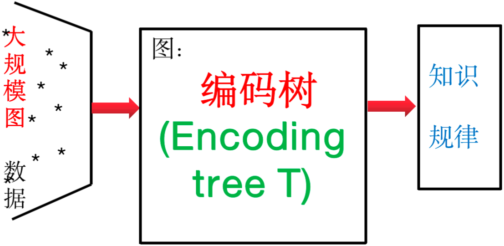
Information Optimization

For every undirected and connected graph G, and for every k, find an encoding tree T of G with height k such that

$$\mathcal{H}^T(G) = \mathcal{H}^k(G).$$

Questions:
(1) Is the k-dimensional optimum encoding tree of G unique?
(2) The complexity of the problem
(3) Polynomial time approximation and randomized algorithms

Information Optimization



$$\mathcal{H}(G) = \min_{T}\{\mathcal{H}^T(G)\}$$

- 限制不同类型的编码树得到不同的信息优化问题，是新的算法、随机算法、近似算法、局部算法、分布式算法与计算复杂性问题

Algebraic Understanding of the Structural Entropy

We have shown that for any undirected and connected graph G, if the two-dimensional structural entropy $\mathcal{H}^2(G)$ is small, then there is a large k such that the k-th largest eigenvalue of the Laplacian of G is small (less than ϵ).
This suggests that the structural entropy of graphs is closely related to the distribution of the eigenvalues of the Laplacian of graphs, leading to a new direction for graph theory.

Local Test of the Structural Entropy

Is there an algorithm to test for a graph G in time polylog n whether or not $\mathcal{H}^2(G)$ is $O(\log\log n)$ (or $\Omega(\log n)$), or far from the graphs with the properties?

Information vs Computation: New Algorithms

As mentioned before,

1. Encoding tree optimizes data structures
2. Encoding eliminates uncertainty

Therefore, encoding tree and in general, encoding is a resource of algorithms. There should exist a new

algorithmic theory using

- Encoding tree
- Encoding

Algorithmic Approach

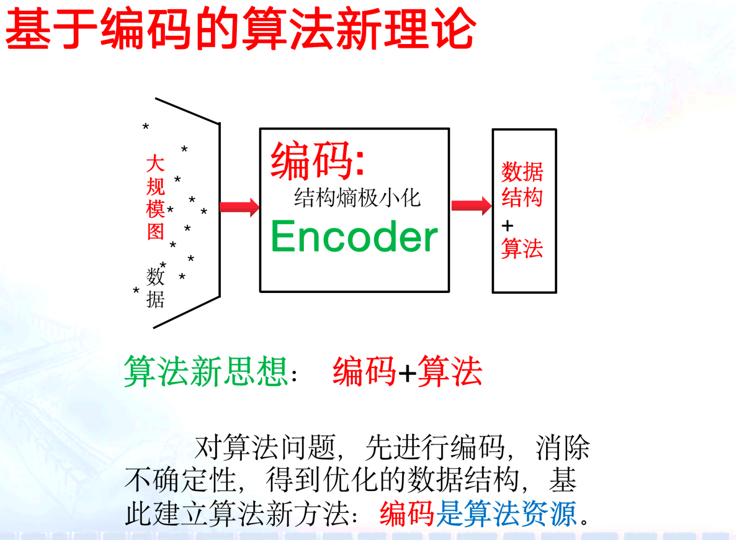


Figure: Algorithm

The Role of Structure in Information Theory

What we have learnt from the structural information theory are:

- Structure plays a key role in information theory
- Structural information finds the encoding tree that minimizes the uncertainty (or random variations) in a system
- However, randomness or noises play a key role in learning and game
  This is the difference between structural information and learning/game
- A natural question: Does structure play a role in AI and game?

Structural Information Theoretical Foundations of Artificial Intelligence

Observations:

1. Intelligence generates in the form of social groups
2. The structure of social groups plays a key role in the generation of intelligence
3. Some structures generate intelligence, but some others fail

Big challenge: **What is the structural generating theory of intelligence**?

Information vs Intelligence

Decoding information is the amount of uncertainty we can reduce, we are intelligent people.
Big challenge:
Is there an information theoretical definition of intelligence?
What is the relationship between information and intelligence?

Structural Information Theoretical Approach to Machine Learning

Statistical learning is perhaps the most successful part of machine learning. However, statistics works only on unstructured probability distributions.
If structure is key to learning, then there should be a structural learning theory.
Big challenge: **Is there a new structural learning theory?**

Game vs Information

Information must be the basis of game. We have shown that structure and randomness are key to the generation of information. The fundamental questions are hence:

1. What are the roles of structure and randomness in game?
2. What is the role of information in game
3. Is there an information theoretical direction of game theory?

References

1. A. Li, Y. Pan, Structural Information and Dynamical Complexity of Networks, IEEE Transactions on Information Theory, **62**, No. 6, pp. 3290 - 3339, 2016.
2. Brooks, F. P., Jr. Three great challenges for half-century-old computer science. Journal of the ACM, **50** (1), pp 25 - 26 (2003).
3. C. Shannon, The lattice of information, IEEE Transactions on Information Theory, **1**, No. 1, pp. 105 - 107, 1953.

Thank You!