Harvesting for Full-Text Retrieval

Fabio Simeoni[1], Murat Yakici[1], Steve Neely[2], and Fabio Crestani[1]

[1] University of Strathclyde
[2] University College of Dublin
fabio.simeoni@cis.strath.ac.uk

Abstract. We propose an approach to Distributed Information Retrieval based on the periodic and incremental centralisation of full-text indices of widely dispersed and autonomously managed content sources.

Inspired by the success of the Open Archive Initiative's protocol for *metadata harvesting*, the approach occupies middle ground between: (*i*) the crawling of content, and (*ii*) the distribution of retrieval. As in crawling, some data moves towards the retrieval process, but it is statistics about the content rather than content itself. As in distributed retrieval, some processing is distributed along with the data, but it is indexing rather than retrieval itself. We show that the approach retains the good properties of centralised retrieval without renouncing to cost-effective resource pooling. We discuss the requirements associated with the approach and identify two strategies to deploy it on top of the OAI infrastructure.

1 Introduction

Our interest is in content-based retrieval of widely dispersed and autonomously managed text sources. This is the central problem of Distributed Information Retrieval (DIR) and, over the past ten years, it has been approached by distributing the process along with the data: queries have been 'pushed' towards the content and the results of their local execution have been centrally gathered and presented to the user. While peer-to-peer models of distribution have recently generated some research interest [3], the traditional DIR approach relies on the simple client/server architecture depicted in Fig.1 (cf. [2]).

Rather independently and over a longer period of time, the Digital Library community has also explored the potential of distributed retrieval in the practice of its information services. Here, retrieval has mainly been interpreted as a deterministic process defined against the explicit structure of descriptive and manually authored metadata. Nonetheless, queries and results have still been exchanged within the client/server architecture in Fig.1; the Z39.50 protocol [14], in particular, has standardised the syntax and semantics of such exchange.

Over the past five years, however, the DL community has progressively favoured the complementary approach of iteratively and incrementally centralising metadata as a pre-condition to the retrieval of the associated data: metadata has been 'moved' towards the queries in advance of their execution (see Fig.2).

E.A. Fox et al. (Eds.): ICADL 2005, LNCS 3815, pp. 204–213, 2005.

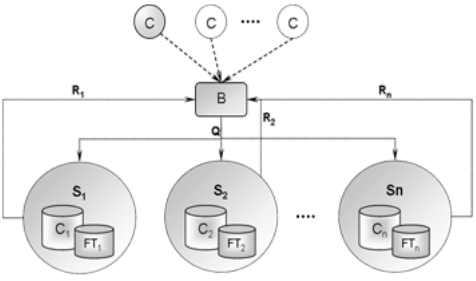

Fig. 1. Client/server distributed retrieval. A search broker B interfaces clients C and dispatches their queries Q to a number of autonomous search engines $S_1, S_2, , S_n$, each of which executes it against an index FT_i of some content C_i before returning results R_i back to B which merges them and relays them to C. Optionally, B optimises query distribution by selecting a subset of the engines based on previously gathered descriptions of their content.

Standardised by the *OAI-PMH*, the Protocol for Metadata Harvesting of the Open Archive initiative (OAI) [9], the *harvesting* model has proved particularly suitable to meet the technical and sociological requirements of retrieval within large-scale Federated Digital Libraries (FDLs)(e.g. [7]). A principled analysis of such success is found in [10] and may be summarised it here as follows.

From a technical perspective, harvesting eliminates the network as a real-time observable of service provision and, with it, a major obstacle to its medium-large scalability within wide-area networks [8]. Bandwidth fluctuations induced by traffic congestions and latency-inducing factors associated with slow, unavailable, or particularly distant data sources have no impact on the continuity, reliability, responsiveness, and even effectiveness of service provision. Retrieval, in particular, may regain the simplicity, generality, and QoS guarantees which are normally associated with local computations.

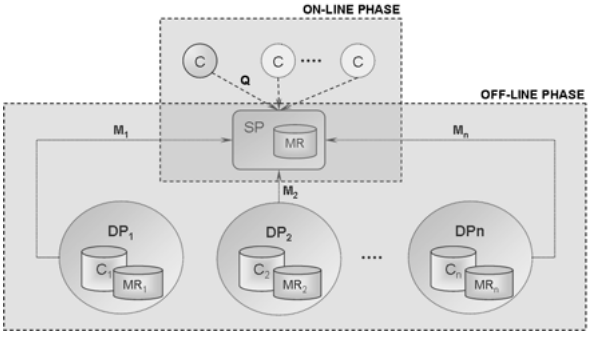

Fig. 2. Metadata harvesting. (a) *off-line phase*: a service provider SP gathers metadata M from data providers $DP_1, DP_2, , DP_n$ and stores it in a metadata repository MR; (b) *on-line phase*: SP interfaces clients C and resolves their queries Q against MR.

