

A Self-organizing System for Large-scale Content-based Information Retrieval

Jan Sedmidubský

Faculty of Informatics, Masaryk University
Brno, Czech Republic



MEMICS
November 14–16, 2008

- Motivation
- Approaches
 - Metric Space
 - Self-organizing Systems
- Metric Social Network
 - Architecture
 - Query Routing
 - Experimental Trials
- Future Work

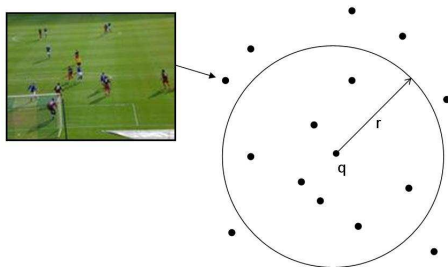
- **Our aim** – to develop an engine suitable for searching in large networks

- **Our aim** – to develop an engine suitable for searching in large networks

- **Problems:**
 - ① Domains of complex objects
 - Non-sortable
 - Similarity-based (content-based) reasoning \Rightarrow **Metric space**
 - ② Huge quantities of data
 - Exponential growth
 - Scalability problem \Rightarrow **Self-organizing systems**

Metric Space

- **Metric space** \mathcal{M} is a pair $\mathcal{M} = (\mathcal{D}, d)$, where:
 - \mathcal{D} is a set of objects – points in the metric space
 - d is a metric function measuring a distance (*similarity*) between two objects



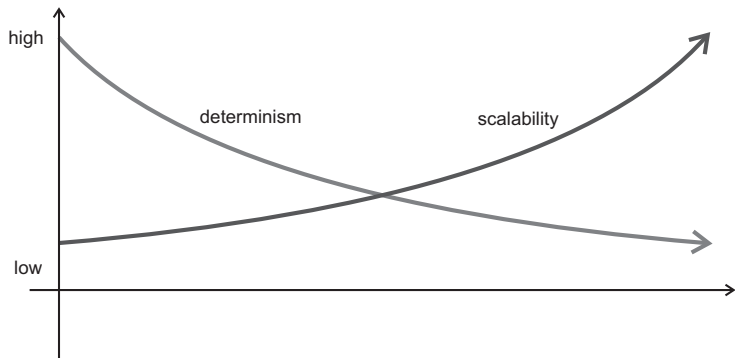
- Queries:
 - Range query $R(q, r)$
 - Nearest-neighbor query $NN(q)$

Search Problem

- **Scalability** – increasing amount of data, number of users (queries)

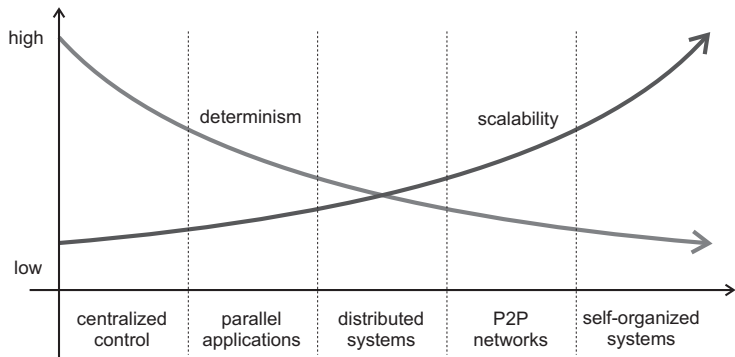
Search Problem

- **Scalability** – increasing amount of data, number of users (queries)
- **Determinism:**
 - exact match → similarity
 - precise answer → approximate answer
 - same answer → satisfactory answer



Search Problem

- **Scalability** – increasing amount of data, number of users (queries)
- **Determinism:**
 - exact match → similarity
 - precise answer → approximate answer
 - same answer → satisfactory answer



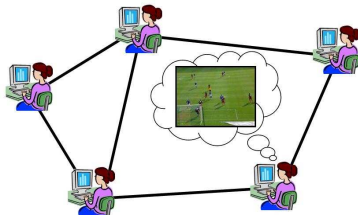
Self-organizing Systems

- A set of interacting components creating a desired outcome
 - Evolves in time and space
 - Inspired in biology, sociology, physics, . . .



