JACOBS
UNIVERSITY

Distributed Case-based Reasoning for Fault Management

Ha Manh Tran and Jürgen Schönwälder
Computer Science, Jacobs University Bremen, Germany

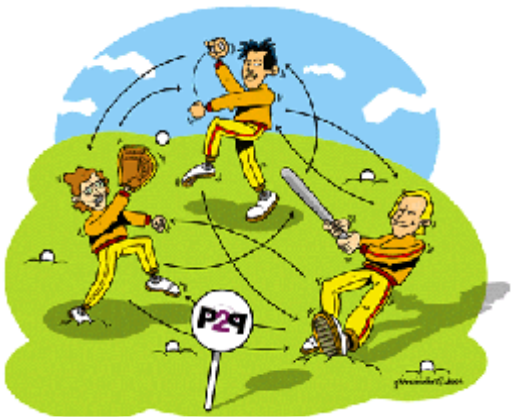
1st EMANICS Workshop on Peer-to-Peer Management
University of Zurich, 3-4 March 2008

Motivation & Goal

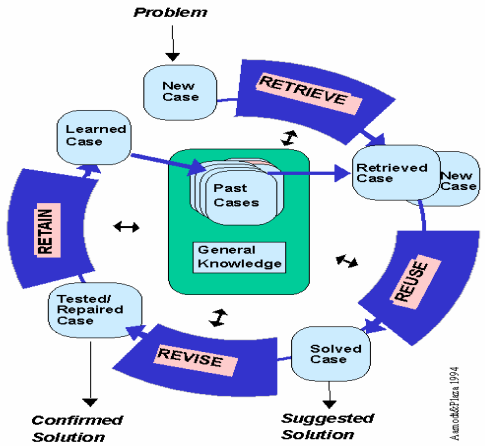
- What if operators find out a fault on a system?
  - Searching for the solution: Google, Yahoo, …
  - An example of upgrading the firmware of Juniper routers
- How operators are assisted in resolving faults?
  - Searching for problem-solving experience shared in decentralized environments

Distributed Case-based Reasoning System

- Peer-to-Peer
  - self-organization
  - scalability and versatility
- Case-based Reasoning
  - Problem-solving method
  - Inference on experience



Peer-to-Peer



Case-based Reasoning

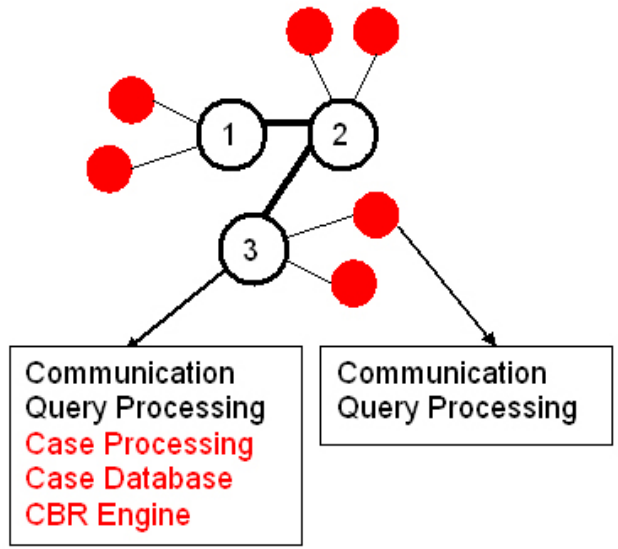
- Peer-to-Peer + Case-base Reasoning = Distributed CBR improving computational performance and case databases maintenance

Outline

- Peer-to-Peer
  - Heuristic Search using a Feedback Scheme in Unstructured Peer-to-Peer Networks [1]
- Case-base Reasoning
  - Fault Representation in Case-based Reasoning
- Work in progress
  - Crawling Bug Tracking Systems: Bugzilla, Mantis, Trac, Debian
  - Fault Resolution in Case-based Reasoning: Probabilistic reasoning

