Data Compression and IoT

Ulises Tirado Zatarain [1]
(ulises.tirado@cimat.mx)

[1] Algorists Group

July, 2016

Outline

What is data compression?

Definition (Data compression)

Remprestation of information using less space than original data. The action to compress data is called **compression** and the opposite actions is called **decompression**. Is a particular case of encoding/decoding information.

- Kinds of compression:
  - Loseless
  - Lossy

Loseless compression

- Information can be retrieved exactly as original data.
- Usually used to text compression
- Some known formats:
  - Zip
  - GZip
  - RAR
  - ACE
  - 7Zip
  - B2Zip
  - ...

Lossy compression

- Information loses some data, that cannot be retrieved exactly as before be compressed.
- Usually used to media compression: images, audio, video.
- Some known formats:
  - JPEG, GIF, PNG, ...
  - MP3, OGG, AAC, ...
  - H264, MPEG-4, VP8, ...

Kinds of information

- **Redundant**
  - Repetitive data
  - Predictable data

Kinds of information

- **Redundant**
  - Repetitive data
  - Predictable data
- **Irrelevant**
  - Invisible data
  - Removing this data don't affect message content

Kinds of information

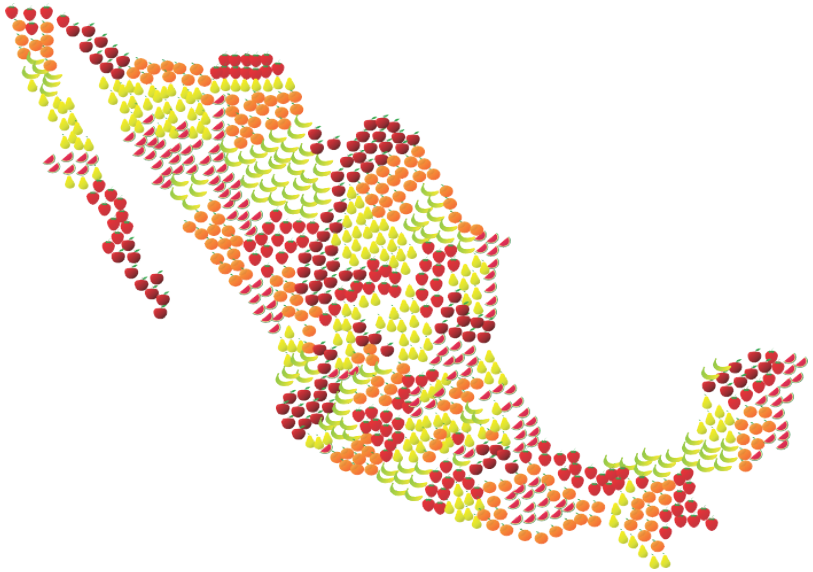
- **Redundant**
  - Repetitive data
  - Predictable data
- **Irrelevant**
  - Invisible data
  - Removing this data don't affect message content
- **Basic**
  - Essential data
  - It's needed to retrive original data
  - It should be trasnmitted

Lets see an example: Fruit 100% random & $\mathfrak{I}(\mathfrak{country})$

There are six popular fruits in an imaginary random country with some states (about 32). People in the country implements an elections system to know: What's the favorite fruit ever in this random, imaginary and 100% hipotetically country?

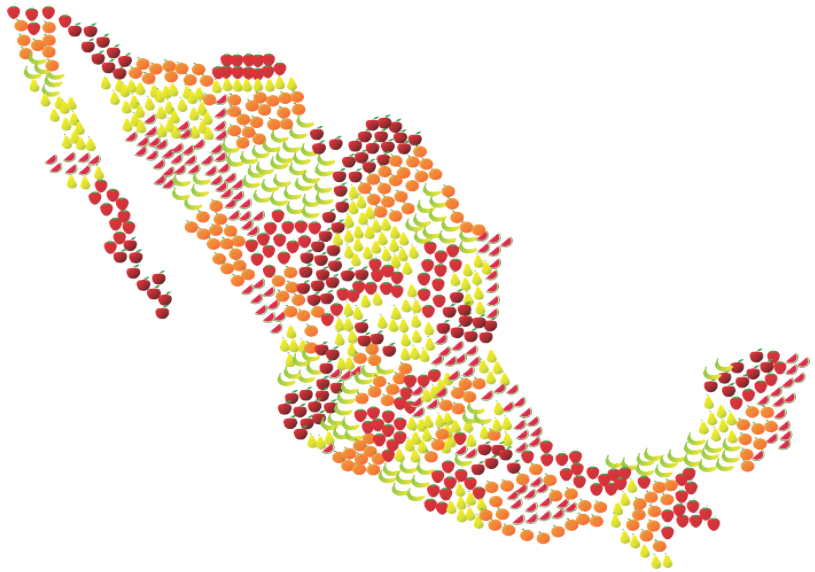
Lets see an example: Fruit 100% random & $\mathfrak{I}(\mathfrak{country})$

