# Ricerca e raccomandazioni di articoli scientifici in reti Peer-to-Peer

Leonardo D'Alonzo

Relatore

Chiar.mo prof. Ing Paolo Ciaccia

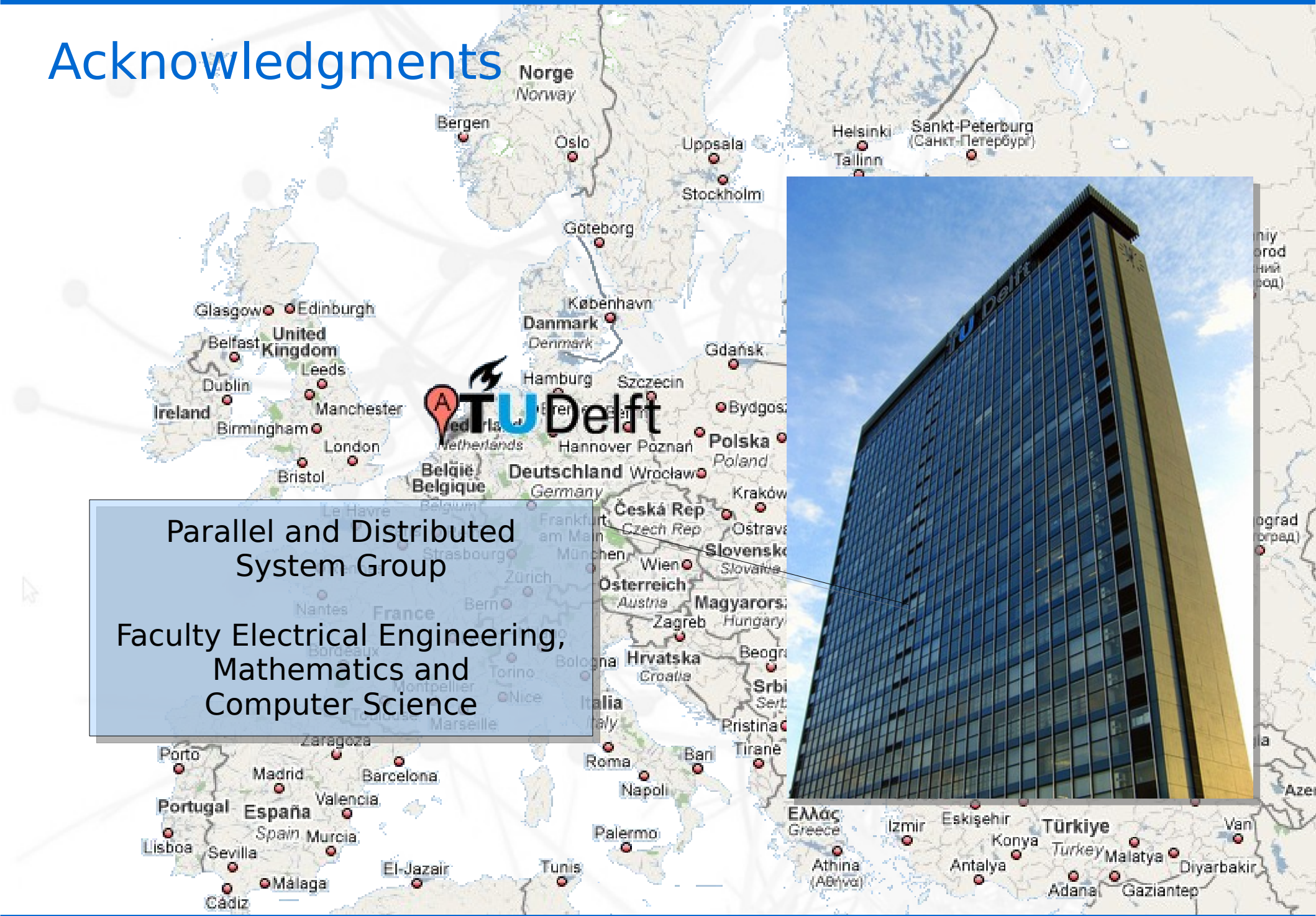
Correlatori

Dr. Ir. Johan Pouwelse

Dr. David Hales

Dr. Tamas Vinko

# Acknowledgments



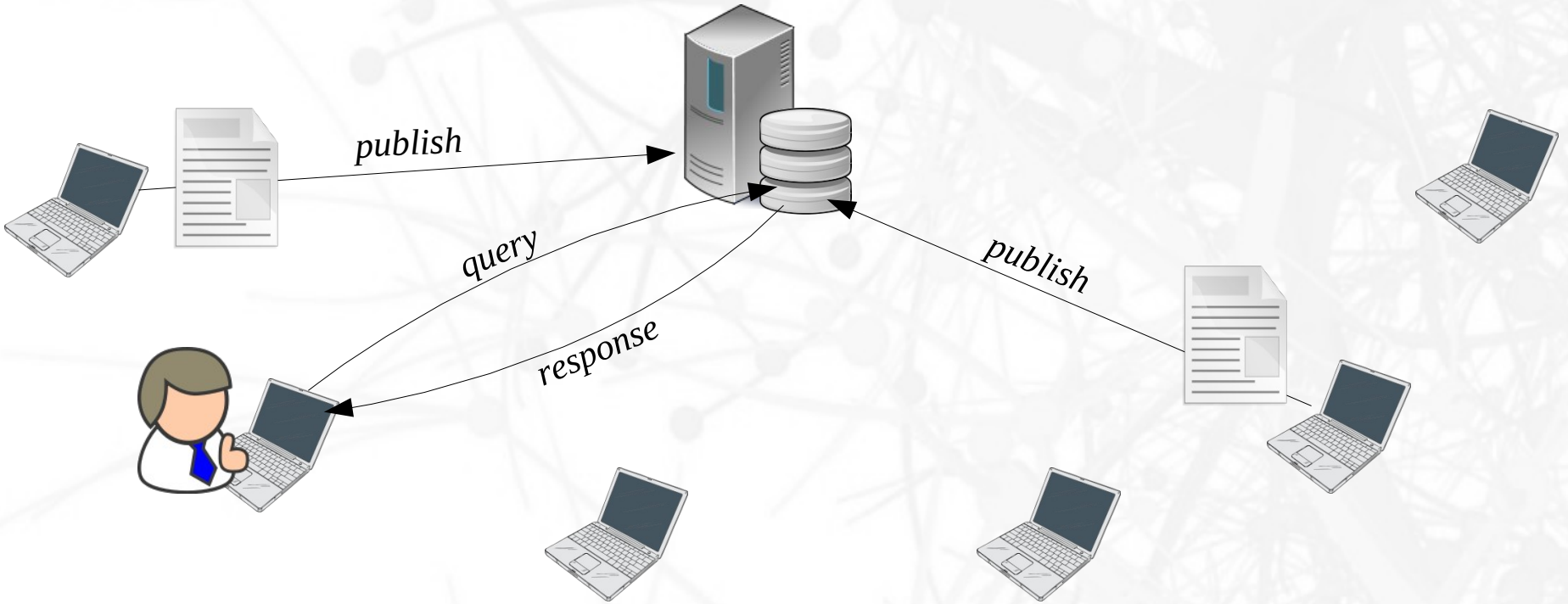
# Overview

- Motivation for a fully-decentralized approach to content location
- Approach to proposed solution
  - Underlying idea: exploiting network topology to ease content location
- A fully-decentralized search and recommender system
  - Aggregation of users with similar information needs
  - Search and recommendation service built on top of a social network
- Demo

# Motivation

Most of p2p file sharing systems rely on centralized information retrieval service

- Global knowledge of available contents