System Overview

- Unstructured P2P network overlay
  - Search performance issue
- Super peers bearing CBR engines
  - Reasonable bandwidth and power processing
- CBR engines proposing fault-matching solutions
  - Local case database and reasoning engine
  - High computational resource consumption



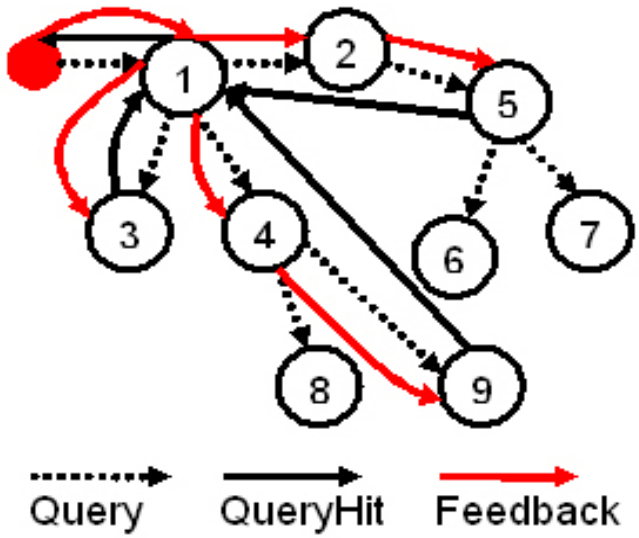
A Heuristic Search

- Finding solutions:
  - Flooding the overlay is avoided
  - Proposed solutions are probably incorrect
  - Finding promising peers by looking at their previously accepted solutions
- Using a feedback scheme to announce the accepted solutions to peers

Feedback Scheme

- A querying peer verifies and accepts solutions among fault-matching solutions, then feedback the accepted solutions to specific peers.

- Upon receiving the feedback, any peer learns solutions and peers for subsequent queries



Algorithms

- Peer learning
  - Learning from feedback
  - Updating the lists of good peers (expertise values) and queries (query information)
- Peer ranking
  - Finding similar queries and corresponding peers
  - Ranking these peers
- Peer selection
  - Selecting peers from lists of good peers, recently active peers, and random peers in order
  - At least, one random peer

Similarity Function

- Learning and ranking algorithms
- Field-value pairs presentation
- Ordered Weighted Averaging [Ronald Yager 1988]

$$\mathbf{sim(q,c)=\sum w_i sim_{\sigma(i)}(q_i,c_i)}$$

q_i, c_i: field i;

w_i: weight i (a monotonic function satisfying $\sum \mathbf{w_i}=1$)

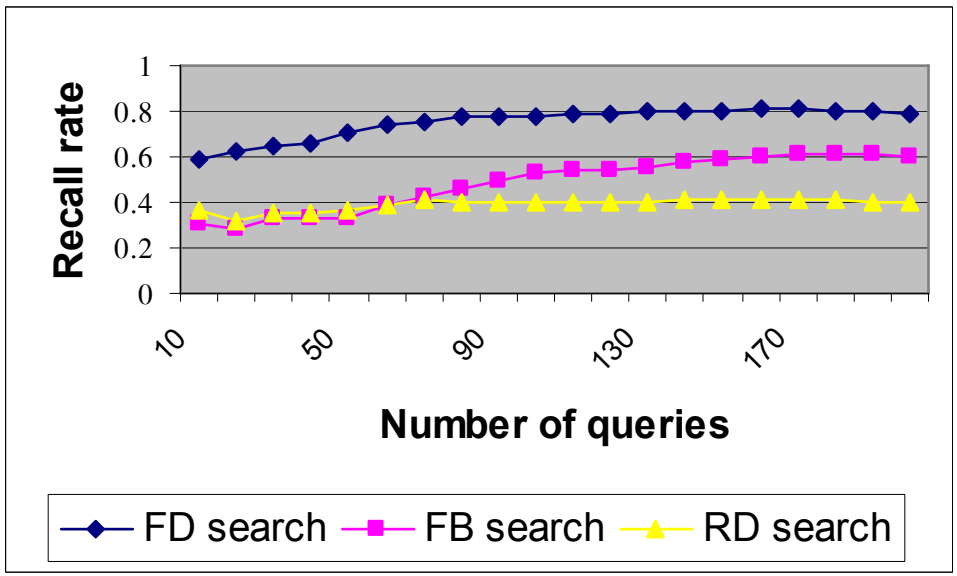
$sim_{\sigma(i)}(q_i,c_i)$: distance of q_i, c_i following a permutation $\sigma(i)$