Self-organizing Search Systems

- **Our aim** – apply principles of self-organization to build a robust search engine
 - A desired outcome = a search engine



- Properties:
 - Scalability
 - Adaptability
 - Robustness

Metric Social Network

Metric Social Network (mSN) =
= Metric space + Self-organization principles

Metric Social Network (mSN) =
= Metric space + Self-organization principles

- Self-organizing network for similarity searching
- Supports range queries (answers are approximate)
- Structure:
 - **Peers** – computers
 - **Relationships** – logical connections between peers
relationships \neq physical connections between peers

Peers

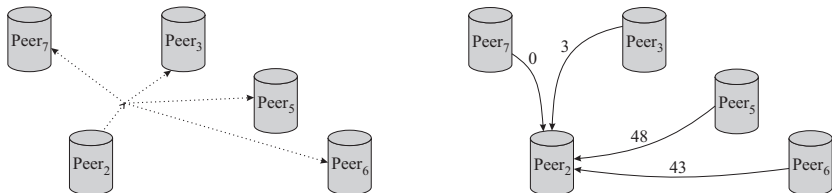
- Data – local data governed by the node, e.g. images
- List of random peers
- **Query history** – experience with previous querying

Relationships

- Exploited by the query-routing algorithm
- Based on the social-network paradigm:
 - Acquaintance relationships – navigation purposes
 - Friend relationships – identify similar peers

Relationships

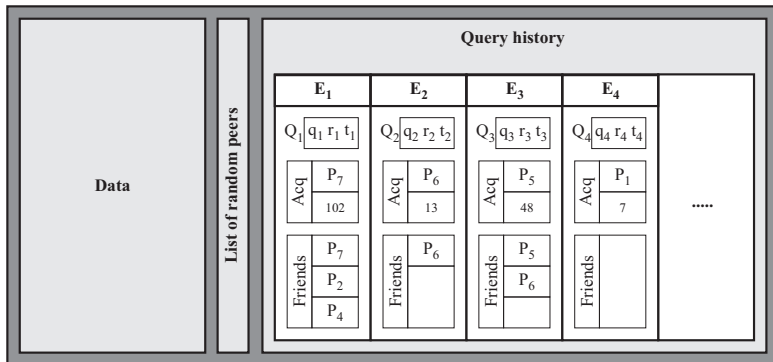
- Created according to **peers' answers to the processed query**



- Acquaintance** – peer with the best quality of the answer
 - $Acquaintance(Q) = Peer_5$
 - Acquaintance relationship: between $Peer_2$ and $Peer_5$
- Friends** – peers with the significant quality of the answer
 - $Friends(Q) = \{Peer_5, Peer_6\}$
 - Friend relationships: between each of two friends

Peer Anatomy

- **Query history** – a list of *entries* E_1, \dots, E_n containing metadata about queries processed so far
 - Query identification (query object, radius, timestamp)
 - Acquaintance
 - List of friends

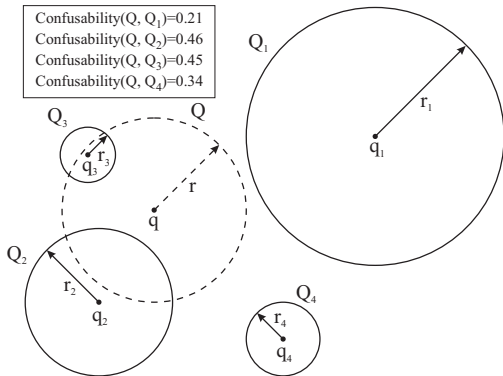


- User poses a query at the query-issuing peer
- Each peer selects the peers to which the query is forwarded
 - **Exploitation strategy**
 - Take the most relevant peers to the query
 - Confusability
 - **Exploration strategy**
 - Take some random peers
- The query is evaluated on local data of contacted peers
- The quality of answers is determined by the query-issuing peer
- The query history is updated