From a sociological perspective, the model captures the disparity of strengths and interests which characterises FDLs; in particular, it clearly distinguishes the roles, responsibilities, and costs of *service providers* from those of *data providers*. Data providers may give broad visibility to their data without having to face the complexity of full service provision; service providers also benefits from simplified participation, for the scope and usefulness of their services may scale beyond previously experienced bounds [6].

1.1 Scope and Motivations

In this paper, we investigate the applicability of the harvesting model to full-text retrieval.

The motivation is two-fold. Firstly, we hope to expand the scope of Distributed Information Retrieval beyond the assumptions which have bound it so far. Secondly, we aim to extend the benefits of the harvesting model within the same domains which to date have successfully but only partially adopted it. Within today's FDLs, a reconciliation of harvesting with full-text retrieval would guarantee homogeneous scope and QoS across both metadata-based and content-based services; using the OAI-PMH for the purpose, in particular, would immediately leverage a widely deployed infrastructure of tools and providers.

Under a generic interpretation, of course, the applicability of harvesting to content-based retrieval need not be questioned: any Web search engine stands as a witness of the feasibility and scalability of moving data towards the retrieval process. Here, however, we focus on a stricter but more advantageous interpretation of harvesting in which retrieval remains predicated on the sole movement of metadata. However, we now give to metadata the technical meaning which it normally assumes in Information Retrieval, and thus focus on automatically generated content statistics rather than manually authored descriptive records. In particular, we assume that the content remains distributed and that a full-text index of the union of the distributed sources is instead centralised[1]. By doing so, we expect to make better use of shared bandwidth and to reduce load at both data and service providers. We also hope to promote scope, for the approach may offer visibility to data which is neither statically published nor publicly accessible; data which is proprietary, costs money, demands access control, or is simply dynamically served, may still be safely disseminated.

Overall, we shift the assumption of distribution from the retrieval process to the indexing process, and thus explore the existence of middle ground between distributed retrieval and content crawling. In doing so, we are guided by the following research questions: can we distribute and incrementally execute the indexing process? And from a more practical perspective: can we leverage the OAI infrastructure for the purpose? We address these questions in Section 2 and Section 3, respectively. We discuss related work in Section 4 before drawing some conclusions in Section 5.

[1] Interestingly, client-server retrieval already relies on a harvesting approach whenever it centralises collection-level descriptions for the purposes of selective query distribution.

2 The Approach

We use an example to clarify the approach and identify the requirements it raises at both ends of the exchange model.

2.1 Harvesting Scenarios

In the standard harvesting scenario, a service provider relies on the OAI-PMH to periodically centralise descriptive metadata from a number of data providers. Independently from dissemination agreements, the providers maintain their metadata in databases and use it routinely to offer local services to their users, including a structure-based retrieval service; some providers also maintain full-text indices on their file systems and use them to complement the retrieval service with keyword-based queries. Models and languages for metadata, indexing, and retrieval are locally defined and locally maintained. At each provider, a dissemination service implements the server side of the OAI-PMH and resolves protocol requests by: (i) executing a fixed range of queries against the metadata database, and (ii) mapping the results expressed in the local metadata model onto instances of a model agreed upon for exchange, say unqualified Dublin Core (DC) [5]. At the service provider, the DC records are normalised and otherwise enhanced (e.g. duplicates are removed and subjects are automatically inferred), and then added to the input of an interactive retrieval service. The service accepts structure-based as well as content-based queries, but it executes both types of query against the harvested DC records.

We propose an extension of the previous scenario in which the descriptive metadata exposed by data providers is augmented with content statistics (see Fig. 3). The providers obtain this information from pre-existing or dedicated full-text indices, rather than databases, but they still map records onto an exchange model. Similarly, at the service provider, the statistics are extracted and used to update a local full-text index, possibly after having been normalised and enhanced to reflect current content statistics and local indexing requirements, respectively. The index is then used to satisfy full-text queries while the descriptive metadata supports the presentation of results. For flexibility, (subsets of) the same content statistics may be used to support more than one model of retrieval (e.g. a vector space model and a language model).

