Document Clustering and Collection Selection

Diego Puppin

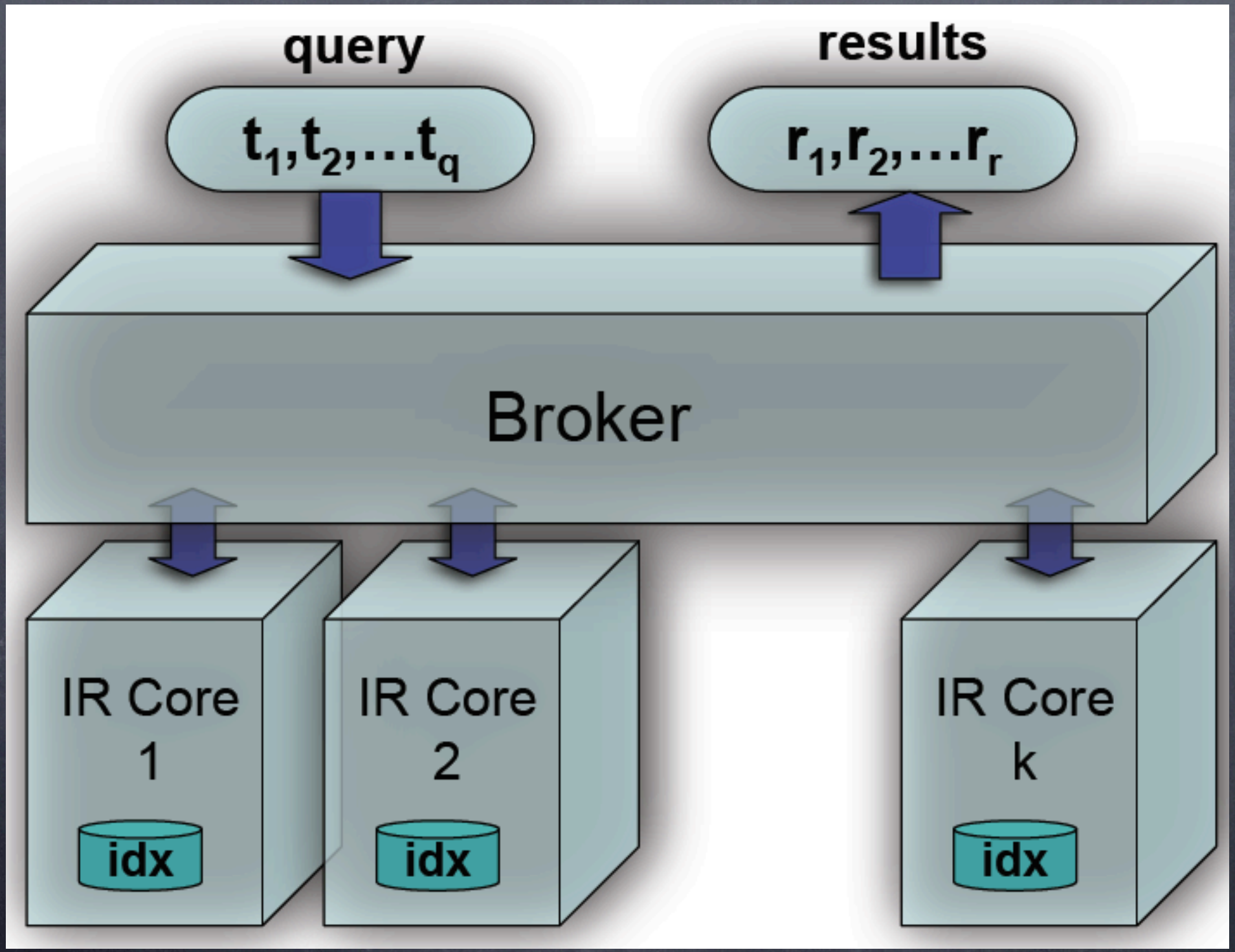
Web Mining, 2006-2007

UNIVERSITÀ DI PISA

Web Search Engines: Parallel Architecture

- To improve throughput, latency, res. quality
- A broker works as a unique interface
- Composed of several computing servers
- Queries are routed to a subset
- The broker collects and merges results

