COSC 290 Discrete Structures

Lecture 3: Sequences and Functions

Prof. Michael Hay
Monday, Jan. 29, 2018
Colgate University

Plan for today

1. Frequent itemset mining: wrap up
2. Sequences
3. Cartesian product
4. Functions
5. Logistics

Frequent itemset mining: wrap up

Frequent itemset mining: slow version

First attempt at algorithm:

- Try every possible combination of items from I.
- For each combo, calculate its support.
- If support is high enough (above some threshold), add it to result.
- Output result.

Frequent itemset mining: slow version

Input: A list of transactions $T := t_1, \dots, t_n$, *minsup*
Output: The set of all frequent itemsets.

Let I be the set of all items that appear in any transaction t_i
$Result := \emptyset$ ▷ *Result* will store all frequent itemsets
for each candidate itemset $c \in \mathcal{P}(I)$ **do**
    $s := \sigma(c)$ ▷ compute *support* for c
    if $s \geq n \times minsup$ **then** ▷ itemset c is frequent
        $Result := Result \cup \{ c \}$ ▷ add c to *Result*
return *Result*

What is inefficient about this approach?

Poll: the set of all items

Suppose we have a collection of n transactions, $t_1, \dots, t_n$. Example:

$$t_1 = \{ \text{soy milk}, \text{coffee} \}$$
$$t_2 = \{ \text{milk}, \text{orange juice}, \text{cocoa puffs} \}$$
$$\dots$$
$$t_n = \{ \text{organic tofu}, \text{broccoli}, \text{coffee}, \text{soy milk} \}$$

Let I represent the set of all items purchased in at least one transaction. Which of the following is a correct definition for I?

A) $I := t_1 \cup t_2 \cup \cdots \cup t_n$
B) $I := t_1 + t_2 + \cdots + t_n$
C) $I := t_1 \cap t_2 \cap \cdots \cap t_n$
D) $I := \{ t_1, t_2, \dots, t_n \}$
E) None of the above / More than one of the above

Poll: frequent itemset

Suppose we have a collection of n transactions, $t_1, \dots, t_n$. Example:

$$t_1 = \{ \text{soy milk}, \text{coffee} \}$$
$$t_2 = \{ \text{milk}, \text{orange juice}, \text{cocoa puffs} \}$$
$$\dots$$
$$t_n = \{ \text{organic tofu}, \text{broccoli}, \text{coffee}, \text{soy milk} \}$$

Suppose that c is a frequent itemset. Consider this statement:

$$c \in \mathcal{P}(t_1 \cup t_2 \cup \cdots \cup t_n)$$

Choose the best answer:

A) This statement must be true.
B) This statement may be true.
C) This statement must be false.
D) This statement is not well defined.

Sequences

Sequences

A *sequence* is an ordered collection of objects. (Also known as lists or tuples.)

Notation: angle brackets (parentheses can be okay too).

Example

Geolocation: first number is degrees north of equator; second number is degrees west of prime meridian.

$$\langle 42.8267, -75.5439 \rangle \quad \text{Hamilton, NY}$$

Order matters:

$$\langle -75.5439, 42.8267 \rangle \quad \text{the edge of Antarctica}$$

Sets vs. Sequences

A *set* is...

- unordered
- does not have duplicate elements
- represented using curly braces
- example: MWF courses being taken this semester $\{$ CORE 151 D, COSC 290, Math 260 $\}$

A *sequence* is...

- ordered
- can have duplicate elements
- represented using angle brackets
- example: MWF course schedule $\langle$COSC 290, Math 260, CORE 151 D$\rangle$

Sequences and data types

In programming, we have many different *data types* that can all be thought of as special cases of sequences.

- A string is a sequence of characters.
- A vector is meant to represent a sequence of $\mathbb{R}$, but each real number is approximated by a float
- In Java, an `int[]` array is a sequence of integers in the interval $[-2^{31}, 2^{31} - 1]$
- In Python, a `list = ['a', 4.3, True]` is a sequence of Python objects.

Cartesian product

Cartesian product

The Cartesian product takes two sets and generates a set of ordered pairs (sequences of length two).

$$A \times B := \{ \langle a, b \rangle : a \in A \text{ and } b \in B \}$$

Ex: let $A := \{ \text{Colgate}, \text{'Gate} \}$ and $B := \{ \text{raiders}, \text{University}, \text{hockey} \}$.

$$\begin{aligned} A \times B = \{ & \langle \text{Colgate}, \text{raiders} \rangle, \\ & \langle \text{Colgate}, \text{University} \rangle, \\ & \langle \text{Colgate}, \text{hockey} \rangle, \\ & \langle \text{'Gate}, \text{raiders} \rangle, \\ & \langle \text{'Gate}, \text{University} \rangle, \\ & \langle \text{'Gate}, \text{hockey} \rangle \} \end{aligned}$$

More than one Cartesian product

The Cartesian product of sets A and B is defined as

$$A \times B := \{ \langle a, b \rangle : a \in A \text{ and } b \in B \}$$

Continuing example, if $A := \{ \text{Colgate}, \text{'Gate} \}$ and $B := \{ \text{raiders}, \text{University}, \text{hockey} \}$ and $C := \{ ! \}$.

