Fault Representation in Case-based Reasoning

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

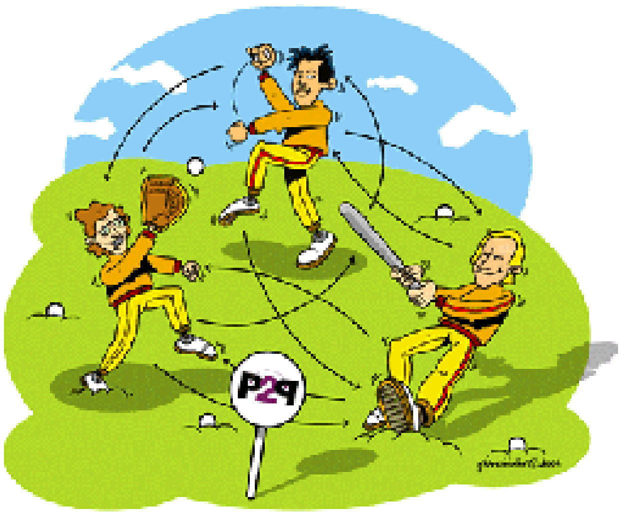
18th IFIP/IEEE International Workshop on
Distributed Systems: Operations and Management
San José, California , 29-31 October 2007

Distributed Case-based Reasoning System

- Assisting network operators in resolving faults
  - Large-scale, diverse communication systems
  - Complex and difficult faults
  - Dynamic and important services
- Searching for solutions by experience sharing in decentralized environments
  - Exploring ubiquitous resources
  - Exploiting relevant resources

Distributed Case-based Reasoning System

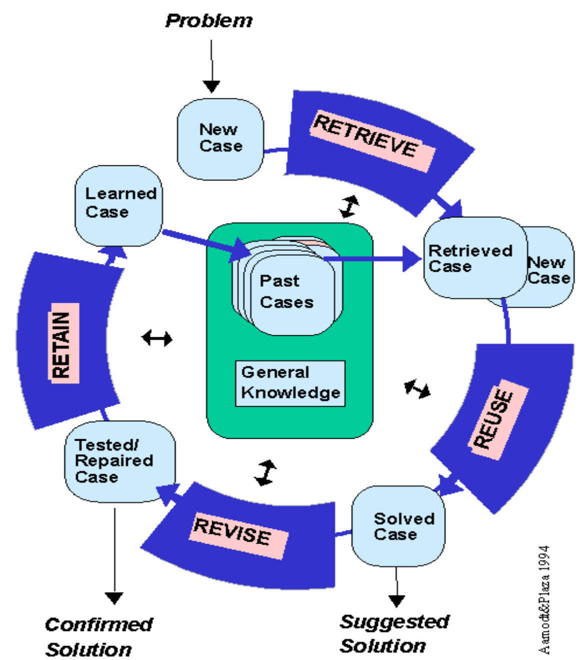
- Peer-to-Peer
  - self-organization
  - scalability in architecture
  - flexibility in search
- Case-based Reasoning
  - Problem-solving method
  - Classification and resolution
  - Inference on experience
- Distributed CBR improving computational performance and case databases maintenance



Peer-to-Peer

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
  - $<field_1:value_1, \ldots, field_n:value_n>$
  - 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