UNIVERSITÀ DI PISA


query
results
$t_1, t_2, \ldots t_q$
$r_1, r_2, \ldots r_r$
Broker
IR Core 1
idx
IR Core 2
idx
IR Core k
idx

	d1	d2	d3	d4	d5	d6	d7	d8
t1	x						x	
t2		x				x		
t3		x				x		x
t4			x				x	
t5	x	x	x	x		x	x	x
t6		x		x		x		
t7	x			x		x		x
t8		x		x			x	
t9								

	d1	d2	d3	d4	d5	d6	d7	d8
t1	x						x	
t2		x				x		
t3		x				x		x
t4			x				x	
t5	x	x	x	x		x	x	x
t6		x		x		x		
t7	x			x		x		x
t8		x		x			x	
t9								

	d1	d2	d3	d4	d5	d6	d7	d8
t1	x						x	
t2		x				x		
t3		x				x		x
t4			x				x	
t5	x	x	x	x		x	x	x
t6		x		x		x		
t7	x			x		x		x
t8		x		x			x	
t9								

Doc-Partitioned Approach

- The document base is split among servers
- Each server indexes and manages queries for its own documents
- It knows all terms of some documents
- Better scalability of indexing/search
  - Each server is independent
  - Documents can be easily added/removed

UNIVERSITÀ DI PISA

Term-Partitioned Approach

- The dictionary is split among servers
- Each server stores the index for some terms
- It knows documents where its terms occur
- Potential for load reduction
- Poor load balancing (some work...)

UNIVERSITÀ DI PISA

Some considerations

- Every time you add/remove a doc
  - You must update MANY servers
- With queries:
  - only relevant servers are queried but...
  - servers with **hot** terms are overloaded

UNIVERSITÀ DI PISA

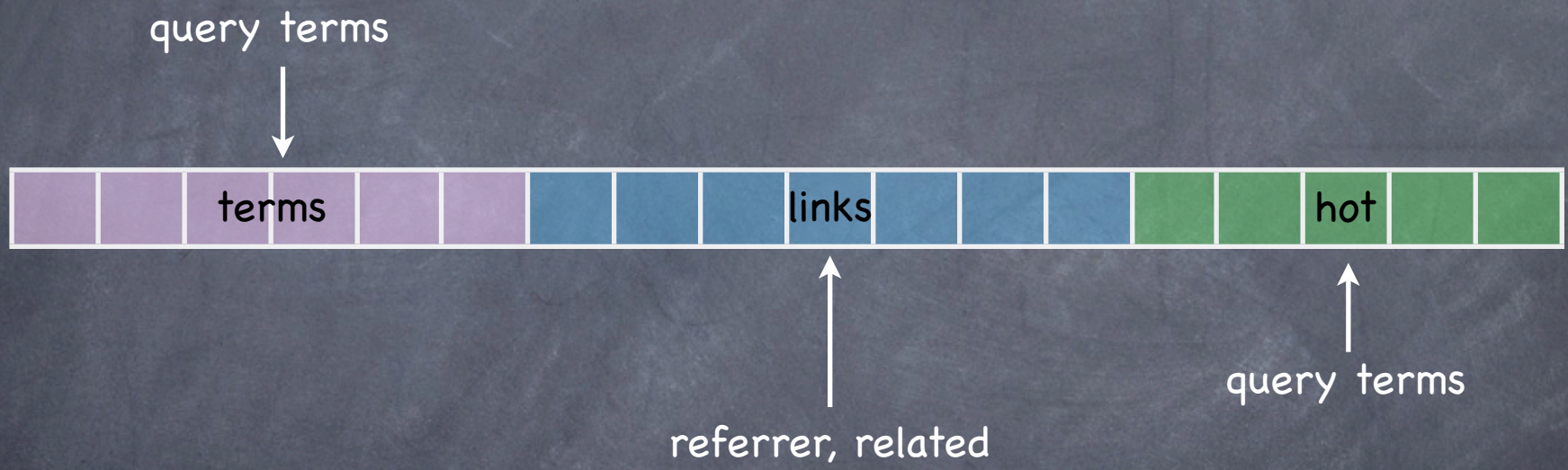
How To Load Balance

- Put together related terms
  - This minimizes the number of hit servers
- Try to put together group of documents with similar overall frequence
  - Servers shouldn't be overloaded
- Query logs could be used to predict

Multiple indexes

- Term-based index
  - Query-vector or Bag-of-word
- Hot text index
  - Titles, Anchor, Bold etc
- Link-based index
  - It can find related pages etc
- Key-phrase index
  - For some idioms

UNIVERSITÀ DI PISA


query terms
terms
links
hot
referrer, related
query terms

How To Doc-Partition?

