Similarity Searching

Pavel Zezula
Vlastislav Dohnal
Michal Batko

Digital Data Explosion

- Everything we **write**, **see**, or **hear** can now be in a **digital** form**!!**
- Estimations:
  - 93% of produced data is digital
  - digital text is important – current technology is functional
  - multimedia, scientific, sensor, etc. is becoming **prevalent**

Searching & Computer Science

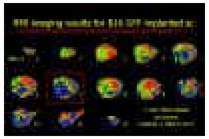
- One of the oldest and important data processing operations

- The problem is constrained by definitions of:
  - **where** to search ⇒ *domain (collection) of data*
  - **how** to search ⇒ *comparison criterion on objects*
  - **what** to retrieve ⇒ *query specification of data subsets*

Requirements of New Applications

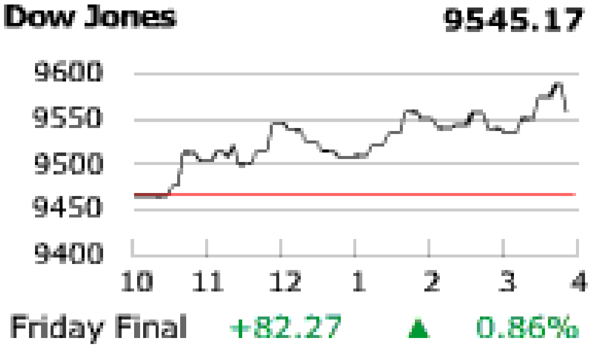
Medicine:

- *Magnetic Resonance Images (MRI)*



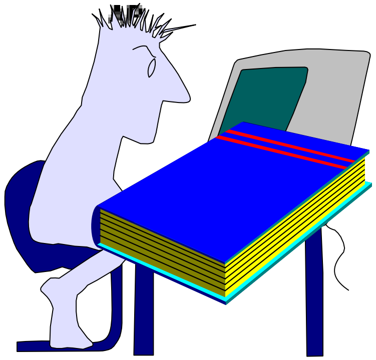
Finance:

- *stocks with similar time behavior*



Digital library:

- *text retrieval*
- *multimedia information retrieval*



Change of the Search Paradigm

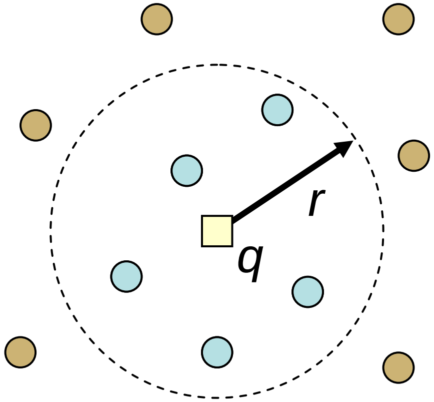
- Traditional YES-NO **keyword** search will not suffice - sortable domains of data (numbers, strings)

- New types of data need **gradual** comparison and/or ranking based on:
  - similarity,
  - dissimilarity,
  - proximity,
  - distance, closeness, etc.

Metric Space

- $\mathcal{M} = (\mathcal{D}, d)$
  - A data domain $\mathcal{D}$
  - A *total (distance) function d*: $\mathcal{D} \times \mathcal{D} \rightarrow$ 💻 (metric function or metric)
- The metric space postulates:
  - Non negativity $\forall x, y \in \mathcal{D}, d(x, y) \geq 0$
  - Symmetry $\forall x, y \in \mathcal{D}, d(x, y) = d(y, x)$
  - Identity $\forall x, y \in \mathcal{D}, x = y \Leftrightarrow d(x, y) = 0$
  - Triangle inequality $\forall x, y, z \in \mathcal{D}, d(x, z) \leq d(x, y) + d(y, z)$

Similarity Range Query

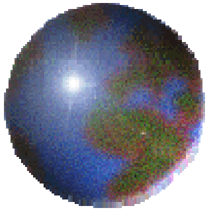


- A range query
  - $R(q,r) = \{ x \in X \mid d(q,x) \leq r \}$

… all museums up to 2km from my hotel …

Index Structures

- Centralized
  - M-tree, D-index
- Parallel
  - Parallel M-tree
- Distributed
  - M-Grid
  - GHT*, M-Chord