- High cost (scalability, dependability, maintenance)
- Prone to data exploitation



# Objective of this thesis

## Goal

Crafting the foundations of a fully-decentralized search and recommender system for text-based documents

Settings

- Highly-populated, highly-dynamic p2p networks
- Contents are provided by peers
- Each node has a partial view of the network
  - Overlay network
- Keyword-based access to information retrieval

Key concern

- Global knowledge of available contents requires high costs ☹

# Organizing network topology around data

Semantic proximity of information should be mapped into an overlay network topology in order to ease content location

## Underlying idea

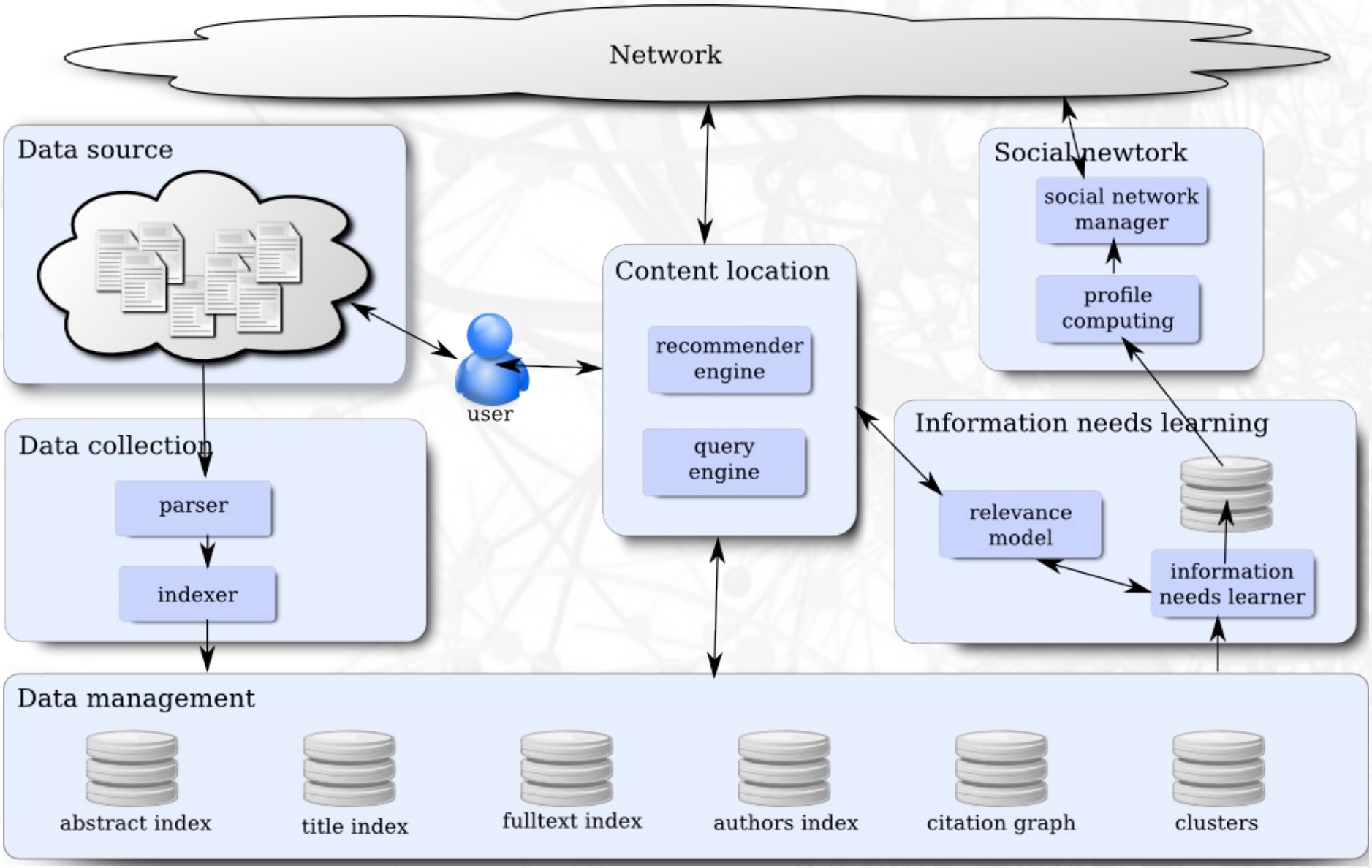
Agents of a population self-organize into a connection topology which reflects similarities between users information needs.