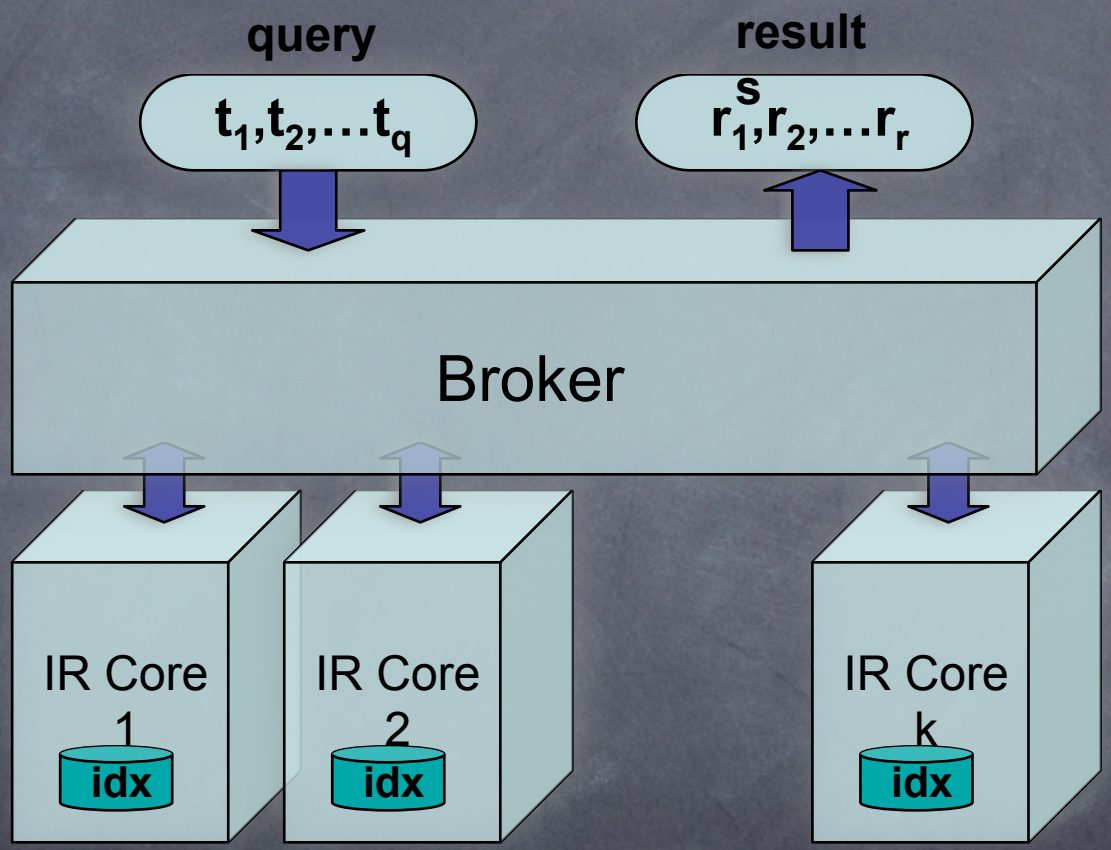
1. Random doc assignment + Query broadcast
2. Collection selection for independent collections (meta-search)
3. Smart doc assignment + Collection selection
4. Random assignment + Random selection

UNIVERSITÀ DI PISA

1. Random + Broadcast

- Used by commercial WSEs
- No computing effort for doc clustering
- Very high scalability
  - Low latency on each server
- Result collection and merging is the heaviest part

