Clustering, classification, indexing

- Given
  - a set of objects in a database and
  - a given query

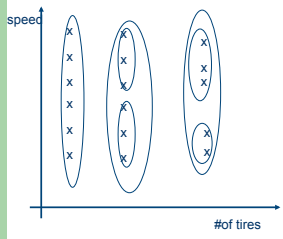
  how do we find the matching objects?

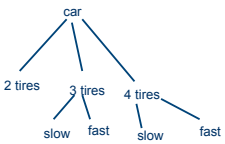
- clustering,
- classification,
- indexing

Class (Merriam-Webster Online Dictionary)

- Main Entry: **1class**
  Pronunciation: 'klas; Function: *noun*; Usage: *often attributive*
  Etymology: French *classe*, from Latin *classis* group called to military service, fleet, class; perhaps akin to Latin *calare* to call -- more at LOW
  1 a : a body of students meeting regularly to study the same subject **b :** the period during which such a body meets **c :** a course of instruction **d :** a body of students or alumni whose year of graduation is the same
  2 a : a group sharing the same economic or social status <the working *class*> **b :** social rank; *especially* **:** high social rank **c :** high quality **:** ELEGANCE
  3 : a group, set, or kind sharing common attributes: as **a :** a major category in biological taxonomy ranking above the order and below the phylum or division **b :** a collection of adjacent and discrete or continuous values of a random variable……

Classification of cars

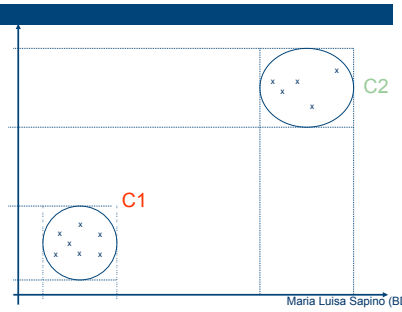




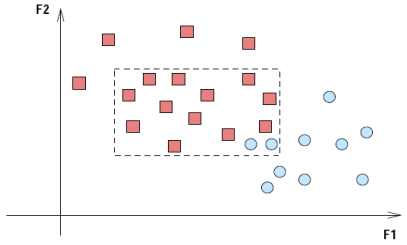
Cluster (Merriam-Webster Online Dictionary)

- Main Entry: **1clus·ter**
  Pronunciation: 'kl&s-t&r
  Function: *noun*
  Etymology: Middle English, from Old English *clyster;* akin to Old English *clott* clot
  : a number of similar individuals that occur together: as **a :** two or more consecutive consonants or vowels in a segment of speech **b :** a group of buildings and especially houses built close together on a sizable tract in order to preserve open spaces larger than the individual yard for common recreation **c :** an aggregation of stars or galaxies that appear close together in the sky and are gravitationally associated
  - **clus·tery** /-t(&-)rE/ *adjective*

Clustering



Clustering vs. classification

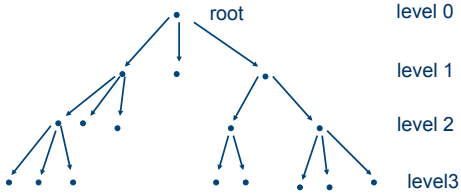


Index (Merriam-Webster Online Dictionary)

- Main Entry: **1in·dex**
  Pronunciation: 'in-"deks; Function: *noun*
  Inflected Form(s): *plural* **in·dex·es** *or* **in·di·ces** /-d&-"sEz/
  Etymology: Latin *indic-, index*, from *indicare* to indicate
  1 a : a device (as the pointer on a scale or the gnomon of a sundial) that serves to indicate a value or quantity **b :** something (as a physical feature or a mode of expression) that leads one to a particular fact or conclusion **:** INDICATION
  2 : a list (as of bibliographical information or citations to a body of literature) arranged usually in alphabetical order of some specified datum (as author, subject, or keyword): as **a :** a list of items (as topics or names) treated in a printed work that gives for each item the page number where it may be found…….....