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

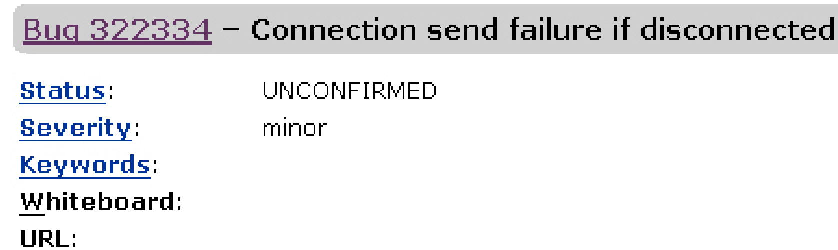
q_i,c_i: value of term i of q_i and c_i

q,c: normalization to 1

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

A Bug Report

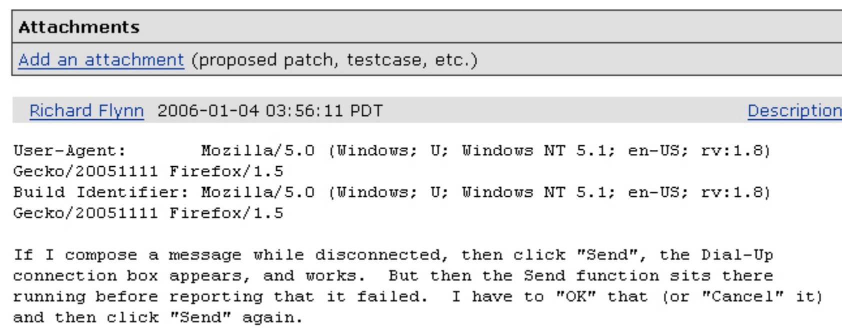
- Header
  - Set of field-value pairs
  - Management information
- Attachment
  - Textual description
  - Problem and discussion information



Bug 322334 – Connection send failure if disconnected

Status:	UNCONFIRMED
Severity:	minor
Keywords:	
Whiteboard:	
URL:	

Bug header



Attachments

Add an attachment (proposed patch, testcase, etc.)

Richard Flynn 2006-01-04 03:56:11 PDT Description

```
User-Agent:       Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8)
Gecko/20051111 Firefox/1.5
Build Identifier: Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8)
Gecko/20051111 Firefox/1.5

If I compose a message while disconnected, then click "Send", the Dial-Up
connection box appears, and works.  But then the Send function sits there
running before reporting that it failed.  I have to "OK" that (or "Cancel" it)
and then click "Send" again.
```

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, packet loss,
- A semantic vector for problem information
  - Term significance and distinction
  - Problem, discussion, solution

Evaluation Function

- Ordered Weighted Averaging [Ronald et al.1988]

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

  q_i,c_i: value of field i of q_i and c_i
  w_i: weight i satisfying $\sum \mathbf{w_i}=1)$
  $sim_{\sigma(i)}(q_i,c_i)$: distance of q_i,c_i following a permutation $\sigma(i)$

  – Monotonic weight function

$$\mathbf{w_i= \begin{cases} 2/(n+2i) \text{ if } i<n/2 \\ 1/2i \text{ if } i \geqslant n/2 \end{cases}}$$

- A decreasing order of important fields: pre-defined fields, user-defined fields.

Evaluation Function

- Cosine function for semantic vectors
  - Indexing the corpus

- Aggregative function:

  $$\mathbf{S_{agg} = \alpha sim_{fv1}(q,c) + \beta sim_{fv2}(q,c) + \gamma cos_{sv}(q,c)}$$