Each user is supported by an agent which learns his information needs

Key concerns

- How to capture user information needs?
  - They are dynamically changing
- How to manage network topology in a decentralized fashion?
  - Low-intrusion principle
  - Fast adaption to dynamics

# Agent architecture



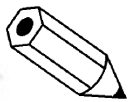
# Collecting data

Research papers which have been read by user are locally collected

- The agent builds a local database

Which data model should be used?

- a fine-grained model allows better definition of the relevance model ☺
- difficult of extracting structured information from visual layout documents ☹
  - machine learning algorithm as state-of-the-art



proof-of-concepts:

We designed a system which extracts title, abstract, fulltext, citations from pdf scholarly papers

- hybrid approach: rule based parser + machine learning algorithm 😐
  - fast prototyping ☺
  - difficult to adapt to new data ☹

# Vector Space Model

## An algebraic model for representing text

- A vocabulary of $t$ words is treated as the basis of a $\mathbb{R}^t$ vector space
- Combination of words (i.e. text) is turned into a vector

Semantics of text can be estimated by relying on statistical model of language

Tf-Idf weighting scheme $w_{ij} = \mathrm{tf}_{ij} \cdot \mathrm{idf}_i$

- Term frequency
  - local weight which depends on number of occurrences
- Inverse document frequency
  - global weight: discriminative power within a collection

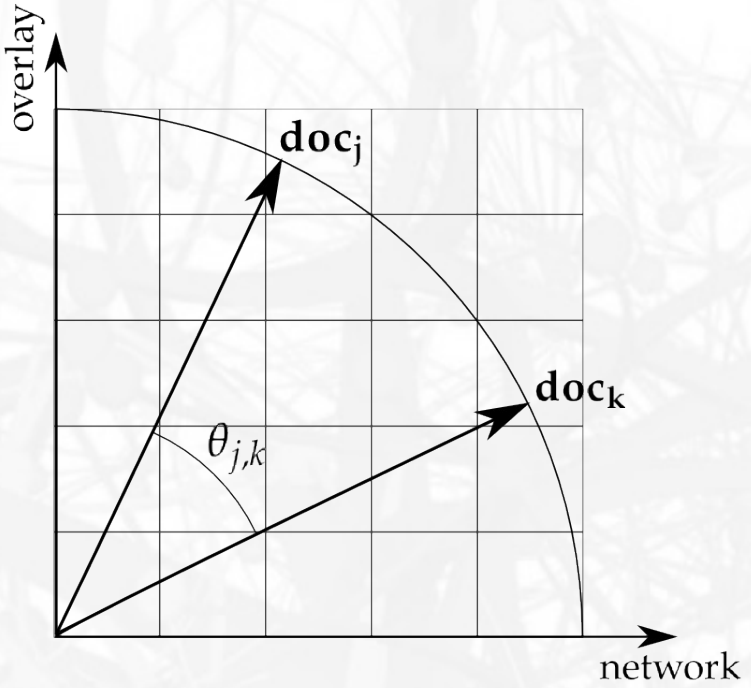
$$\mathbf{A} = \begin{array}{l} \\ \text{content} \\ \text{network} \\ \text{overlay} \\ \text{peer-to-peer} \\ \text{semantic} \end{array} \begin{array}{c} \begin{array}{cccccc} \mathrm{doc}_1 & \mathrm{doc}_2 & \mathrm{doc}_3 & \mathrm{doc}_4 & \mathrm{doc}_5 & \mathrm{doc}_6 \end{array} \\ \begin{pmatrix} 0 & 0.807 & 0 & 0 & 0.938 & 0 \\ 0.653 & 0 & 0.610 & 0.863 & 0 & 0 \\ 0.653 & 0.509 & 0.610 & 0 & 0 & 0 \\ 0 & 0 & 0.357 & 0.505 & 0.346 & 0.707 \\ 0.382 & 0.298 & 0.357 & 0 & 0 & 0.707 \end{pmatrix} \end{array}$$

# Computing similarities

## Similarity between documents

Computed as Euclidean distance between corresponding normalized vectors

- Cosine similarity




Criticism

- Document semantics relies on a lexicon based model
  - Inability to deal with natural language ambiguity
  - Latent Semantic Indexing offers a better model of document semantics by performing linear algebraic operations on the Vector Space Model

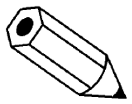
# Learning user information needs

## How to learn user information needs?

The agent should learn user information needs by relying on a model which takes into account a number of objective measurements of user interaction with the system:

- Semantics of documents which have been read
- Time spent while reading documents
- Tracking of issued queries
- ...