2.2 Requirements

From a conceptual perspective, the extension is relatively straightforward. Its only requirement is for the service provider to rely on a model of indexing which allows modular representation of content over space and time. More formally:

> **(Modular Indexing)** If M is an indexing model, C_0 and C_1 two content sources, and I_0 and I_1 their M-indices, then M is *modular* if the difference $\Delta C = C_1 - C_0$ implies a difference $\Delta I = I_1 - I_0$ such that ΔI is computable from I_0 and ΔC only.

Interpreted along a spatial dimension, modularity guarantees the distributivity of the indexing process across independently maintained content; interpreted along a temporal dimension, it guarantees the incremental nature of such process. In turn, modularity is guaranteed by content properties whose measurement may be distributed over document-grained increments. Common indexing models satisfy this requirement, for they either rely on term-related properties which pertain to individual documents – such as in-document term number, frequency, and location – or else pertain to groups of documents and yet may still be progressively derived, such as inverse document frequency [13].

From a pragmatic perspective, however, the enriched semantics of the exchanged data unavoidably adds development complexity and resource consumption. Most noticeably, it assumes data providers which are: (i) sufficiently sophisticated to offer integrated management of descriptive metadata and full-text indices, and (ii) sufficiently rich to sustain the load on computational resources – from storage to memory and network bandwidth – which is induced by the increased size of (per-document) content statistics over descriptive metadata.

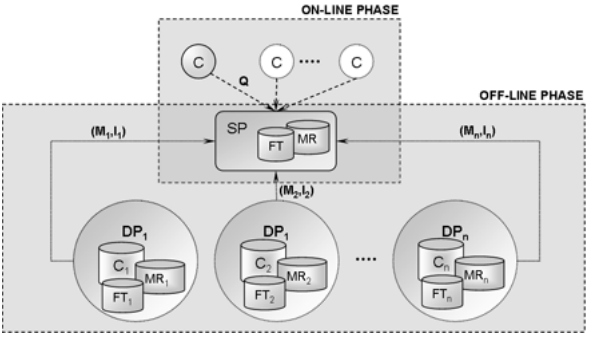

Fig. 3. Full-text index harvesting. (a) *off-line phase*: a service provider SP gathers pairs (M_i, I_i) of metadata and content statistics from data providers $DP_1, DP_2, , DP_n$ and stores them in a metadata repository MR and a full-text index FT, respectively; (b) *on-line phase*: SP interfaces clients C, resolves their queries against FT, and uses the metadata in MR to present the results.

Clearly, issues of data integration and size concern both ends of the exchange scenario. On an absolute scale, problems may seem more acute at the client side of the protocol but the harvesting philosophy indicates that the server side is where adoption and scalability may be more obviously at stake. In particular, data providers must accommodate the cost of generating, maintaining, and serving full-text indices within their resource allocation policies; whenever such costs may not be directly justified in terms of local requirements – i.e. in the assumption of dedicated development – then it may prove too expensive to accommodate the novel dissemination requirements. Cost estimates will vary from case to case and only deployment experience may indicate what level of tool support may help reducing complexity.

As to the issue of size, we expect compression to play an important role at both ends of the protocol. Lossless compression techniques based on optimised representation structures are the first obvious choice, be it for the persistent storage of indices, their in-memory management, or their transfer on the wire[2]. In harvesting, furthermore, compression ratios may be pushed further than they may when decompression is a real-time observable of service provision. Lossy compression techniques may also be conveniently used to complement lossless approaches. Well-known algorithms – ranging from standard case folding, stopword removal, and stemming algorithms, to static index pruning and document summarisation algorithms (e.g. [4]) – may all grant additional size reductions without excessively compromise the final quality of retrieval.

One last pragmatic question concerns the suitability of the OAI-PMH to support the extended exchange semantics. We dedicate the next Section to a possible answer.

3 Protocol Design

We first summarise the features of the OAI-PMH and then assess two strategies to deploy the extended exchange semantics on top of the existing OAI infrastructure.

