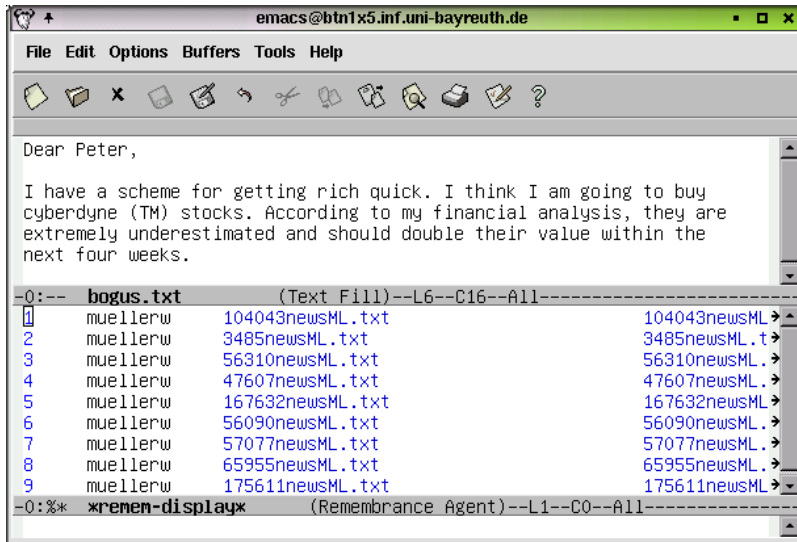




Privacy of Ideas in P2P Information Retrieval Queries

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

Scenario: Personal Information Agent queries P2P net



The screenshot shows an Emacs window titled 'emacs@btn1x5.inf.uni-bayreuth.de'. The main text area contains a message: 'Dear Peter, I have a scheme for getting rich quick. I think I am going to buy cyberdyne (TM) stocks. According to my financial analysis, they are extremely underestimated and should double their value within the next four weeks.' Below the text is a list of search results for the file 'bogus.txt'. The results are as follows:

Line	File Name	File Name
1	mueллерw 104043newsML.txt	104043newsML
2	mueллерw 3485newsML.txt	3485newsML.t
3	mueллерw 56310newsML.txt	56310newsML.
4	mueллерw 47607newsML.txt	47607newsML.
5	mueллерw 167632newsML.txt	167632newsML
6	mueллерw 56090newsML.txt	56090newsML.
7	mueллерw 57077newsML.txt	57077newsML.
8	mueллерw 65955newsML.txt	65955newsML.
9	mueллерw 175611newsML.txt	175611newsML

Queries (100 words close to Cursor)

Peer-to-Peer Information Retrieval Network

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 $k \in K$
→ 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
- $$\text{cost(query)} = \text{cost(privacy risk)} + \text{cost(communication)}$$

- Split query into single words
- Anonymize subqueries