Model should be shaped by relying on a machine learning approach

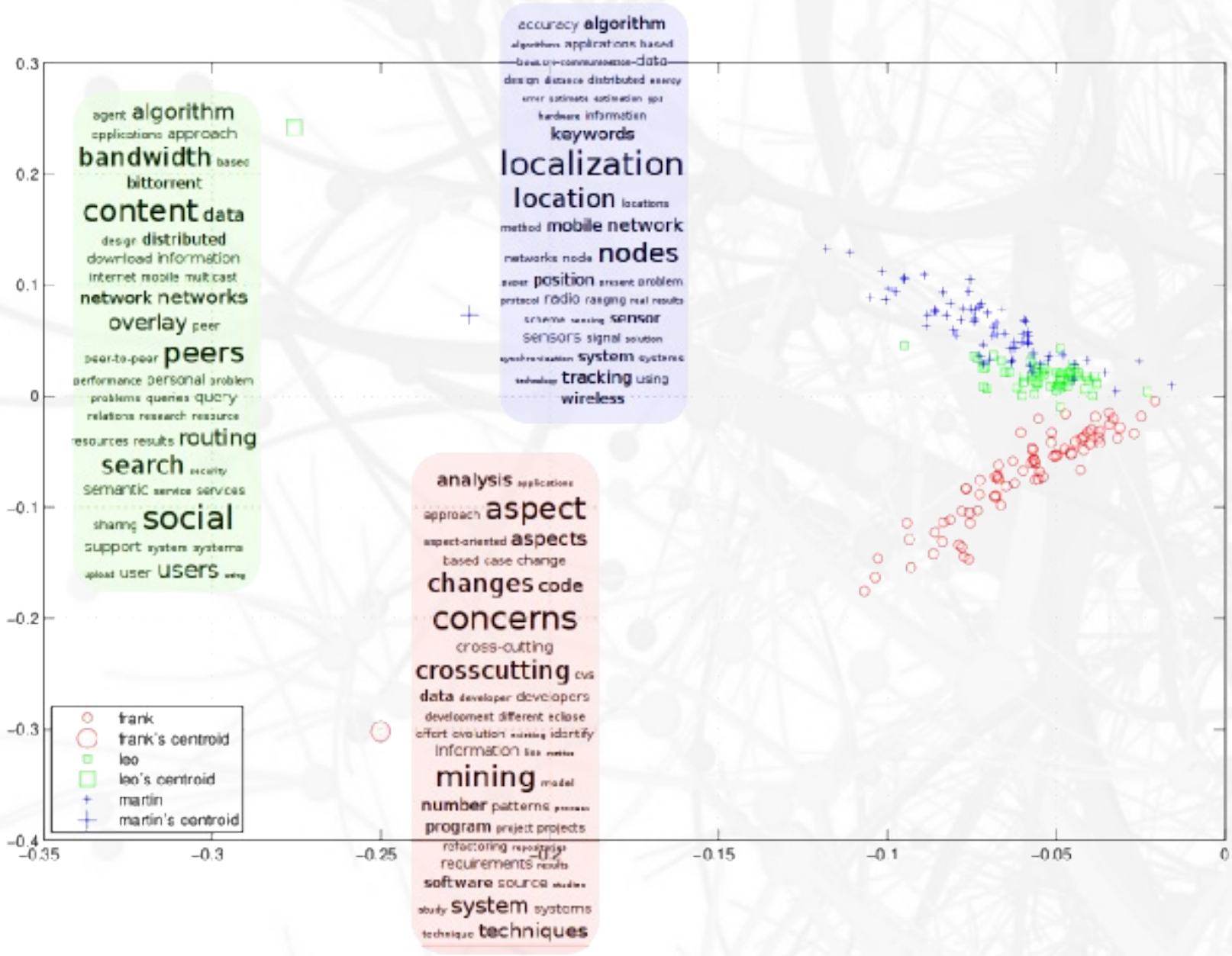


proof-of-concepts:

We rely just on user's collected documents

- User information needs are estimated with the centroid of document vectors computed according to vector space model

# Locality of interests principle



# Topology construction problem

Each agent has:

- a profile
- a ranking function which defines an order over nodes profile
- a partial view of size $c$ of the network