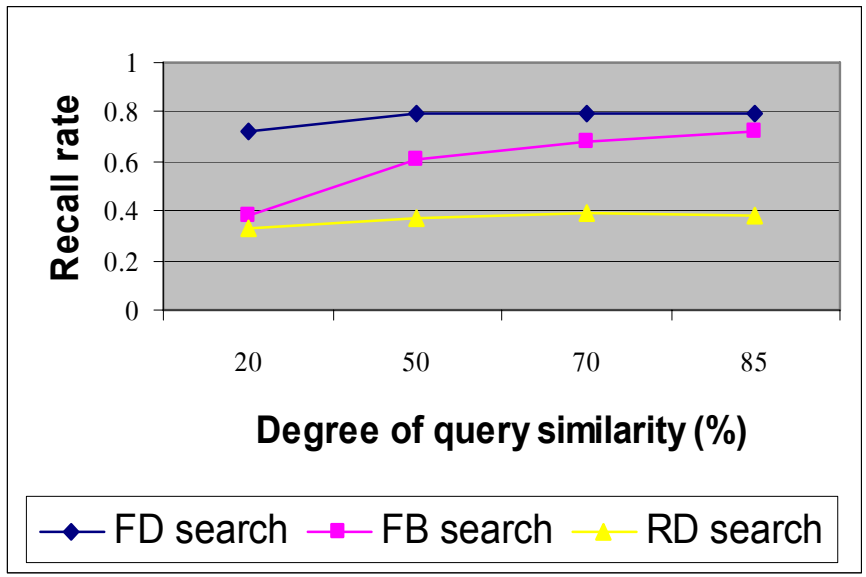
Experiment Setup

- Gnutella network simulation
- SIMILE and CACM bibliographic data-sets
  - 3500 bibtex titles
  - Query sets of 20%, 50% and 85% similarity
- Comparison 3 search mechanisms
  - Flooding-based search (FD, 4 neighbors)
  - Feedback-based search (FB, 3 selected neighbors)
  - Random-based search (RD, 3 random neighbors)

Scheme Evaluation

- Recall rate of retrieved bibtexes
  - Increasing the efficiency of the search mechanism
  - Reaching 77% of the recall rate limit with the query set of 50% similarity



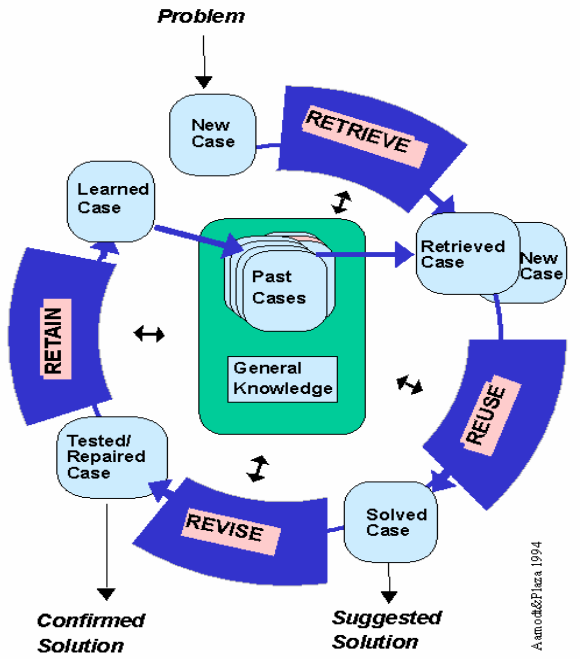


Outline

- Peer-to-Peer
  - Heuristic Search using a Feedback Scheme in Unstructured Peer-to-Peer Networks
- Case-base Reasoning
  - Fault Representation in Case-based Reasoning [2]
- Work in progress
  - Crawling Bug Tracking Systems: Bugzilla, Mantis, Trac, Debian
  - Fault Resolution in Case-based Reasoning: Probabilistic reasoning.