  α, β, γ: parameters
  - Parameters specify the importance of vectors

Evaluation Setup

- Lacking fault data-sets
- 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'));
```

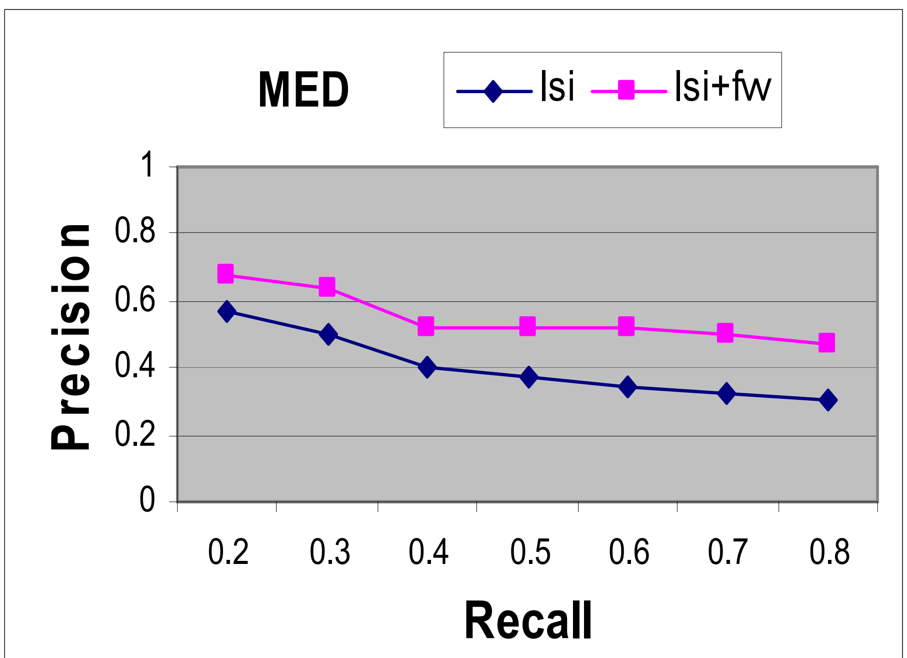
Sample queries

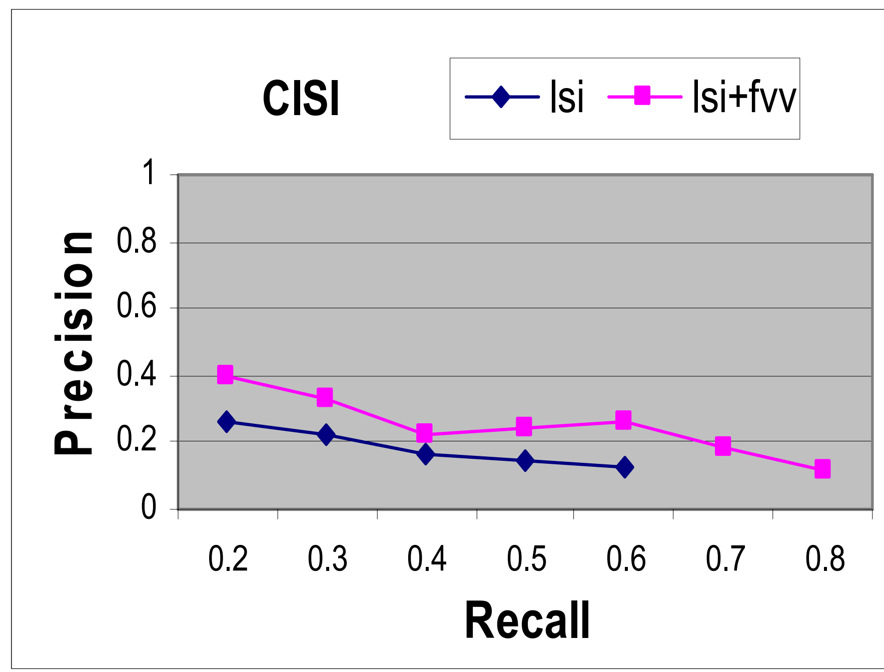
Evaluation Setup

- 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
- Parameters α=0.6 and γ=0.4 (no β)

Precision by Recall (CISI and MED data-sets)

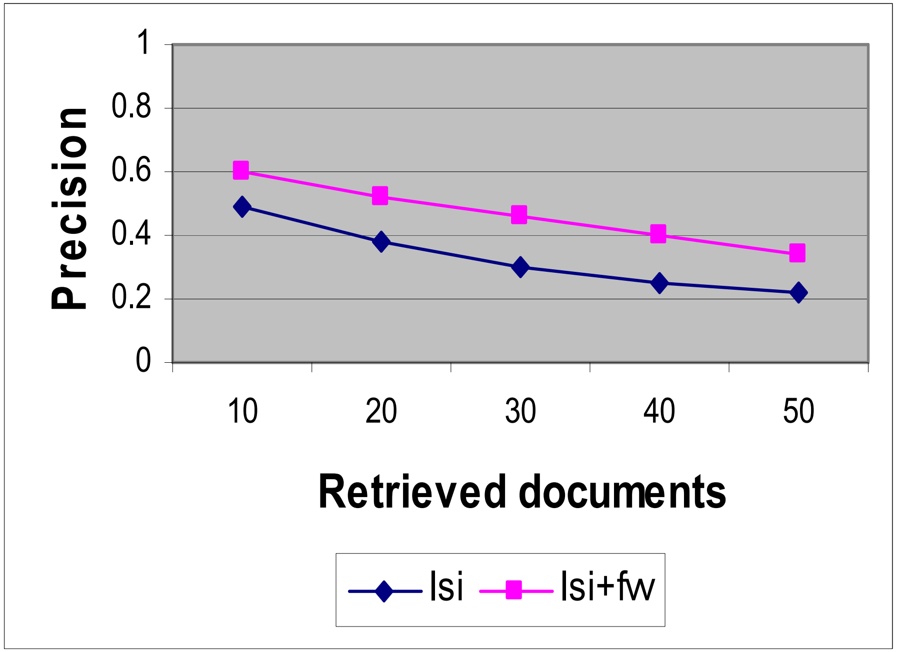