What about

$$(A \times B) \times C = ? \qquad A \times (B \times C) = ?$$

If you look closely at the definition... $(A \times B) \times C$ produces a set of elements of the form $\langle \langle a, b \rangle, c \rangle$.

This is awkward. Instead, we define it as

$$A \times B \times C := \{ \langle a, b, c \rangle : a \in A \text{ and } b \in B \text{ and } c \in C \}$$

n-ary Cartesian product

For sets $A_1, A_2, \ldots, A_n$, the n-ary Cartesian product is defined as

$$A_1 \times A_2 \times \cdots \times A_n := \{ \langle a_1, a_2, \ldots, a_n \rangle : a_1 \in A_1 \text{ and } a_2 \in A_2 \ldots \text{ and } a_n \in A_n \}$$

If $A_1, \ldots, A_n$ are all the same set A, we can use this shorthand:

$$A^n := \underbrace{A \times A \times \cdots \times A}_{n \text{ times}}$$

Back to Example

Let $A := \{ \text{Colgate}, \text{'Gate} \}$ and $B := \{ \text{raiders}, \text{University}, \text{hockey} \}$ and $C := \{ ! \}$.

$$\begin{aligned} A \times B \times C = \{ & \langle \text{Colgate}, \text{raiders}, ! \rangle, \\ & \langle \text{Colgate}, \text{University}, ! \rangle, \\ & \langle \text{Colgate}, \text{hockey}, ! \rangle, \\ & \langle \text{'Gate}, \text{raiders}, ! \rangle, \\ & \langle \text{'Gate}, \text{University}, ! \rangle, \\ & \langle \text{'Gate}, \text{hockey}, ! \rangle \} \end{aligned}$$

$$\begin{aligned} A^2 = \{ & \langle \text{Colgate}, \text{Colgate} \rangle, \\ & \langle \text{Colgate}, \text{'Gate} \rangle, \\ & \langle \text{'Gate}, \text{Colgate} \rangle, \\ & \langle \text{'Gate}, \text{'Gate} \rangle \} \end{aligned}$$

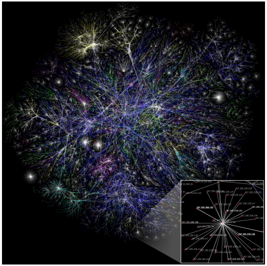


Figure 1: Partial map of the Internet based on the January 15, 2005 data found on opte.org. Each line is drawn between two nodes, representing two IP addresses. The length of the lines are indicative of the delay between those two nodes. By The Opte Project - Originally from the English Wikipedia; description page is/was here., CC BY 2.5, `https://commons.wikimedia.org/w/index.php?curid=1538544`

IP Addresses

Every device on the internet has an IP address. An address is a 32-bit number but is typically represented by four numbers separated by decimals where each number is in the range 0 to 255.

Example

The department website, `cs.colgate.edu`, has an IP address. Represented as a binary number, it is

`10010101 00101011 01010000 00001101`

Represented using the typical approach, it is: `149.43.80.13`

(Aside: above assumes IPv4. In IPv6, addresses are 128-bits and typically represented as 5 blocks of 4 hexadecimal numbers, such as `0:0:0:0:0:ffff:952b:500d`.)

Poll: IP Addresses

An IP address is a 32-bit number but is typically represented by four numbers separated by decimals where each number is in the range 0 to 255.

Example: the IP address `cs.colgate.edu` can be written as `10010101 00101011 01010000 00001101` or `149.43.80.13`

Question: Let S be the set of all valid IP addresses. Which of the following is an acceptable description of S?

A) $\{0, 1, \ldots, 255\} \times \{0, 1, \ldots, 255\} \times \{0, 1, \ldots, 255\} \times \{0, 1, \ldots, 255\}$

B) $\{0, 1, \ldots, 255\}^4$

C) $\{0, 1\}^{32}$

D) More than one of the above

E) None of the above

Paths

Let S represent the set of all IP addresses.

Devices on the internet are connected. Some are directly connected; others are connected through intermediate devices (called routers).

Indirectly connected nodes are connected by a path of direct connections. The *length* of the path is the number of direct connections along the path.

Example

Suppose device A is directly connected to B and B is directly connected to C. Then A and C are indirectly connected by the path (A,B,C). The length of this path is 2.