## Topology management as a membership service middleware

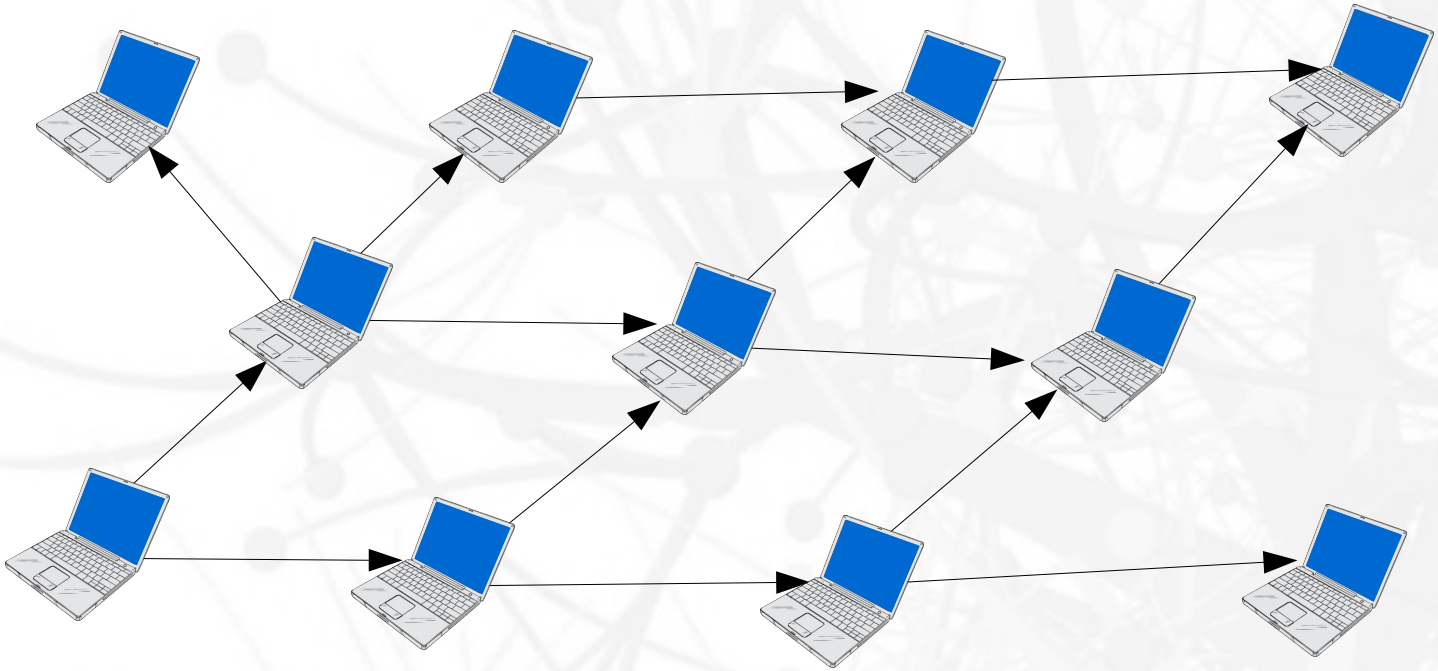
The goal is to build the views of each node such that each view contains the first $c$ elements according to the order defined by the ranking function

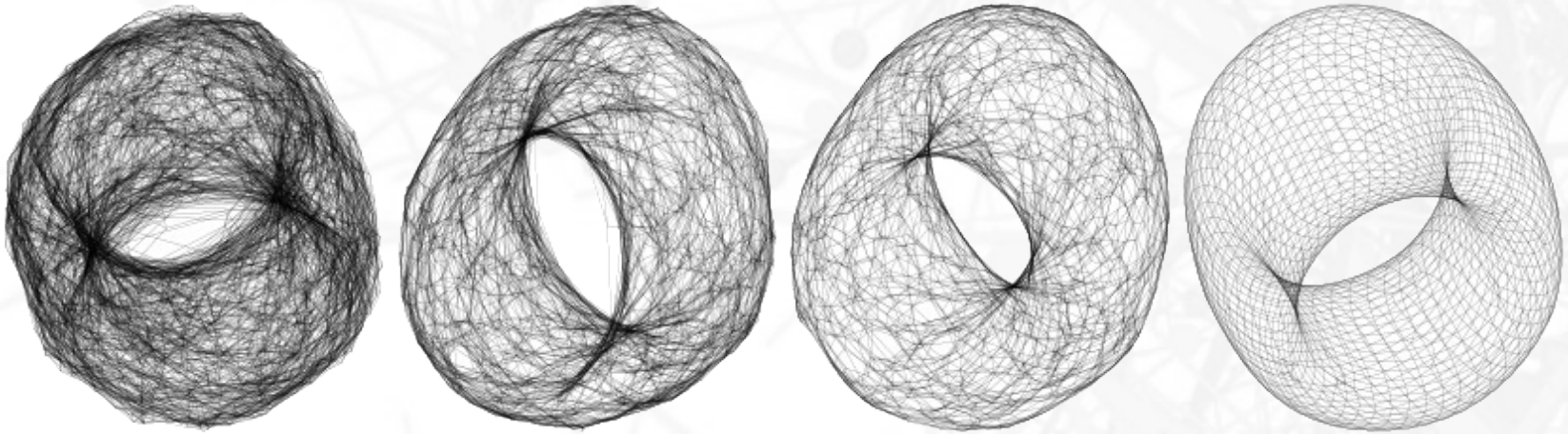
How to manage topology in highly-dynamic, highly-populated network with the minimum intrusion and without the need of a global view?

## Gossip protocol: probabilistic multicast scheme

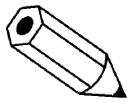
- Model to spread information among a large number of processes with dynamic collection topology
- Robust and scalable even in presence of high rate failures 

# Gossip-based topology management





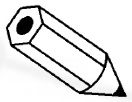
# Aggregating users with similar interests



## User profile

In order to keep light the protocol we rely upon an approximation of user documents centroid

- we retain just the most 30-weighted terms according to td-idf



## Ranking function

Users information needs similarity is computed by cosine similarity between corresponding approximated documents centroids

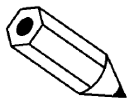
- Documents centroids must refer to the same basis
  - Terms are spread together with their weights

## Social network of researchers

- Agents aggregates users with similar information needs
- Friendships emerge around data