- Retrieving recall rate to compute precision rate
- Lsi+fvv outperforms lsi in both data-sets
- Lsi performs similarly to [Scott et al. 1990]

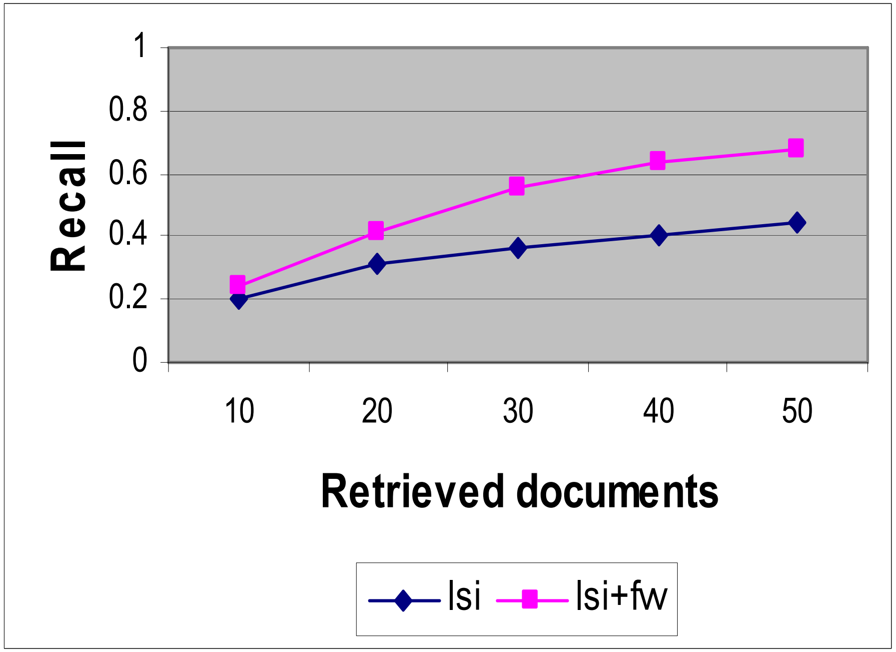




Recall and Precision by Retrieved Documents (MED data-set)

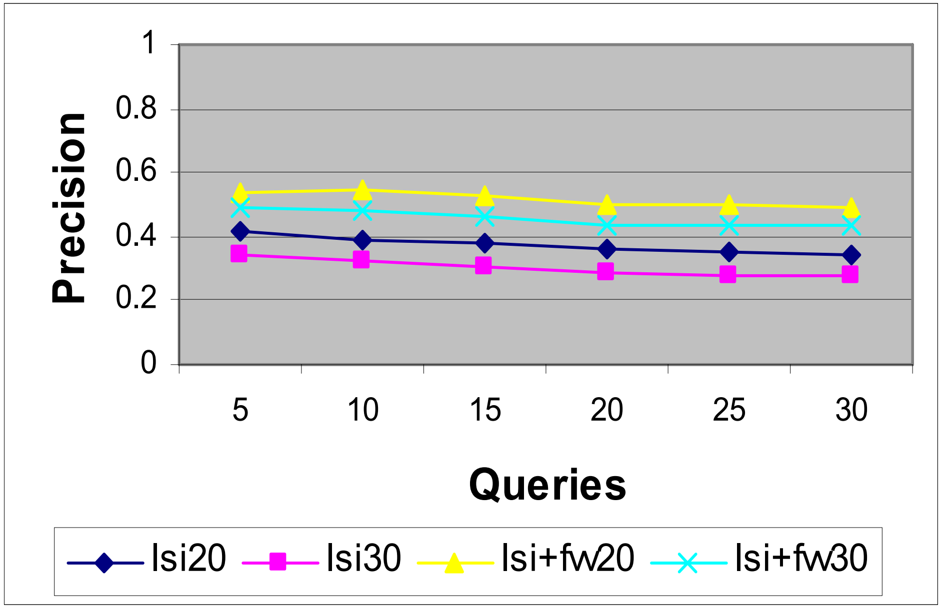
- Accumulative rates for retrieved documents
- Lsi is misled by a *not* operator in queries
- Lsi+fvv performs well with *distinct* keywords