UNIVERSITÀ DI PISA

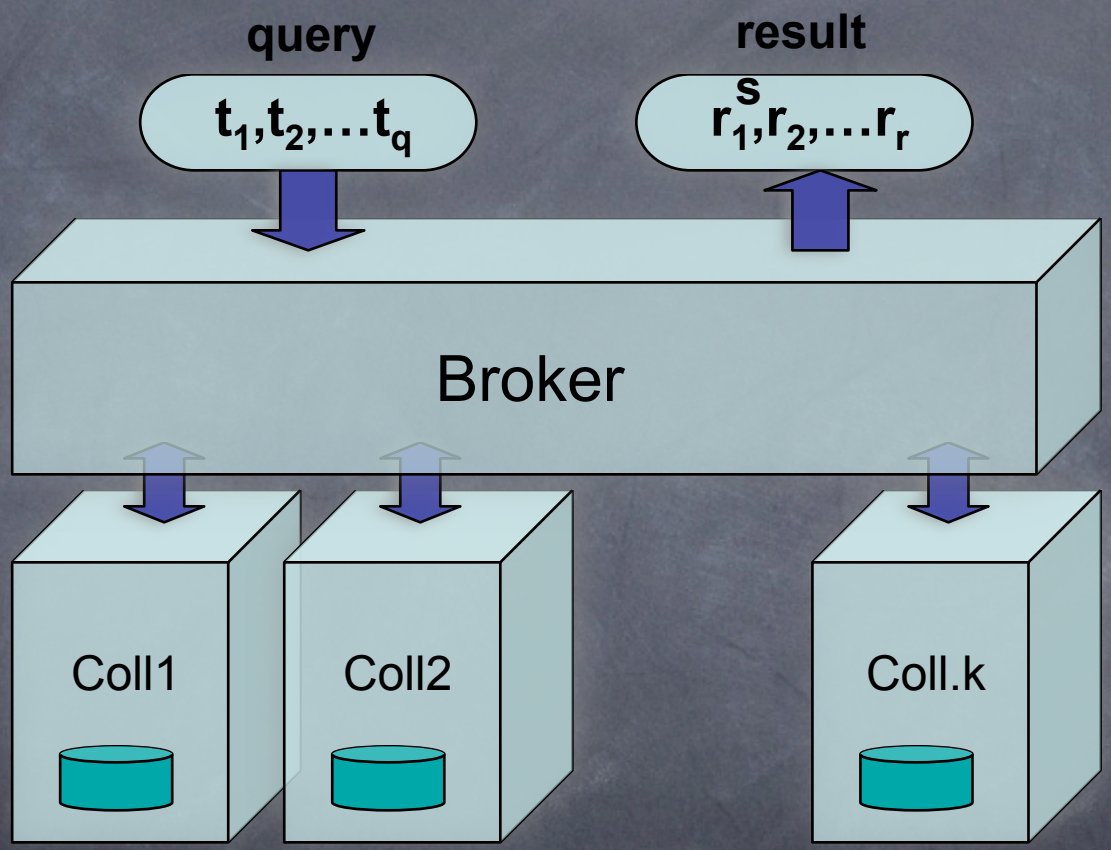


Distributed/Replicated Documents

2. Independent collections

- The WSE uses data from several sources
- It routes the query to the most authoritative collection(s)
- It collects the results according to independent ranking choices (HARD)
- Example: Biology, News, Law