Poll: paths

Let S represent the set of all IP addresses.

The set of all *direct* connections is a subset of $S \times S$.

Which of the following is the best representation for the set of all possible paths having length at most 10?

A) $\mathcal{P}(S)$

B) S^{10}

C) $S \cup S^2 \cup S^3 \cup \cdots \cup S^{10}$

D) S^{11}

E) $S^2 \cup S^3 \cup \cdots \cup S^{10} \cup S^{11}$

Traceroute

You can explore these paths on your own!

Command line:

```
$ man traceroute
```

```
$ traceroute -m 10 -n www.auckland.ac.nz
```

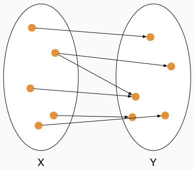
Website: `https://www.locaping.com/`

Traceroute overlaid with geography: `http://csvoss.scripts.mit.edu/traceroute/`

Functions

What is a function?

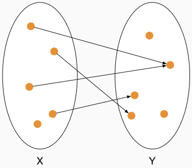
Is this a function from X to Y?



No. Some x is mapped to more than y.

What is a function?

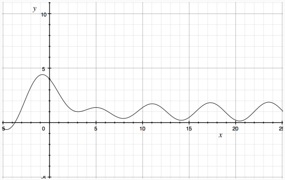
Is this a function from X to Y?



No. Some x is not mapped to any y.

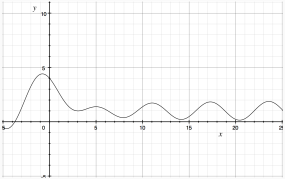
What is a function?

Is this a function from X to Y?



What is a function?

Is this a function from Y to X?



What is a function?

Let X and Y be sets.

A function f from X to Y, written $f : X \to Y$, assigns each input value $x \in X$ to a unique output value $y \in Y$.

We use $f(x)$ to denote the unique value from Y assigned to x by function f.

- For every element $x \in X$, $f(x)$ is always defined.
- Every element $x \in X$ is mapped to only one value in Y.

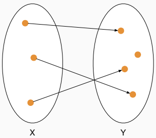
Terminology

- Domain: X
- Codomain: Y
- Image (aka range): the y values that correspond to function outputs.

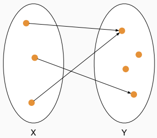
$$\{\, y \in Y : \text{there is some } x \in X \text{ where } f(x) = y \,\}$$

- Function composition: $(f \circ g)(x)$ is same as $f(g(x))$
- Onto, one-to-one, bijective functions
- Polynomials of degree d: $\sum_{k=0}^{d} a_k x^k$

One-to-one



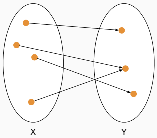
Yes.



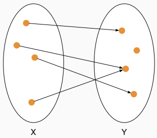
No.

One-to-one: for every $y \in Y$, there is at most one $x \in X$ such that $f(x) = y$.

Onto



Yes.



No.

Onto: for every $y \in Y$, there is at least one $x \in X$ such that $f(x) = y$.

Bijective

A function is bijective if it is both one-to-one and onto.

Poll: analyzing functions, part 0

Let $f : \{0, 1, 2, 3, 4, 5\} \to \{0, 1, 2, 3, 4, 5\}$ where

$$f(x) := 2x \bmod 6$$

This function is...

A) One-to-one
B) Onto
C) Bijective
D) None of the above

One-to-one: for every $y \in Y$, there is *at most one* $x \in X$ such that $f(x) = y$.

Onto: for every $y \in Y$, there is *at least one* $x \in X$ such that $f(x) = y$.

Poll: analyzing functions, part 1

Let $g : \{0, 1, 2, 3, \ldots, 9\} \to \{0, 1, 2, 3, 4\}$ where

$$g(x) := \lfloor x/2 \rfloor \bmod 5$$

This function is...

A) One-to-one
B) Onto
C) Bijective
D) None of the above

One-to-one: for every $y \in Y$, there is *at most one* $x \in X$ such that $f(x) = y$.

Onto: for every $y \in Y$, there is *at least one* $x \in X$ such that $f(x) = y$.

Poll: analyzing functions, part 2

Let $h : \{0, 1, 2, 3, 4\} \to \{0, 1, 2, 3, 4\}$ where

$$h(x) := 2x \bmod 5$$

This function is...

A) One-to-one
B) Onto
C) Bijective
D) None of the above

One-to-one: for every $y \in Y$, there is *at most one* $x \in X$ such that $f(x) = y$.

Onto: for every $y \in Y$, there is *at least one* $x \in X$ such that $f(x) = y$.

Logistics

Logistics

- Tea tomorrow: 11:30am, research lounge
- Problem set 1 – due Tuesday at 11:55pm
- Lab 1 – please read before W!