Inside Case-based Reasoning

- Case retrieval
  - Obtaining relevant cases
  - Case representation
  - Evaluation function
- Case reuse, revision, retaining
  - Reasoning on the retrieved cases
  - Determining the best case
  - Updating the case database
- The focus of this work is on case retrieval



Case-based Reasoning

Representation in Feature Vectors

- Set of field-value pairs
  - <$field_1$:$value_1$,…,$field_n$:$value_n$>
  - Fields are domain-specific
  - Values are binary, numeric or symbolic
- Easy to evaluate similar cases
  - High accuracy
- Difficult to express textual cases
- Popular to several CBR applications

Evaluation of Feature Vectors

- Global similarity [Miquel et al. 2002]

$$\mathbf{sim(q,c)=\sum w_i sim(q_i,c_i)}$$

  q_i, c_i: value of field i of q_i and c_i

  w_i: weight i satisfying $\sum w_i=1$

  $sim(q_i,c_i)$: distance between q_i and c_i

- Logical match [Igor et al. 2003]
  - Using field-value pairs as predicates
- Word similarity [Yuhua et al. 2003]
  - Using the word taxonomy tree

Representation in Semantic Vectors

- Set of terms
  - <“representation”,”semantic”,”vector”>
  - Frequency and distinction of terms
  - Indexing terms to generate semantic vectors
- Suitable for expressing textual cases
- Easy to evaluate similar cases
  - Average accuracy
- Popular to several text-processing applications

Evaluation of Semantic Vectors

- Cosine similarity function

$$\mathbf{cos(q,c)=\sum q_i c_i}$$

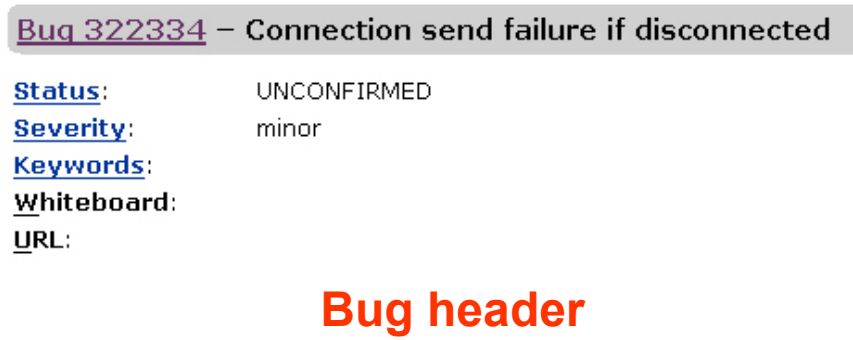
q_i, c_i: value of term i of q_i and c_i

q,c: normalization to 1

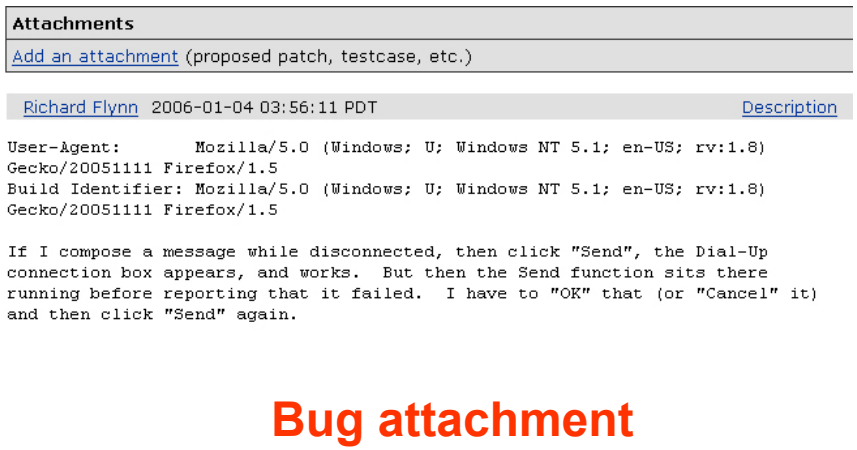
- Accuracy issues
  - Indexing the huge corpus
  - Using approximate methods
  - Very small q_i, c_i