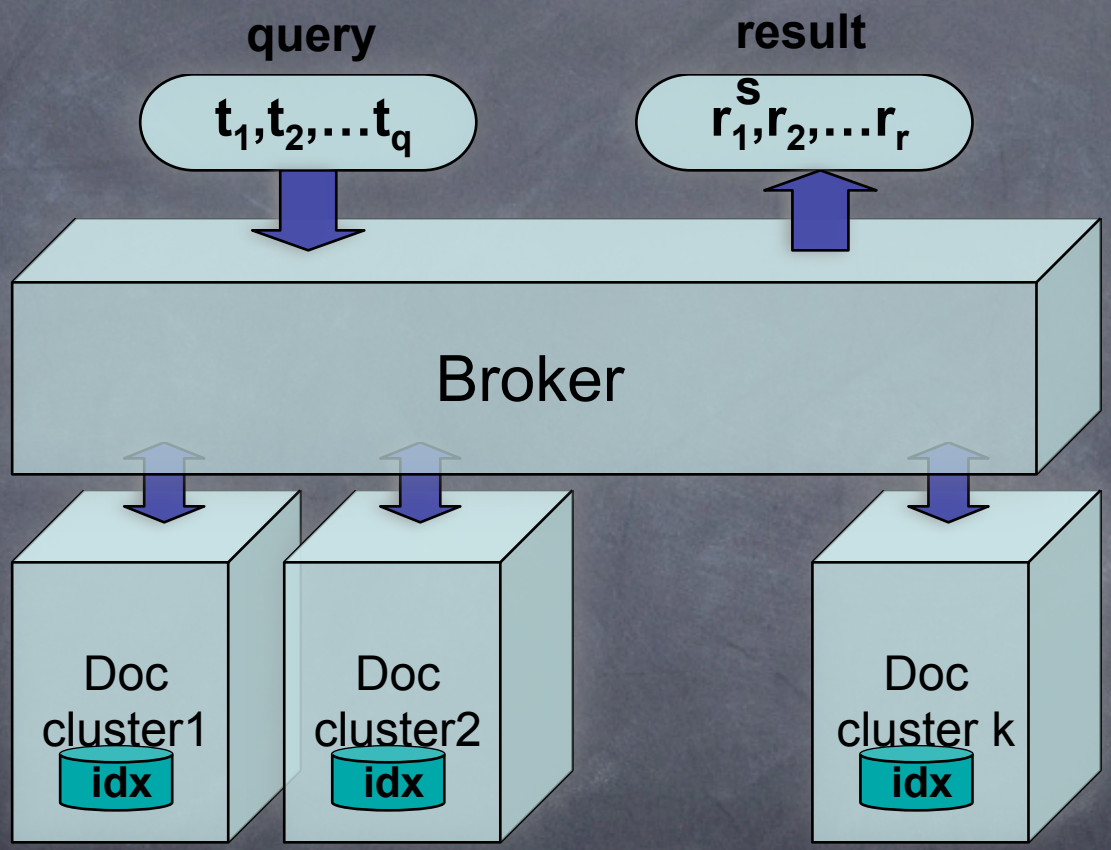
UNIVERSITÀ DI PISA


query
result
$t_1,t_2,\ldots t_q$
s
$r_1,r_2,\ldots r_r$
Broker
Coll1
Coll2
Coll.k

3. Assignment + Selection

- WSE creates document groups
- Each server holds one group
- The broker has a knowledge of group placement
- The selection strategy routes the query suitably