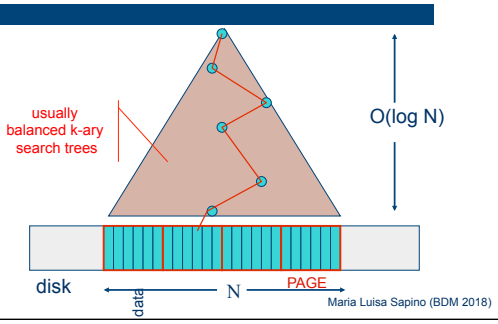
Index structures are usually trees

- Tree: connected directed graph where every node - except the root- has at most one incoming edge



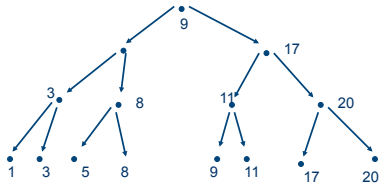
- Balanced tree: all leaves are at the same level
- k- ary tree: each node has at most k outgoing edges (children)
- Given a balanced k- ary tree with N nodes, the depth of the tree is $O(\log_k N)$

Index structures



Maria Luisa Sapino (BDM 2018)

Search tree



- It is hard to keep trees balanced due to updates: insertion/deletion
- The tree has to be RESTRUCTURED (if the tree is not balanced, search is O(N)

Maria Luisa Sapino (BDM 2018)

B-tree

- Is a balanced tree where each node (except the root) has at most n, and at least $\lceil n/2 \rceil$ children
- (leaves have 0 children)

Maria Luisa Sapino (BDM 2018)

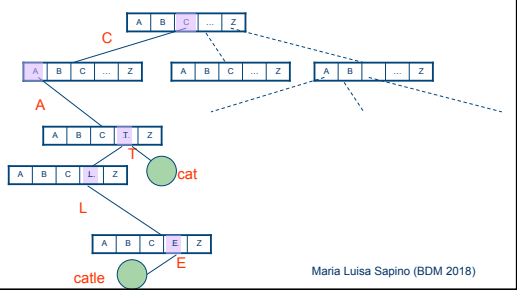
Keyword Search

- Full text scanning…

Keyword Search

- Inversion
  - For each document maintain a list of matching documents
    - B-tree
    - Hashing
    - trie

Trie (prefix)…



Suffix Trees and Arrays

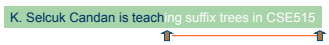
- Tries work well if the data consists of keywords…
- What if we do not have keywords?

Suffix Trees and Arrays

- Tries work well if the data consists of keywords…
- What if we do not have keywords?
- Suffix trees and suffix arrays
  - Input text: a single long string
  - each position in the text gives a suffix

Suffix Trees and Arrays

- Suffix trees and suffix arrays
  - Input text: a single long string
  - each position in the text gives a suffix



- Text of length of *N* gives *N* suffixes

Suffix Trees and Arrays

- Suffix trees and suffix arrays
  - Input text: a single long string
  - each position in the text gives a suffix

K. Selcuk Candan is teaching suffix trees in CSE515

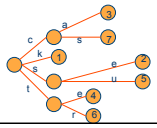
- Text of length of *N* gives *N* suffixes
- ..alternatively, text with W words give W suffixes,

Suffix Trees

- Suffix trees
  - Input text: a single long string
  - each word position in the text gives a suffix

K. Selcuk Candan is teaching suffix trees in CSE515

1 2 3 4 5 6 7

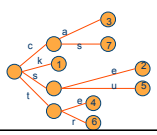


Suffix Trees

- Suffix trees
  - Input text: a single long string
  - each word position in the text gives a suffix

K. Selcuk Candan is teaching suffix trees in CSE510

1 2 3 4 5 6 7

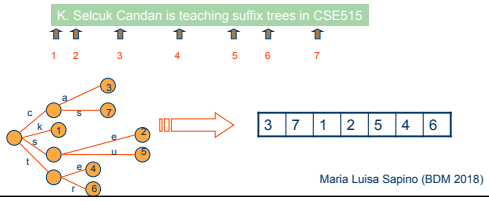


Patricia trie is a trie where all unary paths are compressed

This is also a Patricia trie

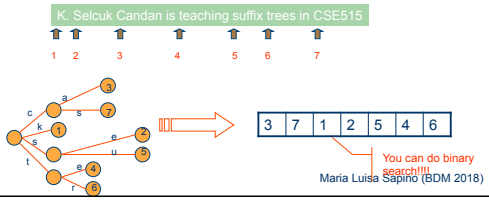
Suffix Arrays

- Suffix trees
  - Input text: a single long string
  - each word position in the text gives a suffix