3.1 OAI-PMH

At its heart, the OAI-PMH is a client-server protocol for the selective exchange of self-describing data. Six types of requests are available to clients: three to discover capabilities of servers (*auxiliary requests*) and three to solicit data from servers in accordance with their capabilities (*primary requests*). To support incremental harvesting, primary requests may be temporally scoped with a granularity of days or seconds; selective harvesting relies instead on the optional definition of a hierarchy of potentially overlapping datasets. Simple session management mechanisms support large data transfers in the face of transaction failures. For ease of deployment, the overall semantics of exchange – including error semantics – is 'tunnelled' within HTTP's, while XML provides syntax and high-level semantics for response payloads.

The exact semantics of the exchanged data is formally undefined but, by design, it is expected to fall within the domain of content metadata; indeed, all servers are required to produce DC metadata on request. In particular, an exchange model associates servers with repositories of *resources* and resources with one or more metadata descriptions, or *records*; the latter form the basic unit of exchange. The model says little about resources but it offers a layered model of metadata in which records are format-specific instantiations of fully abstract resource descriptions, or *items*. The identification of items and formats is explicit; the protocol suggests an implementation scheme for item identifiers (e.g.

[2] Transport-level compression, in particular, is already within the scope of standard OAI-PMH exchange semantics.

`oai:dp:hep-th/9901001`) and defines an extensible lists of format identifiers (e.g. `oai_dc` for the required DC). Individual records are instead implicitly identified by their format and the item they instantiate; they are nonetheless explicitly associated with datestamps and thus may change independently from their items. As an example of OAI-PMH data exchange, the following HTTP GET request:

```
http://www.dp.org/oai?
verb=ListRecords&MetadataPrefix=oai_dc&from=2005-01-01
```

asks a server available at `http://www.dp.org/oai` to return all the DC records which have changed since the beginning of the current year. The following is a sample response[3]:

```
<OAI-PMH>
  <responseDate>2005-01-01T19:20:30Z</responseDate>
  <request verb="ListRecords" from="2005-01-01"
    metadataPrefix="oai_dc">http://www.dp.org/OAI</request>
  <ListRecords>
    ...
    <record>
      <header>
        <identifier>oai:dp:hep-th/9901001</identifier>
        <datestamp>2005-02-18</datestamp>
      </header>
      <metadata>
        <dc>
          <title>Opera Minora</title>
          <creator>Cornelius Tacitus</creator>
          <identifier>http://www.dp.org/res/9901001.html</identifier>
          ...
        </dc>
      </metadata>
    </record>
    ...
  </ListRecords>
</OAI-PMH>
```

3.2 Design Strategies

The increasing popularity of the OAI-PMH has generated some interest in using the protocol beyond its original design assumptions. Building on the generality of the data model, original use has sometimes been predicated on creative instantiations of the modelling primitives [11]. In other cases, the exchange semantics has been extended to accommodate additional functionality (e.g. [12]). Both design routes are available for our protocol: we could conceive it as an *application* or as an *extension* of the OAI-PMH.

The first solution may be simply predicated on: (i) a specialisation of the protocol's data model, and (ii) the definition of a dedicated format for the integrated exchange of descriptive metadata *and* content statistics. The model would simply introduce constraints on the notion of resource, namely: (a) resources have at least one digital and text-based manifestation, and (b) a distinguished manifestation, the *primary manifestation*, satisfies (a) and is designated to represent the content of the resource for harvesting purposes. The format would instead bind descriptive metadata and content statistics of primary manifestations to

[3] For clarity, namespace information is omitted in this and following examples.

individual request/response interactions, so as to avoid the synchronisation problems which may arise if each form was harvested independently from the other. The solution is appealing for it proves the concept whilst requiring no change to the protocol and its deployment infrastructure. While it may immediately serve the needs of specific communities, however, its design is rather ad-hoc and requires the definition of dedicated formats for each variation in the shape of descriptive metadata and/or content statistics. This induces a 'combinatorial' approach to standardisation which may unnecessarily compromise interoperability across communities of adoption.

To illustrate the full potential of the approach, we concentrate instead on the definition of a more modular exchange mechanism which may gracefully accommodate arbitrary forms of descriptive metadata and content statistics. Specifically, we retain the data model specialisation defined above, as well as the binding of metadata and content statistics within individual request/response interactions. However, we now identify each form of data independently from the other and thus assume that a record includes both a metadata part and an index part. In particular, we expect requests to specify a format for the metadata part and a format for the index part.