There are six popular fruits in an imaginary random country with some states (about 32). People in the country implements an elections system to know: What's the favorite fruit ever in this random, imaginary and 100% hipotetically country?



Lets see an example: Fruit 100% random & $\mathfrak{I}(\mathfrak{country})$

There are six popular fruits in an imaginary random country with some states (about 32). People in the country implements an elections system to know: What's the favorite fruit ever in this random, imaginary and 100% hipotetically country?



Can you see the different kinds of information?

Fruit 100% random & $\mathfrak{I}(\mathfrak{country})$: Game rules

The election system has following rules:

- Each citizen has an unique ID scanned from his/her ID card.

Fruit 100% random & $\mathfrak{I}(\mathfrak{country})$: Game rules

The election system has following rules:

- Each citizen has an unique ID scanned from his/her ID card.
- Each citizen can vote only once and only by one fruit.

Fruit 100% random & $\mathfrak{I}(\mathfrak{country})$: Game rules

The election system has following rules:

- Each citizen has an unique ID scanned from his/her ID card.
- Each citizen can vote only once and only by one fruit.
- If somebody tries to vote twice or more, then all votes from this citizen will be invalidated.

Fruit 100% random & $\mathfrak{I}(\mathfrak{country})$: Game rules

The election system has following rules:

- Each citizen has an unique ID scanned from his/her ID card.
- Each citizen can vote only once and only by one fruit.
- If somebody tries to vote twice or more, then all votes from this citizen will be invalidated.
- Any citizen can vote in any state.

Fruit 100% random & $\mathfrak{I}(\mathfrak{country})$: Game rules

The election system has following rules:

- Each citizen has an unique ID scanned from his/her ID card.
- Each citizen can vote only once and only by one fruit.
- If somebody tries to vote twice or more, then all votes from this citizen will be invalidated.
- Any citizen can vote in any state.
- There is a central system publishing partial live results.

Fruit 100% random & $\mathfrak{I}(\mathfrak{country})$: Game rules

The election system has following rules:

- Each citizen has an unique ID scanned from his/her ID card.
- Each citizen can vote only once and only by one fruit.
- If somebody tries to vote twice or more, then all votes from this citizen will be invalidated.
- Any citizen can vote in any state.
- There is a central system publishing partial live results.
- Each state has a system to votes counting and this reports to the central system. This systems only can report (to central system) votes from citizens who are natives from that state.

Fruit 100% random & $\mathfrak{I}(\mathrm{country})$: Game rules

The election system has following rules:

- Each citizen has an unique ID scanned from his/her ID card.
- Each citizen can vote only once and only by one fruit.
- If somebody tries to vote twice or more, then all votes from this citizen will be invalidated.
- Any citizen can vote in any state.
- There is a central system publishing partial live results.
- Each state has a system to votes counting and this reports to the central system. This systems only can report (to central system) votes from citizens who are natives from that state.
- In anytime the systems in each states can communicate with the other state systems to report votes from non-native citizens.

Logistics (brainstorming)

- How people can vote?

Logistics (brainstorming)

- How people can vote?
- Is an app needed?

Logistics (brainstorming)

- How people can vote?
- Is an app needed?
- How people can vote outside of their state? (non-native people)

Logistics (brainstorming)

- How people can vote?
- Is an app needed?
- How people can vote outside of their state? (non-native people)
- Infraestructure?