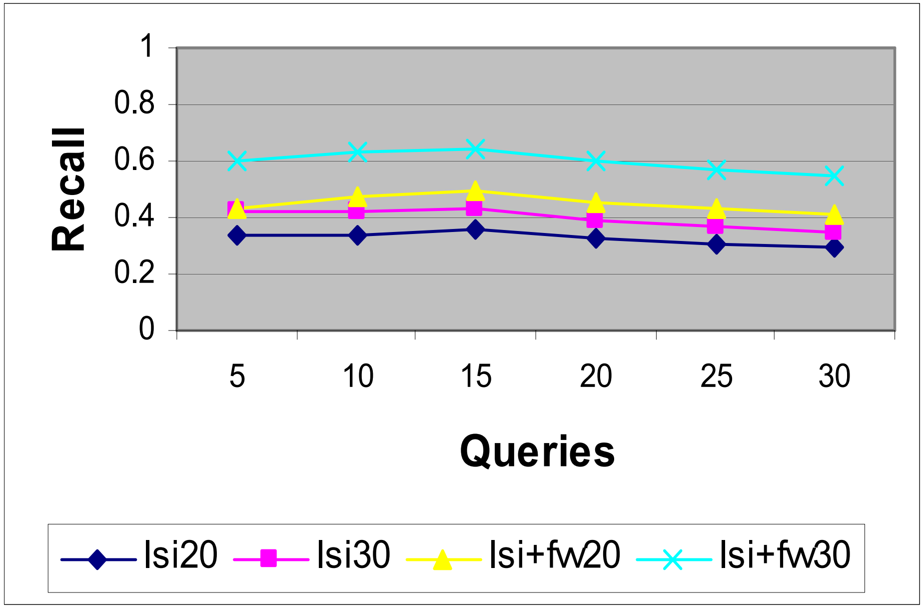




Recall and Precision by Queries (MED data-set)

- 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
- Multi-vector representation helps retrieving more relevant resources
  - Exploit semi-structured data
- Evaluation of multi-vector representation on bibliographic data-sets provides positive results
- Evaluation of multi-vector representation on fault data-sets is work in progress

Thank you for your attention

Questions?