Distributed Information Retrieval: an approach based on harvesting

Fabio Crestani°

In collaboration with Fabio Simeoni*, Murat Yakici* and Steve Neely*

*University of Strathclyde, Glasgow, UK
*University College of Dublin, Dublin, Ireland
°University of Lugano, Lugano, Switzerland
outline

- Problem Domain
  - content-based wide-area distributed Information Retrieval
- Approach
  - from distributed retrieval to index harvesting via metadata harvesting
- Design Strategies
  - expanding the OAI-PMH infrastructure: protocol applications and protocol extensions
- Conclusions
  - where next?
problem domain

- Information Retrieval (IR)
  - primary content, not metadata
  - unstructured content and queries
  - queries evaluated probabilistically, not deterministically

- Distributed IR (DIR)
  - content is distributed across mutually remote collections

- Wide-Area DIR
  - content collections are widely dispersed
    - latencies, bandwidth fluctuations, network failures, connectivity issues
  - content collections are autonomously managed
    - disparity of strength and motivations
distributed retrieval contd.

- **strategy**
  1. distribute process across its inputs
     - ‘push’ queries towards collections
  2. centralise remotely produced outputs
     - ‘pull’ results of local query executions

- **two phases**
  - synchronous
  - real-time wrt user interaction

- **common assumptions**
  - brokered client/server architectures
  - textual content
distributed retrieval contd.
distributed retrieval

contd.

- considerable amount of research
  - collection description, collection selection, result fusion
    - cooperative and uncooperative techniques
  - test-beds & evaluation

- hot areas
  - from client/server to peer-to-peer architectures
    - hybrid, multi-tiered
  - from textual to multi-media content
    - cf. MIND and PENG Projects
  - from ad-hoc to GRID-enabled infrastructures
    - cf. the DILIGENT Project

- applications
  - metasearch engines on the Web
  - Federated Digital Libraries
metadata harvesting

- from Z39.50…to the OAI-PMH

- strategy
  - centralise input in advance of process execution
    - incrementally and iteratively
  - execute process against its input
    - locally

- two phases
  - asynchronous
  - one batch, one real-time wrt user interaction

- common assumptions
  - input: manually authored, descriptive metadata records
  - queries: fielded and deterministically evaluated
metadata harvesting contd.
metadata harvesting contd.

- **technical advantages:**
  - wide-area not observable during service provision
    - consistency, reliability, responsiveness, effectiveness, generality, simplicity
    - encourages medium, medium-large scalability

- **sociological advantages:**
  - *data providers*: greater visibility
    - without cost of full service provision
    - even for sensitive and dynamically published data
  - *service providers*: wider reach

- **disadvantages:**
  - minimal cooperation required
  - input potentially stale

- …and yet a common assumption in large-scale DL developments
index harvesting

- strategy
  - centralise content statistics automatically generated at data providers…
    - e.g. term histograms
    - possibly filtered (e.g. stopword removal)
    - possibly normalised (e.g. stemming)
    - incrementally and iteratively
    - according to some exchange model
  - …as well as descriptive metadata records
    - according to some exchange model
  - ingest both into local index at service provider
    - possibly normalising statistics wrt to current index statistics
    - possibly enhancing/normalising metadata records
  - execute queries at service provider
    - against local index of remote collections
    - using the harvested metadata to present query results
index harvesting contd.
index harvesting 

- between distributed retrieval and content crawling
  - some process is distributed…but indexing not retrieval
    - reap benefits of metadata harvesting
  - some data is centralised…but content statistics not content
    - more efficient bandwidth consumption
    - reduced load at data and service providers

- expand scope of content-based DIR research
  - content distribution need no longer imply distribution of retrieval or centralisation of content

- complement existing harvesting-based DL services
  - from metadata-based services to content-based services
  - leveraging the OAI-PMH infrastructure
    - a protocol application
    - a protocol extension
OAI-PMH recap

- client/server protocol for exchange of self-describing data
- 6 requests available to clients
  - 3 auxiliary requests, to discover server capabilities (Identify, ListMetadataFormats, ListSets)
  - 2 primary requests, to solicit data according to capabilities (GetRecord, ListRecords, ListIdentifiers)
- support for incremental harvesting
  - based on data time-stamping
- support for selective harvesting
  - based on hierarchies of potentially overlapping datasets
- support large data transfers in the face of transaction failures
  - simple session management mechanism based on resumption tokens
infrastructural issues outside protocol semantics
  - authentication, load balancing, compression, etc. resolved in a broader scope (e.g. at HTTP level)

abstract data model
  - servers maintain repositories of resources
  - resources have 1+ abstract descriptions, or items (basic unit of identification)
  - descriptions have 1+ format-specific instantiations, of records (basic unit of exchange and time-stamping)
    - support for Dublin Core mandatory
applying the OAI-PMH

- application strategy
  - extended data model
    - resources have at least one digital and text-based manifestation
    - one such manifestation, the *primary manifestation*, represents the resource content for harvesting purposes
  - dedicated format
    - for manually authored metadata *and* content statistics
    - statistics extracted from primary manifestation
- appealing solution…
  - no change to protocol and its development infrastructure
  - may serve immediately specific communities
- …but ad-hoc
  - different format for any combination of metadata and content statistics formats
  - need more modular, infrastructural approach
extending the OAI-PMH

- extension strategy
  - retain extended data model
  - identify metadata and content statistics independently
    - a record has now both a ‘metadata part’ and an ‘index part’
  - requests specify desired formats for both parts

- extension elements
  - extra auxiliary request `ListIndexFormats`
    - mirrors `ListMetadataFormats`
  - extra primary request parameter `indexPrefix`
    - mirrors `metadataPrefix`
  - extra `<index>` element to server responses
    - follows `<metadata>` element
  - sample format `tf_basic` for the index part
    - captures name and frequency of occurrence of indexing terms
evaluation

- proof-of-concept prototype
  - extensive testing
  - release of extended PMH to the OAI community
- testing
  - used the Aquaint TREC corpus across two institutions in different countries
    - tested the emulated heterogeneity of collections
    - tested the behaviour of incremental and periodical harvesting
  - efficiency
    - very small difference in resources required to index the global collection wrt index the harvested index data
  - effectiveness
    - same level of effectiveness of the global collection
conclusions

- **main points**
  - the harvesting model may be profitably applied to content-based retrieval
    - or there exists appealing middle ground between distributed retrieval and content crawling
  - the OAI-PMH infrastructure may be profitably leveraged for the purpose
    - immediately, via a protocol application
    - flexibly, via a protocol extension

- **future work**
  - exploit 2-phase model for asynchronous resource discovery in mobile and context-aware computing
question?

- more detailed presentation