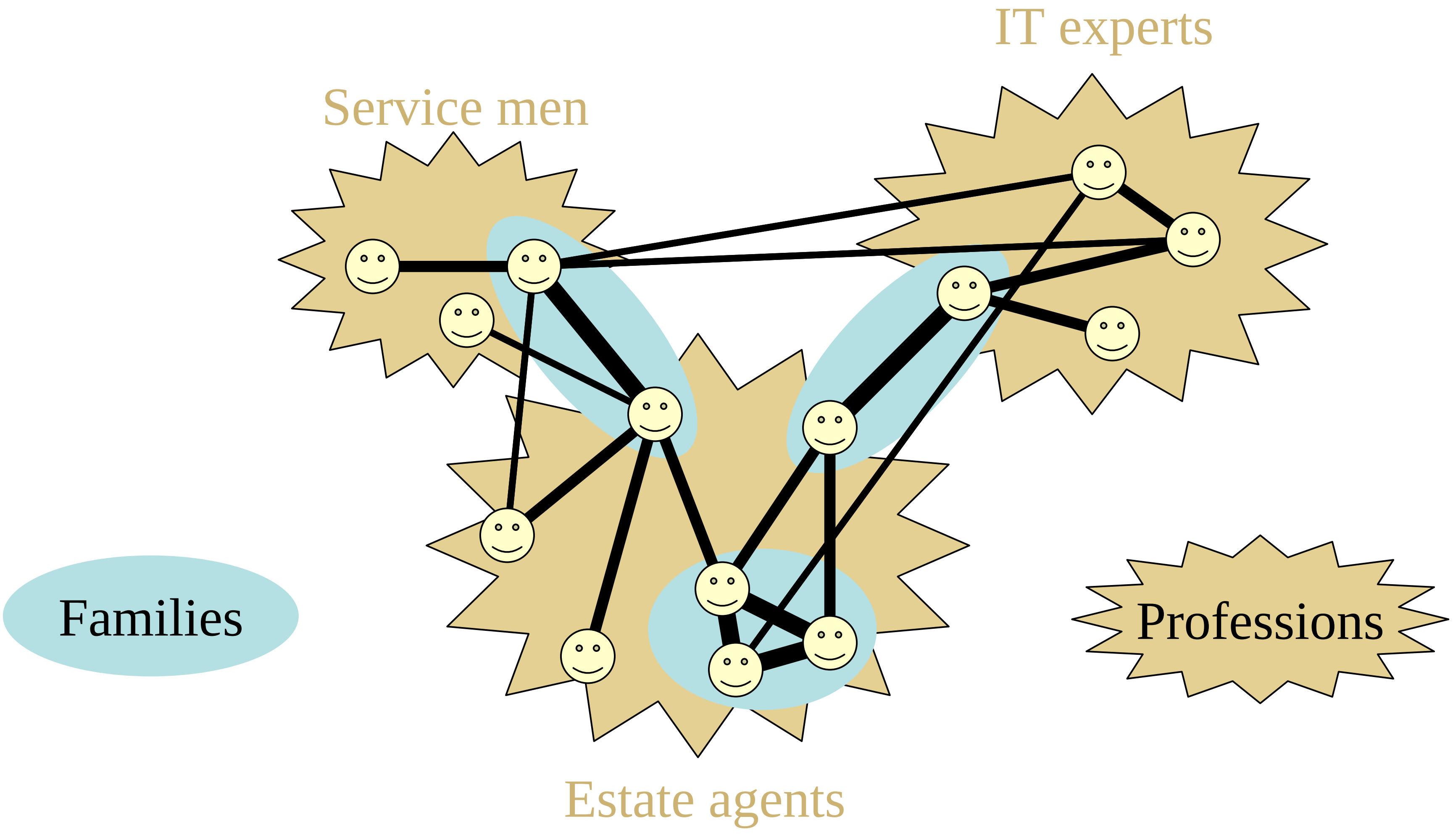
Metric Society

A new concept of indexing

Social Networks

- A social structure consisting of nodes
  - Individuals, organizations
- Connections (ties) between nodes
  - Social familiarities
    - Casual acquaintances
    - ...
    - Close family bounds