A Bug Report

- Header
  - Set of field-value pairs
  - Management information
- Attachment
  - Textual description
  - Problem and discussion information



Bug header



Bug attachment

Multi-Vector Representation

- A feature vector for classification information
  - Domain-specific fields
  - Connectivity, performance, configuration
- A feature vector for diagnosis information
  - Symptoms, typical parameters
  - Error message, debug information
- A semantic vector for problem information
  - Term significance and distinction
  - Problem, discussion, solution

Evaluation Setup

- Lacking fault datasets
- Using bibliographic data-sets
  - CISI: 1460 titles and MED: 1033 titles
  - Field-value pairs and textual descriptions
  - Keyword-specific and textual queries
- Using recall and precision metrics

.I 133
.T
The Annual Review of Information Science and Technology
.A
Cuadra, C.A.
.B
1967
.W
This volume is the second in a series of Annual Reviews of progress in the field of Information Science and Technology. Like its predecessor, it attempts to describe, compare, and evaluate the most significant work that has been reported in the field during the past year. The effort has been undertaken in the belief that such taking stock of accomplishments provides a valuable service to the specialists in the information science field. The chapters on New Techniques for Publication and Distribution of Information, on New Developments in Chemical Documentation, and on Applications in Medicine.
.X
471 2 133
565 2 133
28 2 133
40 2 133

A sample bibtex

A textual query:
What problems and concerns are there in making up descriptive titles? What difficulties are involved in automatically retrieving articles from approximate titles? What is the usual relevance of the content of articles to their titles?

A keyword-based query:

```
#q1= #and ('titles', #or ('automatically',
          'retrieving', 'problems', 'concerns',
          'descriptive', 'approximate',
          'difficulties', 'content', 'relevance',
          'articles'));
```