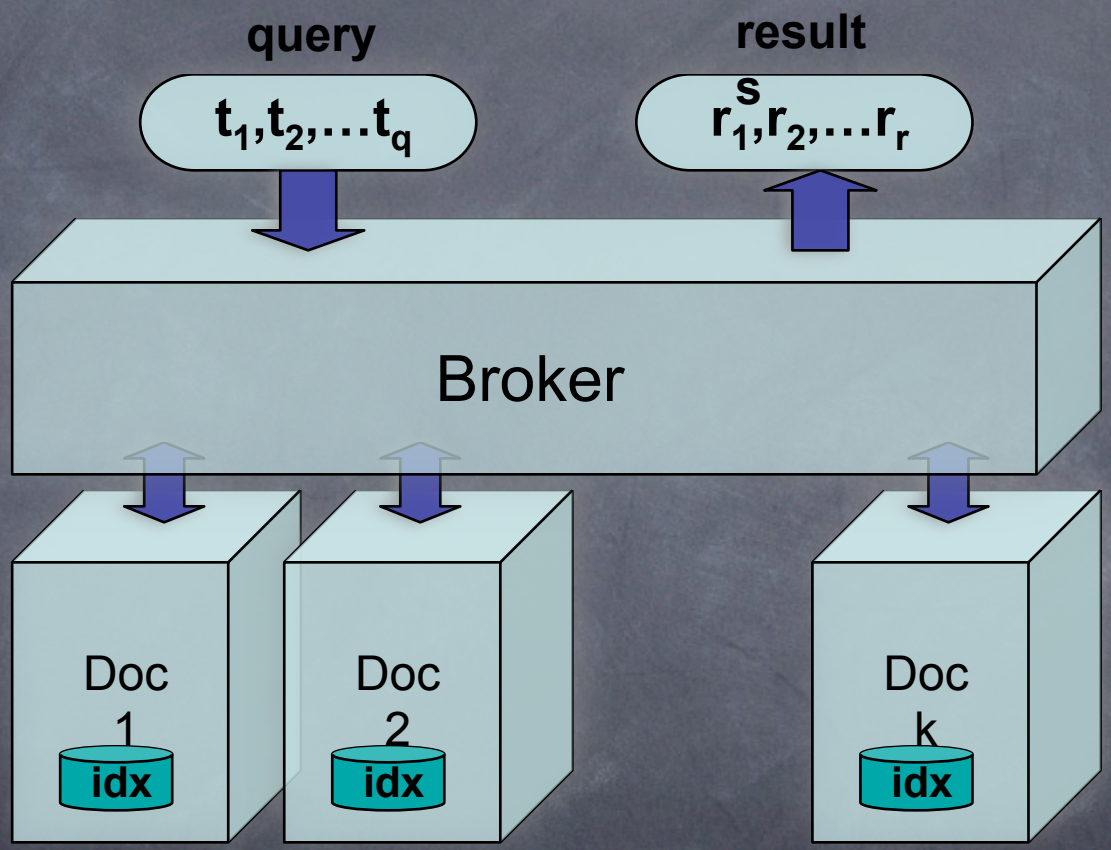
UNIVERSITÀ DI PISA


query
$t_1, t_2, \ldots t_q$
result
s
$r_1, r_2, \ldots r_r$
Broker
Doc cluster1
idx
Doc cluster2
idx
Doc cluster k
idx

4. Random + Random

- If data (pages, resources...) are replicated, interchangeable, hard to index
- Data are stored in the server that publishes them
- We query a few servers hoping to get something

UNIVERSITÀ DI PISA


query
$t_1,t_2,\ldots t_q$
result
s
$r_1,r_2,\ldots r_r$
Broker
Doc 1
idx
Doc 2
idx
Doc k
idx

CORI

- The Effect of Database Size Distribution on Resource Selection Algorithms, **Luo Si and Jamie Callan**
- Extends the concept of TF.IDF to collections

UNIVERSITÀ DI PISA

$$tf = \frac{n_i}{\sum_k n_k}$$

$$idf = \log \frac{|D|}{|d_i \supset t_i|}$$

$$tfidf = tf \times idf$$

$$T = \frac{df}{df + 50 + 150 * cw_i / avg\_cw}$$

$$I = \frac{\log(\frac{|DB| + 0.5}{cf})}{\log(|DB| + 1.0)}$$

$$p(r_k \mid c_i) = b + (1 - b) * T * I$$

df is the number of documents in db_i that contain r_k;
cf is the number of databases that contain r_k;
$|DB|$ is the number of databases to be ranked;
cw_i is the number of words in db_i;
avg_cw is the average cw of the databases to be ranked; and
b is the default belief, usually set to 0.4.

CORI

- Needs a deep collaboration from the collections:
  - Data about terms, documents, size
- Unfeasible with independent collections
  - Statistical sampling, Query-based sampling
- Term-based: no links, no anchors
- Very large footprint