# Searching on top of a social network

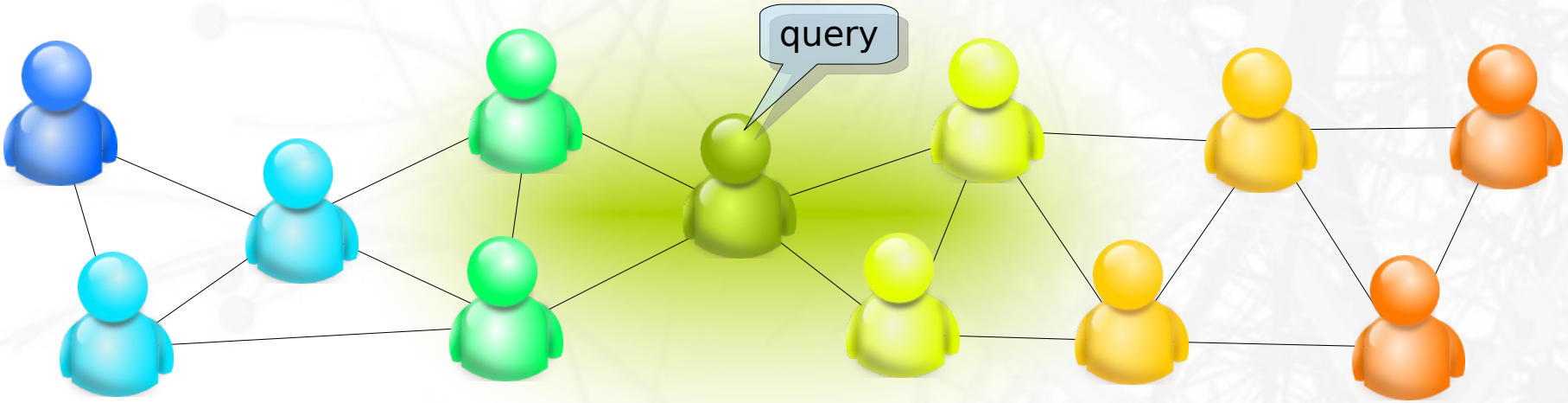
Network topology facilitates content location



## Querying the system

User queries are routed by exploiting the network topology

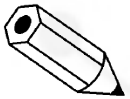
- Relevant information is expected to be located in the user neighborhood
  - We limit the search radius to the first connection level
- Effectiveness of the model resides in the ability of estimating user information needs and in his locality of interests



# Recommendation as collaborative filtering

## Collaborative filtering

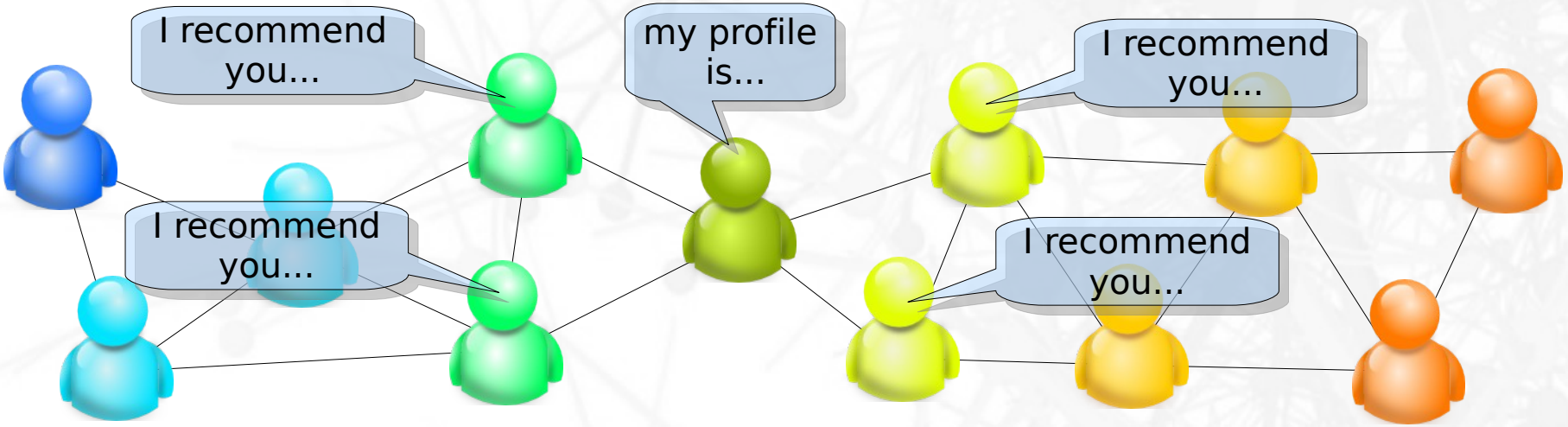
Agents adapt available information to each user information needs



### Recommendation model

User's neighborhood recommends their contents which are most similar to user information needs

- We rely on a linear combination of cosine similarity and link analysis



# Conclusions

Centralized information retrieval systems requires sizeable capital investments and are prone to data exploitation

## Contribution

We designed a proof-of-concept prototype of a fully-decentralized search and recommender system for scientific literature

- Search and recommendation as a service developed on top of a proactively managed social network
  - Aggregation of users with similar interests
- user interests are automatically learned by tracking his reading habits

Future work should improve:

- relevance model
- user information needs estimation
- agent profile and ranking function in order to achieve desirable clustering of information

# Demo and Question time

Hi **prof_Bassi**. Enter your query...

## Search on your local index

Find articles with **at least** one of the words [optical] in [abstract] Search!
Find articles **written by** [] Search!