Architecture and design (brainstorming)

- First, think in the small case (i.e. one server by state)

Architecture and design (brainstorming)

- First, think in the small case (i.e. one server by state)
- Solve for this case

Architecture and design (brainstorming)

- First, think in the small case (i.e. one server by state)
- Solve for this case
- Improve to solve big case (i.e. dividing each states by districts)

What about data transferring? (brainstorming)

Do you have some ideas for the system?

- One vote a once?

What about data transferring? (brainstorming)

Do you have some ideas for the system?

- One vote a once?
- Several votes at once?

What about data transferring? (brainstorming)

Do you have some ideas for the system?

- One vote a once?
- Several votes at once?
- What technology can we use?
  - XML
  - JSON
  - Our own coding method?

Data transferring: XML? (brainstorming)

```
1  <?xml version="1.0" encoding="UTF-8" ?>
2  <!DOCTYPE FruitCountry SYSTEM "votes.dtd">
3  <state id="25">
4      <vote>
5          <citizen id="111999" />
6          <by>Apple</by>
7      </vote>
8       ...
9      <vote>
10         <citizen id="333777" />
11         <by>Strawberry</by>
12     </vote>
13 </state>
```

Data transferring: JSON? (brainstorming)

```
{
    state: 25,
    votes: [
        { citizen: 111999, by: 'Apple' },
        { citizen: 222888, by: 'Pear' },
        { citizen: 222888, by: 'Banana' },
        { citizen: 222888, by: 'Watermelon' },
        ...
        { citizen: 333777, by: 'Strawberry' }
        { citizen: 333777, by: 'Orange' }
    ]
}
```

Data transferring: Our own coding method? (brainstorming)

- What if we use some abreviations?
  - A: Apple
  - B: Banana
  - O: Orange
  - P: Pear
  - S: Strawberry
  - W: Watermelon

Data transferring: Our own coding method? (brainstorming)

- What if we use some abreviations?
  - A: Apple
  - B: Banana
  - O: Orange
  - P: Pear
  - S: Strawberry
  - W: Watermelon
- Do we really need to send the citizen ID?

Data transferring: Our own coding method? (brainstorming)

- What if we use some abreviations?
  - A: Apple
  - B: Banana
  - O: Orange
  - P: Pear
  - S: Strawberry
  - W: Watermelon
- Do we really need to send the citizen ID?
- Do we really need to send the state ID?

Data transferring: Our own coding method? (brainstorming)

- What if we use some abreviations?
  - A: Apple
  - B: Banana
  - O: Orange
  - P: Pear
  - S: Strawberry
  - W: Watermelon
- Do we really need to send the citizen ID?
- Do we really need to send the state ID?
- Fixed width messages?

Data transferring: Our own coding method? (brainstorming)

- What if we use some abreviations?
  - A: Apple
  - B: Banana
  - O: Orange
  - P: Pear
  - S: Strawberry
  - W: Watermelon
- Do we really need to send the citizen ID?
- Do we really need to send the state ID?
- Fixed width messages?
- A possible message from state to central system:

$$25 \quad AAAAPPPPPBBBBBWWWSSOOOOOAAA$$

Run-Length Encoding (basic idea)

RLE Algorithm

The idea is counting the times that each character appears consecutively. For example, for a string:

$$S = \mathrm{aaaabbbbbbbbbaaaaaabbbbbbccccccbb}$$

its compressed representation will be:

$$\tilde{S} = \mathrm{a4b8a5b6c5b2}$$

Run-Length Encoding (algorithm v1.0)

```
function char * compress(const char *input)begin
    char *str ← input;
    char *output ← new char;
    int length ← 0;
    while *str ≠ 0 do
        char x ← *str;
        push-back(output,x);
        int k ← 1;
        while x = *(++str) do k++;
        push-back(output,to-alpha(k));
        length ← length + k + 1;
    end
    return strlen(output) < length ? output : input;
end
```

Run-Length Encoding (inconvenients)

- What about decompression?

Run-Length Encoding (inconvenients)

- What about decompression? Very simple, duh!