UNIVERSITÀ DI PISA

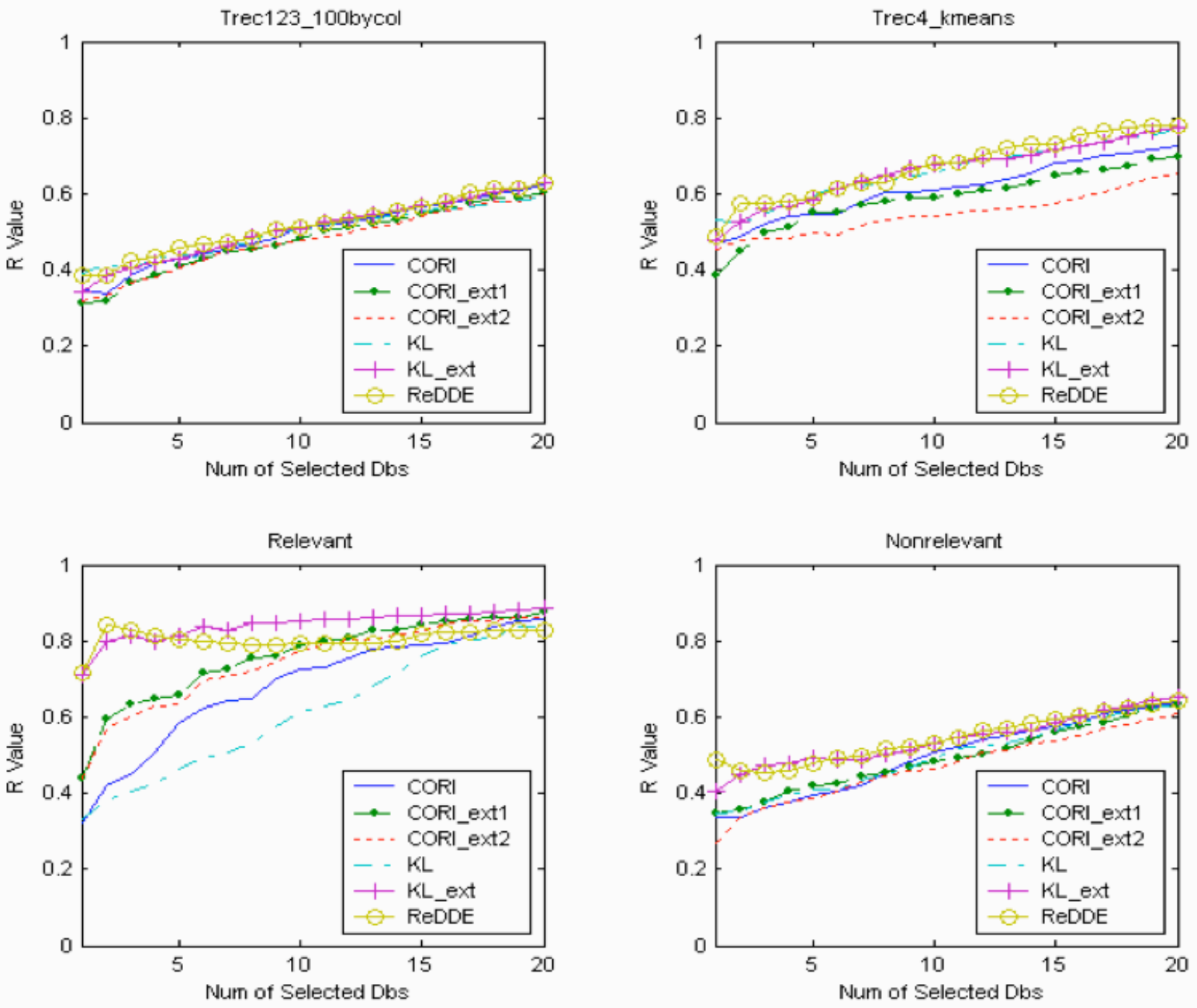


Fig. 1. The performance of the resource selection algorithms on four different testbeds.

Querylog-based Collection Selection

- Using Query Logs to Establish Vocabularies in Distributed Information Retrieval
- Milad Shokouhi; Justin Zobel; S.M.M. Tahaghoghi; Falk Scholer
- The collections are sampled using data from a query log
- Before: Queries over a dictionary (QBS)

UNIVERSITÀ DI PISA

- Recall
  - Number of found documents (not for web)
- Precision at X (P@X)
  - Number of relevant documents out of the first X results (X= 5, 10, 20, 100)
- Average precision
  - Average of Precision for increasing X
- MAP (Mean Average Precision)
  - The average over all queries

UNIVERSITÀ DI PISA

Table 1
Comparison of the QL and QBS methods on a subset of the WT10g data; QL consistently performs better. Differences that are statistically significant based on the t-test at the 0.05 and 0.01 level of significance are indicted by † and ‡ respectively. "CO" is the cutoff number of servers from which answers are retrieved.