Social Networks



IT experts

Service men

Estate agents

Families

Professions

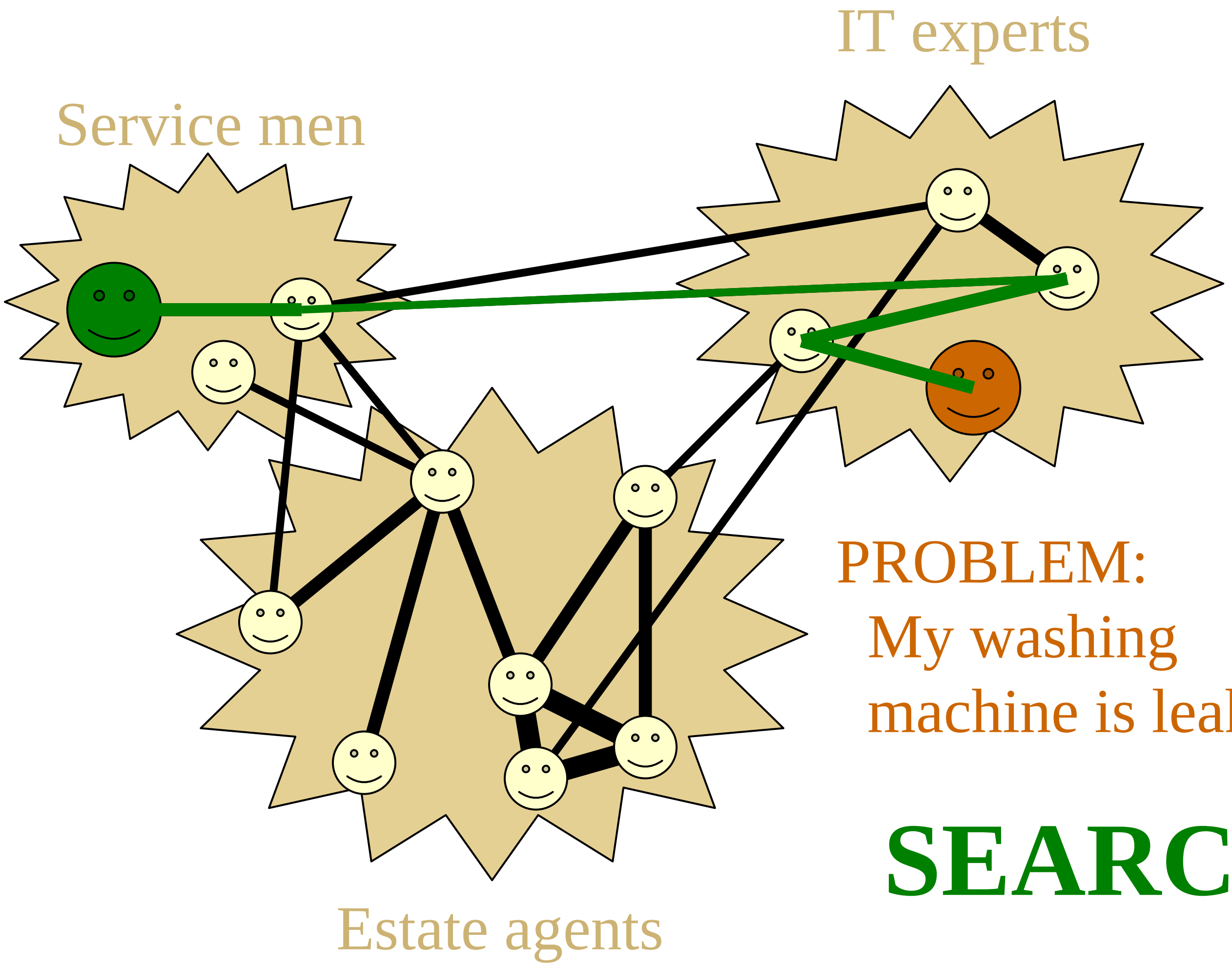
Social Networks

- Usefulness of the network
  - Search for help
    - Information, ...
  - Depends on the shape
    - Small/tight networks vs. lots of loose connections (weak ties)
- Different from structured P2P indexes
  - Nodes does not need to give up controlling their own data.
    - Nodes store data and others are allowed to search in it.