Run-Length Encoding (inconvenients)

- What about decompression? Very simple, duh!
- What happens if we got a compressed (and ambiguos) string like following?

$$\tilde{S} = a3\mathbf{15}b3$$

Run-Length Encoding (inconvenients)

- What about decompression? Very simple, duh!
- What happens if we got a compressed (and ambiguos) string like following?

$$\tilde{S} = a315b3$$

- Maybe, original string was like:

$$S = aaa11111bbbb$$

Run-Length Encoding (inconvenients)

- What about decompression? Very simple, duh!
- What happens if we got a compressed (and ambiguos) string like following?

$$\tilde{S} = a315b3$$

- Maybe, original string was like:

$$S = aaa11111bbbb$$

- Or maybe was:

$$S = aaa \cdots aaabbbb$$

Run-Length Encoding (inconvenients)

- What about decompression? Very simple, duh!
- What happens if we got a compressed (and ambiguos) string like following?

$$\tilde{S} = a315b3$$

- Maybe, original string was like:

$$S = aaa11111bbbb$$

- Or maybe was:

$$S = aaa \cdots aaabbbb$$

- Is this algorithm effective with XML or JSON?

Run-Length Encoding (improved v2.0)

- Maybe, we can use a separator/delimiter character?

Run-Length Encoding (improved v2.0)

- Maybe, we can use a separator/delimiter character?
  - What character can we use to «,», «|», ...?

Run-Length Encoding (improved v2.0)

- Maybe, we can use a separator/delimiter character?
  - What character can we use to «,», «|»,...?
  - What if this character is in the original message?

$$S = |||,,, \Rightarrow \tilde{S} = |4,,3 \text{ or } \tilde{S} = |4|,3$$

Run-Length Encoding (improved v2.0)

- Maybe, we can use a separator/delimiter character?
  - What character can we use to «,», «|», ...?
  - What if this character is in the original message?

$$S = |||,,, \Rightarrow \tilde{S} = |4,,3 \text{ or } \tilde{S} = |4|,3$$

- What if we only have two kinds of characters?

Run-Length Encoding (improved v2.0)

- Maybe, we can use a separator/delimiter character?
  - What character can we use to «,», «|», ...?
  - What if this character is in the original message?

$$S = |||,,, \Rightarrow \tilde{S} = |4,,3 \text{ or } \tilde{S} = |4|,3$$

- What if we only have two kinds of characters?
  - Thinking in binary :)

Run-Length Encoding (improved v2.0)

- Maybe, we can use a separator/delimiter character?
  - What character can we use to «,», «|»,...?
  - What if this character is in the original message?

$$S = |||,,, \Rightarrow \tilde{S} = |4,,3 \text{ or } \tilde{S} = |4|,3$$

- What if we only have two kinds of characters?
  - Thinking in binary :) The Beattles - Let it «bit»! XD

Run-Length Encoding (improved v2.0)

- Maybe, we can use a separator/delimiter character?
  - What character can we use to «,», «|»,...?
  - What if this character is in the original message?

$$S = |||,,, \Rightarrow \tilde{S} = |4,,3 \text{ or } \tilde{S} = |4|,3$$

- What if we only have two kinds of characters?
  - Thinking in binary :) The Beattles - Let it «bit»! XD

$$S = e >< v = \quad 00100101 \quad 00111110 \quad 00111100 \quad 01110110$$
$$\tilde{S}_{0,3} = 01000101000100100101010111000110110010010001$$
$$\tilde{S}^2_{0,2} = 01,01,11,$$

Run-Length Encoding (improved v2.0)

```
function char * compress(const char *input)begin
    char *str ← input;
    char *output ← new char;
    int length ← 0;
    while *str ≠ 0 do
        char x ← *str;
        push-back(output,x);
        int k ← 1;
        while x = *(++str) do k++;
        push-back(output,to-alpha(k));
        length ← length + k + 1;
    end
    return strlen(output) < length ? output : input;
end
```

Run-Length Encoding (improved v3.0)

- The idea is...

Run-Length Encoding (improved v4.0 PRO)

- The idea is...

Be creative (v1.0)

- The idea is...

Be creative (v2.0)

- The idea is...

Be creative (v3.0)

- The idea is...

Having fun & tradding off! (v4.0)

- The idea is...

Probability

- The idea is...

Priority Queues

- The idea is...

Tries

- The idea is...

Huffman Encoding

- The idea is...

Statistical

- The idea is...

Principal Component Analysis

- The idea is...

Grayscale images

- The idea is...

Color images

- The idea is...

References I

- Standford University
- HackerRank
- Code Forces
- Code Chef
- Wikipedia