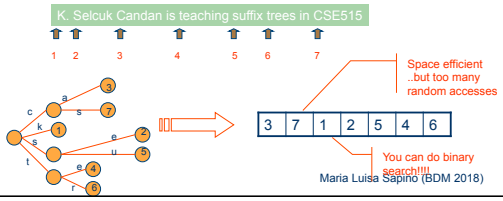
Suffix Arrays

- Suffix trees
  - Input text: a single long string
  - each word position in the text gives a suffix



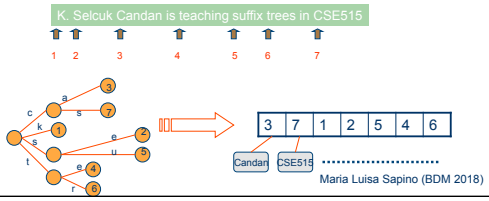
Suffix Arrays

- Suffix trees
  - Input text: a single long string
  - each word position in the text gives a suffix



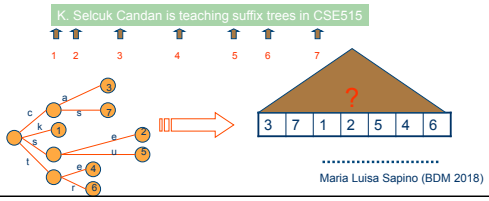
Suffix Arrays

- Suffix trees
  - Input text: a single long string
  - each word position in the text gives a suffix



Suffix Arrays

- Suffix trees
  - Input text: a single long string
  - each word position in the text gives a suffix



…no arrays..no suffixes??

- Can we do search without a data structure on text?

…no arrays..no suffixes??

- Can we do search without a data structure on text?
- Create a data structure on the query!!!!

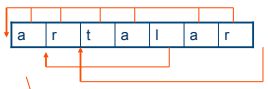
…no arrays..no suffixes??

- Given a text of length N....and pattern M
- Brute force: O(NM)

Average behavior closer to O(N)
-errors are found quick!!!

…no arrays..no suffixes??

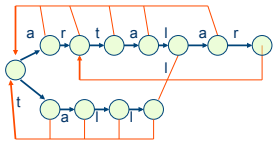
- Given a text of length N....and pattern M
- Knuth-Morris-Pratt: O(N)



A data structure
on the query!!!!

…no arrays..no suffixes??

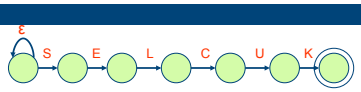
- Given a text of length N....and multiple patterns
- Aho-Corasick Trie: O(N)



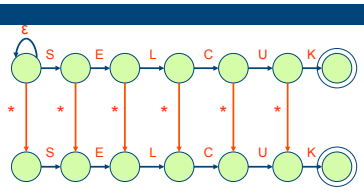
…no arrays..no suffixes??

- Given a text of length N....and pattern M
- Boyer-Moore: O(N(logM)/M)

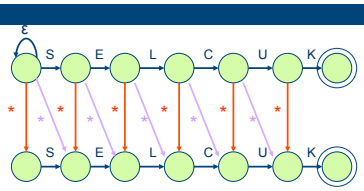
NFA…upto 0 errors



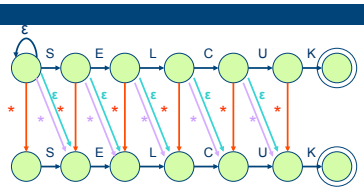
NFA…upto 1 insertion



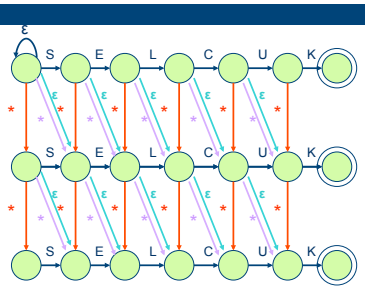
NFA…upto 1 insertion\replacement



NFA…upto 1 ins.\rep.\deletion



NFA…upto 2 ins.\rep.\deletion



Maria Luisa Sapino (BDM 2018)