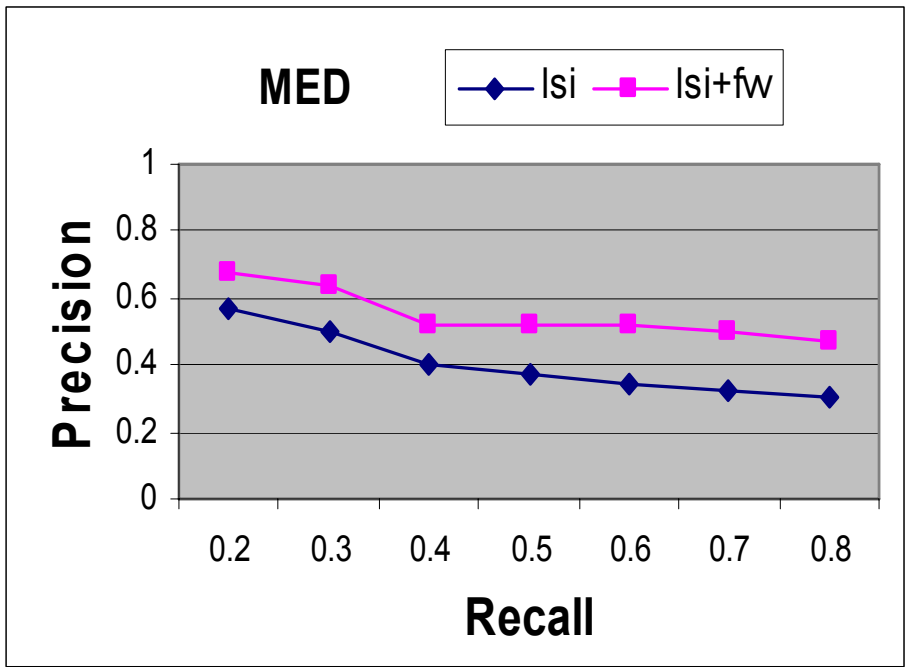
Sample queries

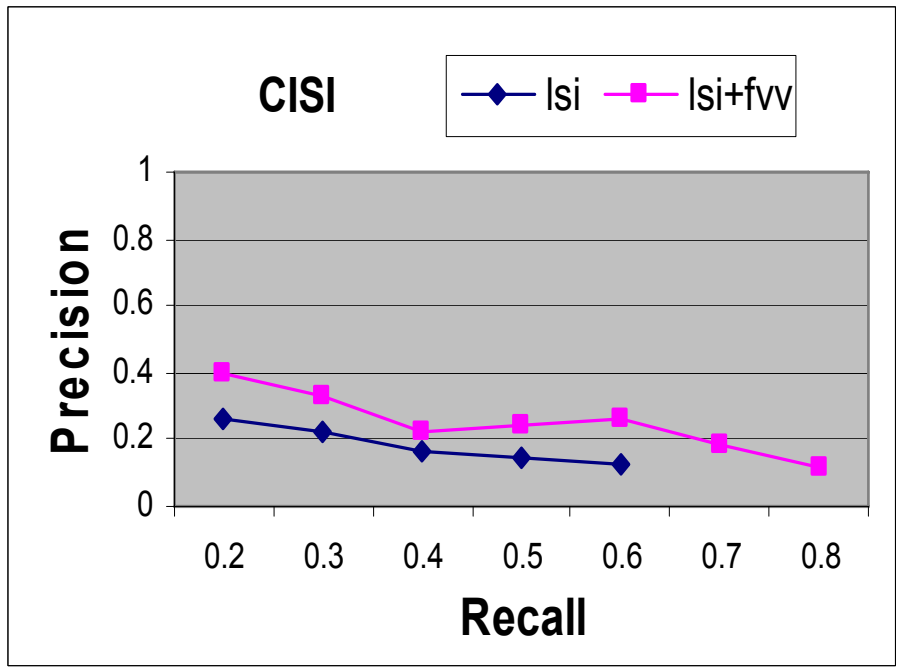
Evaluation Setup

- Performance comparison
  - Semantic vectors (lsi) vs. multiple vectors (lsi+fvv)
- Feature vectors
  - Significance of specific keywords using the *term* x *document* matrix
  - Weight of keywords using *and*, *or*, *not* operators
- Semantic vectors
  - Jacobi method for indexing terms
  - More accuracy, but slow computation

Precision by Recall (CISI and MED data-sets)

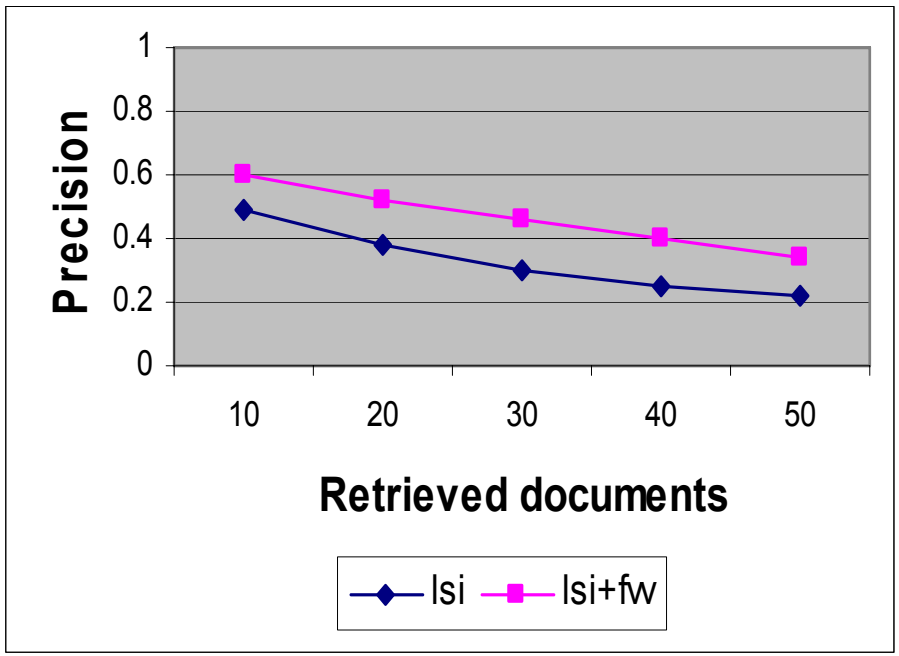
- Retrieving recall rate to compute precision rate
- Lsi+fvv outperforms lsi in both data-sets
- Lsi performs similarly to [Scott et al. 1990]

