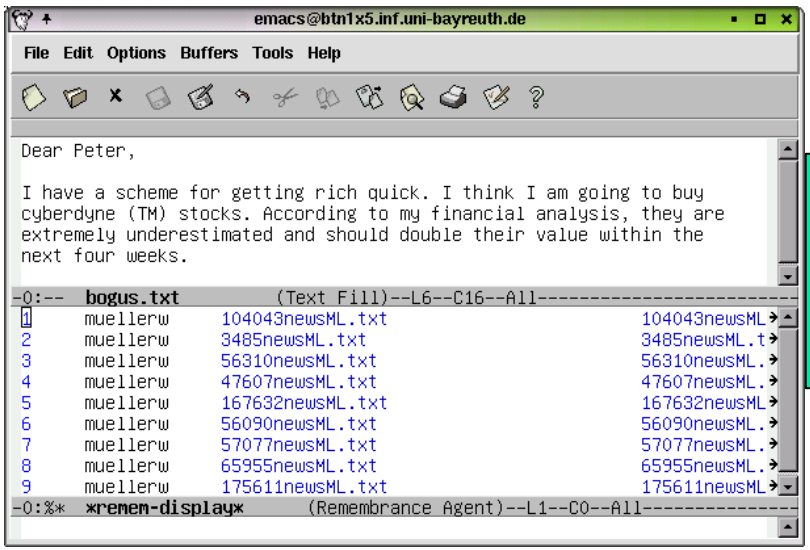
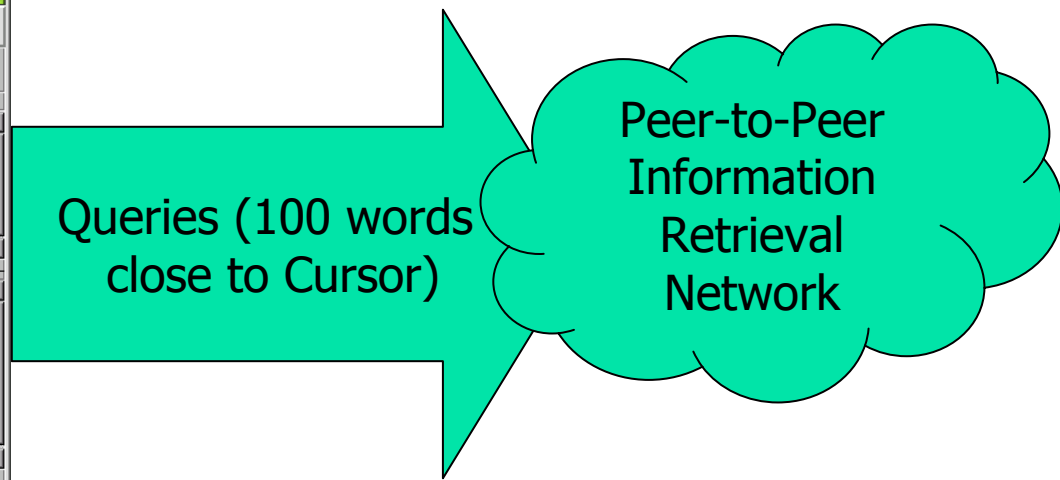
Privacy of Ideas in P2P Information Retrieval Queries

Wolfgang Müller, Andreas Henrich
Angewandte Informatik I
Universität Bayreuth
wmueller@btn1x1.inf.uni-bayreuth.de

Scenario: Personal Information Agent queries P2P net





Personal information agent (automatic query formulation, proactive presentation of useful information)

Issue: ideas transferred along with query

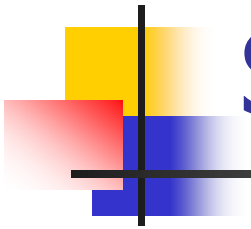
Relation to previous work

- Observation
  - Publisher/Reader anonymity hot topic
  - Private IR hides query, Yet:
    PIR [Chor *et al.,* 1995…]→ either
    - Distributed servers or
    - Costly calculation

→ Motivation for
  → less private than PIR
  → less costly than PIR,

IR,

via weaker, yet useful variant of query anonymity



Setting

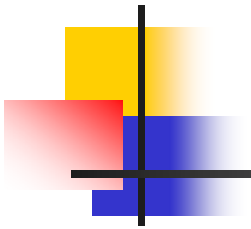
- Queries about sensitive data in P2P network
  - Unknown query processors
  - Difficult to track rogue peers
- Privacy concerns:
  - **Not:** Downloads (we don‘t care)
  - Don‘t want to leak **ideas** behind the query **to other peers**

What is a (new) idea?

- In the strong sense:
  A piece of information whose semantic meaning is not present in the document collection C

- ➔ too hard to measure
- ➔ „Working definition“:
  K be set of Keywords.
  No single document in C contains all k ∈ K
  ➔ K is a new idea with respect to C

Approach



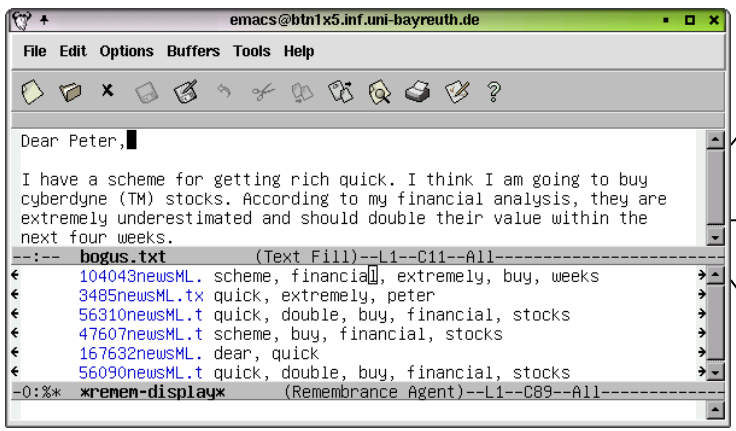
- Avoid querying revealing new ideas by
  - **Splitting** the query into subqueries of single words
  - **Anonymizing** each subquery to avoid linking
  - **Merging** results
- Issue
  - Many queries with low selectivity→costly

➔Try to improve on communication cost

➔Split into fewer, longer queries; minimize

cost(query)=cost(privacy risk)+
cost(communication)

→Split query into single words
→Anonymize subqueries



„cyberdyne stock underestimated"