**12** documents have been indexed.
Cluster centroid **RSS** = 10.87

achieved analysis bandgap bandgaps bandpass bandwidth based control/analysis coupling crystal crystals demonstrated designing device devices different dual-frequency effects features fiber fibers filled filter filtering frequency grating guided highly integrated lc light liquid molecules nm nonlinear novel optical photonic platform polarization poled power present robust spectral stop-band temperature tunability tunable tuning

## P2P search

Find articles with **at least** one of the words [] in [title] Search!

You have **2** taste buddies:

| Name | IP address | Similarity |
|---|---|---|
| **prof_Diversi** | @localhost:16205 | **5.4099**% |
| **prof_Sartori** | @localhost:16201 | **1.6042**% |

Buddycast cluster centroid approximation = **63.458**% of the norm

## Recommendations

**prof_Diversi** recommends you:

### A Dual Filtering Approach in MEMS based Dynamic Attitude Estimation

Roberto Guidorzi , Roberto Diversi , Umberto Soverini

VSM similarity ~42% (0.429), PageRank ~8% (0.0833)

*Abstract:* [536] - The problem considered in this paper is the design of a low cost MEMS based attitude estimation unit to be used in ultralight, experimental and sport pilot aircrafts as auxiliary safety tool in VFR flight conditions. The proposed approach relies on a new data fusion scheme based on a dual Kalman filter design and on

Hi **prof_Ciaccia**. Enter your query...

## Search on your local index

Find articles with **at least** one of the words [ ] in [title] Search!
Find articles **written by** [ ] Search!

**33** documents have been indexed.
Cluster centroid **RSS** = 43.49

## P2P search

Find articles with **at least** one of the words [ ] in [title] Search!

You have **2** taste buddies:

| Name | IP address | Similarity |
|---|---|---|
| **prof_Corradi** | @localhost:16204 | **4.1283**% |
| **prof_Sartori** | @localhost:16201 | **10.2813**% |

Buddycast cluster centroid approximation = **58.108**% of the norm

## Recommendations

**prof_Corradi** recommends you:

**Integrating Mobile Agent Infrastructures with CORBA-based Distributed Multimedia Applications**
Paolo Bellavista , Antonio Corradi , Domenico Cotroneo , Stefano Russo
VSM similarity ~29% (0.294), PageRank ~2% (0.0203)
*Abstract:* [898] - The increased computing power and the enhanced connectivity of current open computing systems are encouraging the deployment of new classes of services both centered around dynamically changing

approach **browsing** ceteris complex cp-nets data database db distance distributed dt dtw efficient features **image** images information integration issues management objects order paribus partial pattern **patterns** personalization phase pibe preference preferences preliminary present **queries** query represent results retrieval semantics set **similarity** skyline system systems techniques terms **time** user using warp

Hi **prof_Corradi**. Enter your query...

## Search on your local index

Find articles with **at least** one of the words [ ] in [title] Search!
Find articles **written by** [ ] Search!

**22** documents have been indexed.
Cluster centroid **RSS** = 29.76

## P2P search

Find articles with **at least** one of the words [ ] in [title] Search!

You have **2** taste buddies:

| Name | IP address | Similarity |
|---|---|---|
| **prof_Sartori** | @localhost:16201 | **5.9100**% |
| **prof_Verdone** | @localhost:16206 | **11.9920**% |

Buddycast cluster centroid approximation = **49.627**% of the norm

## Recommendations

**prof_Sartori** recommends you:

**Description Logics for Semantic Query Optimization in Object-Oriented Database Systems**
Domenico Beneventano , Sonia Bergamaschi , Claudio Sartori
VSM similarity ~31% (0.316), PageRank ~3% (0.0362)

access agent anomaly application-level architecture area awareness based capabilities client clients connectivity consumption context continuity environment facility hand handoff interfaces internet interoperability ma management mesis middleware mobile mobility multimedia network node nodes personalized provide provisioning proxies quality requirements resources semantic sensor service services streaming suitable support systems ubiquity wi-fi wireless

Hi **prof_Diversi**. Enter your query...

## Search on your local index

Find articles with **at least** one of the words [ ] in [title] Search!
Find articles **written by** [ ] Search!

**12** documents have been indexed.
Cluster centroid **RSS** = 14.94

## P2P search

Find articles with **at least** one of the words [ ] in [title] Search!

You have **2** taste buddies:

| Name | IP address | Similarity |
|---|---|---|
| **prof_Sartori** | @localhost:16201 | **11.5104**% |
| **prof_Verdone** | @localhost:16206 | **8.4270**% |

Buddycast cluster centroid approximation = **59.482**% of the norm

## Recommendations

**prof_Sartori** recommends you:

### Relevant Values: New Metadata To Provide Insight On Attribute Values At Schema Level

Sonia Bergamaschi , Mirko Orsini , Francesco Guerra , Claudio Sartori

VSM similarity ~33% (0.337), PageRank ~3% (0.0362)

*Abstract:* [988] - Research on data integration has provided languages and systems able to guarantee an integrated intensional representation of a given set of data sources. A significant limitation common to most

additive algorithm allows amounts approaches ar ararx autoregressive based case channels design different disturbance dynamic enhancement estimate estimation filter filtering frisch identification input kalman means method methods minimal model models new observations optimal output parameters performance presence procedure procedures results scheme signal smoothing speech system theoretical unknown variables variance variances

Hi **prof_Sartori**. Enter your query...

## Search on your local index

Find articles with **at least** one of the words [ ] in [title] Search!
Find articles **written by** [ ] Search!