Social networks

- Attributes (data) of individuals
  - Determine a node’s participation in close (tight) relationships only
- Loose (weak) relationships
  - More important when searching
    - Because the group of friends who only do things with each other already share the same knowledge and opportunities.

Social Networks



PROBLEM:
My washing
machine is leaking

SEARCH

Social Networks

- Small World Problem
- Six Degrees of Separation
  - Stanley Milgram, 1967
  - 60 letters to various recruits in Omaha, Nebraska who were asked to forward the letter to a stockbroker living at a specified location in Sharon, Massachusetts.
  - Two random US citizens are connected on average by a chain of six acquaintances.
  - Completion ratio 5%

Social Networks

- The perceived value of the letter or parcel was a key factor in whether people were motivated to pass it on or not.
  - Later, researchers achieved as high as 97% completion.

- Most of the forwarding (i.e. connecting) was being done by a very small number of "stars" with significantly higher-than-average connectivity.

More Formal Model

- Searching for data using a social network

- Nodes
  - Stores data/information
  - Ask and answer queries
    - Using stored data/information
  - Forward queries to other nodes

More Formal Model

- Topology
  - Relationships between nodes
    - Tightness measurable (too complex)
  - Friends
    - Tight relationship
    - Nodes with similar data/information
  - Acquaintances
    - Loose relationship
    - Knowledge about the acquaintance’s domain

Searching in Social Network

- A query is posed to a node
  - Identify best experts
    - Deduced from previous answers
    - Similarity between question domains
      - Influences experts' relevance
- A query is either answered or forwarded
  - If a more relevant expert than me is known

Metric Society

- Use metric space similarity paradigms
  - To measure closeness between nodes
    - Friends
  - To measure relevance of query answers
    - Acquaintances
  - To measure similarity between queries
    - Routing algorithms

Friends

- The best friend of node P with respect to a given query $R(q,r)$ is a node P_{frd}
  - The similarity of their answers is high.

$$R_{(q,r)}(P) \approx R_{(q,\, r)}(P_{frd})$$

- Nodes maintain a list of friends
  - Updated by notifications from the query originator
    - The nodes that sent similar answers are notified that they are probably friends.

Acquaintances

- With respect to a query
  - Query = the node’s domain of expertise
- The best acquaintance for a given query $R(q,r)$ is a node P_{acq}
  - If the answer from P_{acq} is the most similar to the complete answer from all nodes.

$$R_{(q,r)}(P_{acq}) \approx R_{(q,\ r)}$$

Search Algorithm ***for known R(q,r)***

- Get the best acquaintance *A* for *R(q,r)*
  - One or more?
- Get the current node's result R_{cur} for *R(q,r)*
- Compare characteristics
  - Acquaintances, our result, our friends
- Forward to the node with the best characteristics of result
  - More than one, heuristics (friends/acq.)

Search Algorithm for unknown R(q,r)

- We must pick the best similar known query R(q,r)
- We may adjust characteristics
  - Estimate characteristics of the unknown query using known ones

Comparison Measurements

- Number of objects
  - Can be relative to full result set
- Query ball intersection overlaps
- Error on position
  - From approximate search
- Sum of absolute differences of histograms peaks
  - Can be weighted according to the distance from *q*
- Earth movers distance

Conclusion

- Implementation
  - Range search algorithm
  - Basic similarity measures
- Experiments
  - Reveal weaknesses
  - Show which similarities are unsatisfactory
  - Study recall / precision
  - Compare with “traditional” approaches