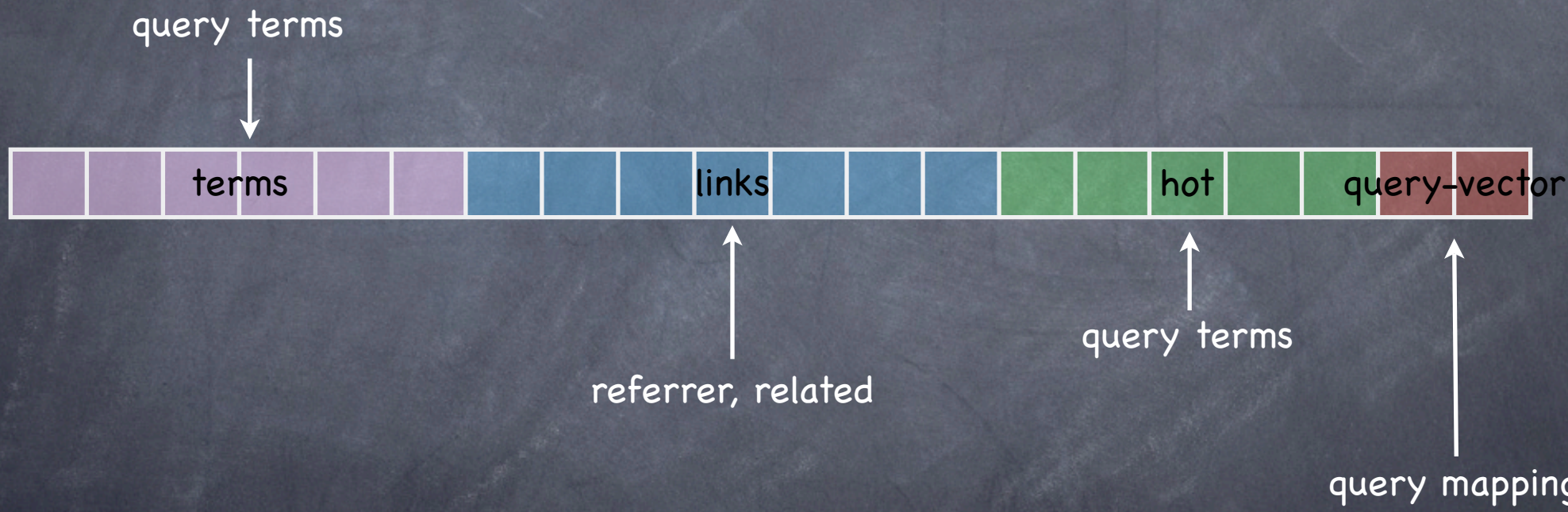
CO	MAP		P@5		P@10		R-Precision	
	QBS	QL	QBS	QL	QBS	QL	QBS	QL
1	0.0668	0.0902	0.1302	0.1721	0.0744	0.0988	0.0744	0.0988
10	0.1562	0.2515‡	0.3057	0.4322‡	0.2011	0.3023‡	0.2011	0.3023‡
20	0.1617	0.2811‡	0.3149	0.4621‡	0.2115	0.3437‡	0.2115	0.3437‡
30	0.1540	0.2655‡	0.2941	0.4471‡	0.2106	0.3259‡	0.2106	0.3259‡
40	0.1812	0.2639‡	0.3200	0.4306‡	0.2459	0.3212‡	0.2459	0.3212‡
50	0.1868	0.4188‡	0.3341	0.4188†	0.2506	0.3176‡	0.2506	0.3176‡

And now for something completely different!

Important features

- It can use any underlying WSE
  - Links, snippets, anchors...
- Good or bad as the WSE it uses!
- Small footprint
  - It can be added as another index

UNIVERSITÀ DI PISA


query terms
terms
links
hot
query-vector
referrer, related
query terms
query mapping

Developments

- Query suggestions
  - The system finds related queries
- Result grouping
  - Documents are already organized into groups
- Query expansion
  - Can find more complex queries still matching

UNIVERSITÀ DI PISA

	d1	d2	d3	d4	d5	d6	d7	d8
Q1	x						x	
Q2		x				x		
Q3		x				x		x
Q4			x				x	
Q5	x	x	x	x		x	x	x
Q6		x		x		x		
Q7	x			x		x		x
Q8		x		x			x	
Q9								

	d1	d2	d3	d4	d5	d6	d7	d8
Q1	x						x	
Q2		x				x		
Q3		x				x		x
Q4			x				x	
Q5	x	x	x	x		x	x	x
Q6		x		x		x		
Q7	x			x		x		x
Q8		x		x			x	
Q9								

	d1	d2	d3	d4	d5	d6	d7	d8
Q1	x						x	
Q2		x				x		
Q3		x				x		x
Q4			x				x	
Q5	x	x	x	x		x	x	x
Q6		x		x		x		
Q7	x			x		x		x
Q8		x		x			x	
Q9								

Still missing

- Using the QV model as a full IR model:
  - Is it possible to perform queries over this representation?
- New query terms cannot be found
  - CORI should be used for unseen queries
- Better testing with topic shift

UNIVERSITÀ DI PISA

Possible seminars/ projects

- Advanced collection selection
  - Statistical sampling, Query-based sampling
- Partitioning a LARGE collection (1 TB)
- Load balancing for doc- and term-partitioning
- Topic shift
  - Query log analysis

UNIVERSITÀ DI PISA