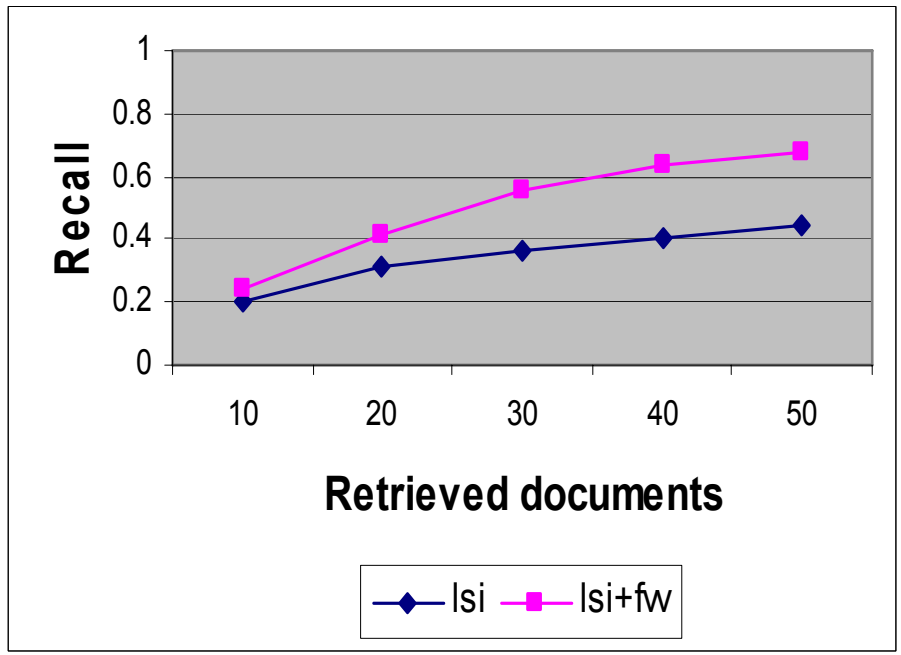




Recall and Precision by Retrieved Documents (MED data-set)

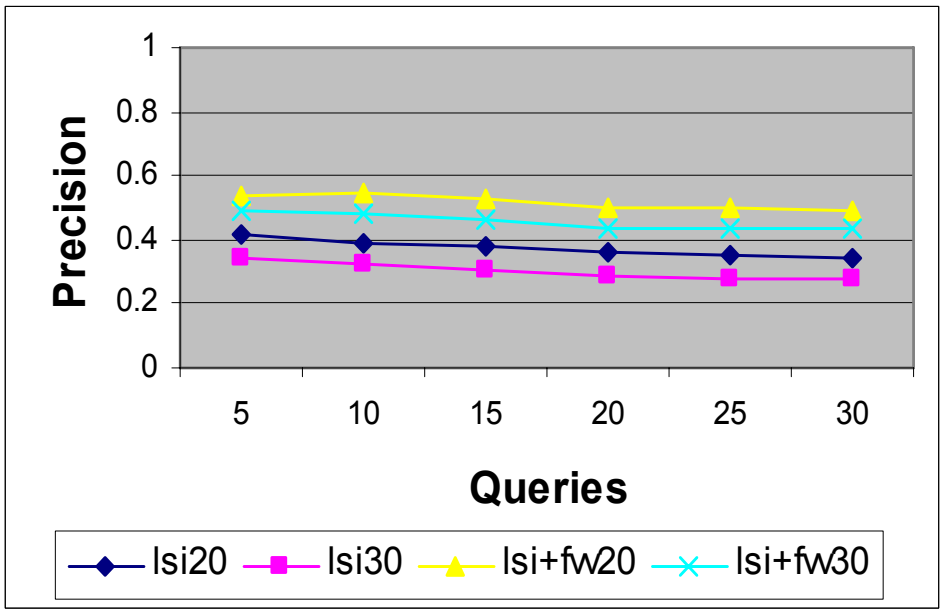
- Accumulative rates for retrieved documents
- Lsi is misled by a *not* operator in queries
- Lsi+fvv performs well with *distinct* keywords