**11** documents have been indexed.
Cluster centroid **RSS** = 5.19

## P2P search

Find articles with **at least** one of the words [ ] in [title] Search!

You have **2** taste buddies:

| Name | IP address | Similarity |
|---|---|---|
| **prof_Diversi** | @localhost:16205 | **11.5104**% |
| **prof_Verdone** | @localhost:16206 | **14.5397**% |

Buddycast cluster centroid approximation = **71.092**% of the norm

able allowing attribute based channel component data days detection development different discovery domain effective enrich entire events exceptional framework fully integration intensional just knowledge mining model networks news newspapers operations paper presentation problem process propose published query related relevant resulting results routing sensor sources strategies technique tool user values visual

## Recommendations

**prof_Diversi** recommends you:

### Residual Generation And Disturbance De-Coupling For A Chemical Process

Roberto Diversi , Silvio Simani

VSM similarity ~33% (0.331), PageRank ~8% (0.0833)

*Abstract:* [756] - The paper presents some results concerning fault diagnosis for dynamic processes using dynamic system identification and disturbance de-coupling techniques. The first step of the considered approach consists of exploiting input-output descriptions of the monitored system. In particular, the disturbance term of

Hi **prof_Verdone**. Enter your query...

## Search on your local index

Find articles with **at least** one of the words [ ] in [title] Search!
Find articles **written by** [ ] Search!

**15** documents have been indexed.
Cluster centroid **RSS** = 16.21

## P2P search

Find articles with **at least** one of the words [ ] in [title] Search!

You have **2** taste buddies:

| Name | IP address | Similarity |
|---|---|---|
| **prof_Corradi** | @localhost:16204 | **11.9920**% |
| **prof_Sartori** | @localhost:16201 | **14.5397**% |

Buddycast cluster centroid approximation = **55.449**% of the norm

air algorithms average beacon-enabled channel chs coverage data dca ddsp delay design distributed energy estimation fh field given impact la level lifetime mac mathematical mobile network networks node nodes number performance processing scalar scheduling sensor sensors signal sink sinks strategy supervisor technique throughput trade-off transmission tree users using video wsn

## Recommendations

**prof_Corradi** recommends you:

**Context-aware handoff middleware for transparent service continuity in wireless networks**
Paolo Bellavista , Antonio Corradi , Luca Foschini
VSM similarity ~25% (0.252), PageRank ~2% (0.0203)
*Abstract:* [1357] - Advances in wireless networking and content delivery are enabling new challenging provisioning scenarios where a growing number of users access continuous services, e.g., audio/video streaming,

Applications Places System 23 °C Tue Jul 21, 01:25:26

leo@leo-laptop: ~/workspace/code/p2p-search/src

File Edit View Terminal Tabs Help

leo@leo-laptop... leo@leo-laptop... leo@leo-laptop... leo@leo-laptop... leo@leo-laptop... leo@leo-laptop... leo@leo-laptop...

```
...
        (0.1275) prof_Sartori@localhost:16201
        (0.0843) prof_Verdone@localhost:16206
[buddyBuilder-TH_04] friendships of prof_Diversi@<ServerProxy for localhost:16205/RPC2> have been updated.

[buddyBuilder-TH_05] attempting to contact prof_Sartori@<ServerProxy for localhost:16201/RPC2> for updating his taste buddies
...
        (0.1567) prof_Verdone@localhost:16206
        (0.1275) prof_Diversi@localhost:16205
[buddyBuilder-TH_05] friendships of prof_Sartori@<ServerProxy for localhost:16201/RPC2> have been updated.

[buddyBuilder] 6 peers in the network, 0 are alone, 6 have friends for a total of 12 friendships.
[buddyBuilder] friendships updated in 0 s, 273408 us.
[superPeerServer] prof_Sartori@localhost:16201 posted its profile (30 keywords)...
[superPeerServer] Profiles similarity matrix:
        prof_Bassi      1.000  0.000  0.000  0.054  0.016  0.000
        prof_Ciaccia    0.000  1.000  0.041  0.022  0.103  0.027
        prof_Corradi    0.000  0.041  1.000  0.021  0.059  0.120
        prof_Diversi    0.054  0.022  0.021  1.000  0.115  0.084
        prof_Sartori    0.016  0.103  0.059  0.115  1.000  0.145
        prof_Verdone    0.000  0.027  0.120  0.084  0.145  1.000
[buddyBuilder] calculating friendships (each peer has at most 2 taste buddies)...
[buddyBuilder] Taste buddies adjacency matrix:
        prof_Bassi      0  0  0  1  1  0
        prof_Ciaccia    0  0  1  0  1  0
        prof_Corradi    0  0  0  0  1  1
        prof_Diversi    0  0  0  0  1  1
        prof_Sartori    0  0  0  1  0  1
        prof_Verdone    0  0  1  0  1  0
[buddyBuilder] Taste buddies similarity adjacency matrix:
        prof_Bassi      0       0       0       0.0541  0.0160  0
        prof_Ciaccia    0       0       0.0413  0       0.1028  0
        prof_Corradi    0       0       0       0       0.0591  0.1199
        prof_Diversi    0       0       0       0       0.1151  0.0843
        prof_Sartori    0       0       0       0.1151  0       0.1454
        prof_Verdone    0       0       0.1199  0       0.1454  0
```

Graphviz - Mozill... demo - File Browser leo@leo-laptop: ~... *citgraph.dot (~/... neatoguide-1.pdf