Confusability

- Confusability** expresses the similarity between two range queries Q_1 and Q_2 with timestamps

$$\text{Confusability}(Q_1, Q_2) = w_D \cdot D(Q_1, Q_2) + w_I \cdot I(Q_1, Q_2) + w_T \cdot T(Q_1, Q_2)$$



Q ₁	Q ₂	Q ₃	Q ₄
q ₁ r ₁ t ₁	q ₂ r ₂ t ₂	q ₃ r ₃ t ₃	q ₄ r ₄ t ₄
Acq Peer ₇ 102	Acq Peer ₆ 13	Acq Peer ₅ 29	Acq Peer ₁ 7
Friends Peer ₇ 102	Friends Peer ₆ 13	Friends Peer ₅ 29	Friends
Peer ₂ 88		Peer ₃ 18	
Peer ₄ 83			
			Friends

- 1 Retrieve five most confusable entries from the query history
- 2 Determine *max_confusability* of these five entries
- 3 Route the query
 - **Exploitation strategy**
 - Select acquaintances picked from n most confusable entries
 - The value of n depends on *max_confusability*

$max_confusability \geq$	0.90	0.65	0.40	0.15	0.00
n	1	2	3	4	5

- **Exploration strategy**
 - Select an arbitrary peer with a probability $1 - max_confusability$ from the list of random peers

Query Routing (cont.)

- Query forwarding stops when:
 - Maximum hop count is reached
 - More quality acquaintance does not exist
- The query is evaluated on local data of:
 - The most relevant acquaintances
 - Friends of the most relevant acquaintances respecting the query
- The answers are returned to the query-issuing peer
 - The qualities of answers are computed
 - New relationships are established

- Networks

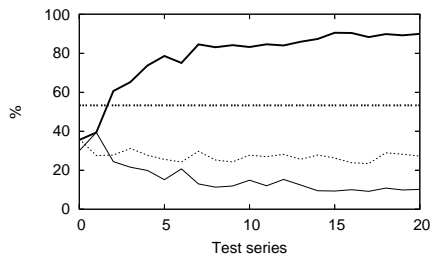
- ① 500 peers indexing 2,500,000 images
- ② 2,000 peers indexing 10,000,000 images

- Measures

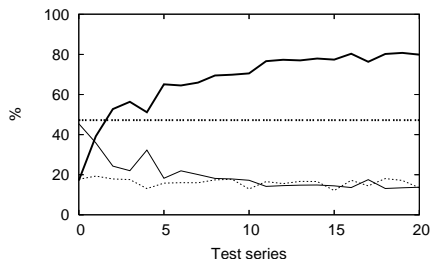
- **Costs** – a ratio between the number of accessed peers and the number of all peers in the network in percents
- **Recall** – a ratio between the sizes of mSN answer and the precise answer in percents
- **EOPP** – a normalized error on peers' positions expressing the inaccuracy of approximate answer of mSN

Experimental Trials (cont.)

- A batch of random 50 range queries between each of two test series
- Each test series consisted of fixed 20 range queries
- Results compared to *M-Chord* – a structured P2P network



mSN recall ——— mSN costs
mSN EOPP ——— M-Chord costs



mSN recall ——— mSN costs
mSN EOPP ——— M-Chord costs

- Dynamicity
 - Massive peers' churning
 - Joining two networks
- Knowledge management
 - Positive / negative feedback of querying

Thank you for your attention.

- Supported by:
 - **SAPIR** – the EU IST FP6 project 045128 – Search In Audio Visual Content Using Peer-to-peer Information Retrieval
 - **PhD grant** – the Integrated Approach to Education of PhD Students in the Area of Parallel and Distributed Systems 102/05/H050