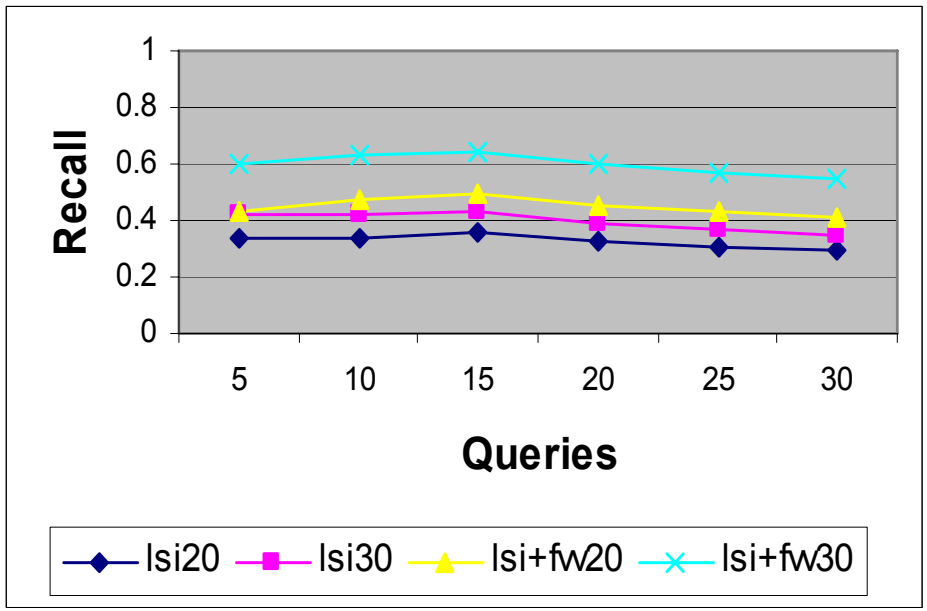




Recall and Precision by Queries (MED data-set)

- Retrieving 20 and 30 documents per query
- Each query obtains the similar ratio of relevant documents to retrieved documents





Summary

- Distributed CBR system aims to search for relevant resources for resolving problems
- Feedback scheme improves search in the proposed system
- Multi-vector representation improves semi-structured resource retrieval in the proposed system
- On-going work involves crawling bug tracking systems for datasets and providing reasoning methods for CBR engines

Possible Issues

- The issues of P2P overlay network
  - Performance
    - Efficient routing mechanism ? Centralized or decentralized reasoning ?
  - Security and privacy
    - Sharing resources with honest peers ? Anonymous peers requesting resources ?
- The issues of CBR
  - Case Representation
    - Expressive format for retrieval and reasoning ?
  - Datasets
    - Testbed for reasoning ?

References

[1] H.M. Tran, J. Schönwälder: *Heuristic Search using a Feedback Scheme in Unstructured Peer-to-Peer Networks*. 5th International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2007), Vienna, September 2007. Springer LNCS. To appear.

[2] H.M. Tran, J. Schönwälder: *Fault Representation in Case-based Reasoning*. 18th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management (DSOM 2007), San Jose, October 2007. Springer LNCS 4785

Thank you for your attention

Questions?