This leads to a protocol extension defined by: (i) the addition of an auxiliary request `ListIndexFormats` with associated response format; (ii) the addition of an optional parameter `indexPrefix` to primary requests; and (iii) the addition of an optional `index` child to the `record` elements contained in responses to primary requests. `ListIndexFormats` is used to discover the index formats supported by servers, and as such it extends the semantics of `ListMetadataFormats`. Similarly, `indexPrefix` specifies the format of the index part of records and thus mirrors `metadataPrefix` and its associated error semantics. Finally, `index` elements contain the index part of records and follow the standard `metadata` elements.

The extension of the sample request/response pair shown in Sect. 3.1 may then be the following::

```
http://www.dp.org/oai?
verb=ListRecords&metadataPrefix=oai_dc&indexPrefix=tf_basic&from=2005-01-01

<OAI-PMH>
  ...
  <ListRecords>
    ...
    <record>
      ...
      <metadata>
        <dc>...</dc>
      </metadata>
      <index>
        <terms>
          ...
          <term name="opera" freq="26">
          <term name="minora" freq="36">
          ...
        </terms>
      </index>
    </record>
    ...
  </ListRecords>
</OAI-PMH>
```

Here, `tf_basic` is the identifier of a simple format which captures the name and frequency of occurrence of the terms chosen to represent primary manifestations (possibly after stemming and stop-word removal). The underlying model serves the purpose of a proof of concept but supports most of the indexing models which may be employed at the client side. Variations are of course possible; for example, a format which captures only term names and document lengths would decrease resource consumption and still support simple models of boolean retrieval. On the other hand, a model which includes positional information for each term occurrence would increase resource consumption but also support proximity searches at the client side.

4 Related Work

The relationship between the proposed approach, distributed retrieval, content crawling, and existing implementations of the harvesting model has been extensively discussed in previous Sections. Here, we concentrate on what - to the best of our knowledge - is the only work which directly shares some of our motivations.

The Harvest system [1] was initially proposed in the mid-nineties as a sophisticated end-to-end solution for content-based retrieval over the inter-network. Harvesting is a central component of the system's architecture and its technical contribution to the OAI initiative has been repeatedly acknowledged in the literature. Unlike the OAI-PMH, however, the system abstracts over the precise semantics of the harvested data, which may range from manually authored, descriptive metadata, to automatically derived, and type-specific content statistics. Text-based formats, in particular, are processed along lines similar to those advocated in this paper.

Our work, however, frames the approach within an evolved infrastructural context, where later developments - particularly XML and the role-based model OAI-PMH itself - are leveraged towards a more general data exchange mechanism than what may be found buried within a closed system. In particular, we operate in a context in which interoperability is predicated on protocol-based solutions, rather than end-to-end implementations. Further, Harvest focuses on the indexing of type-specific content summaries, which represents just one of many possible applications of the approach. Overall, our work motivates, contextualises, and generalises the good properties of an architectural model which has been previously implemented and yet has to receive widespread acceptance.

5 Conclusions

A topological separation between the processes of indexing and retrieval suits DIR systems in which content is widely distributed and autonomously managed. Indexing is conceptually distributed along with the content and remains the only responsibility of content providers; located elsewhere on the network, retrieval is centralised around a periodic and incremental harvest of the indexes produced at

each provider. A protocol-based infrastructure for harvesting descriptive metadata in support of structured retrieval has already been widely and successfully deployed and we have shown how it may be leveraged for full-text retrieval. As a proof-of-concept, we have tested the approach in a prototype for multi-model retrieval of distributed and potentially unmanaged file collections; due to lack of space, however, we leave a report on the implementation to future work.

References

1. Bowman, C.M., Danzig, P.B., Hardy, D.R. et al.: Harvest: A Scalable, Customizable, Discovery and Access System. Technical Report TR CU-CS-732-94, Department of Computer Science, University of Colorado-Boulder, 1994.
2. Callan, J.: Distributed information retrieval. In W.B. Croft, editor, Advances in information retrieval, chapter 5, pages 127-150. Kluwer Academic Publishers, 2000.
3. Callan, J., Fuhr, N., Nejdl, W. (Eds.): Proceedings of the SIGIR Workshop on Peer-to-Peer Information Retrieval, 27th Annual International ACM SIGIR Conference, July 29, 2004.
4. Carmel, D, Cohen, D. et al.: Static Index Pruning for Information Retrieval Systems. In Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval, pages 43-50, 2001.
5. The Dublin Core Metadata Initiative: Dublin Core Metadata Element Set, Version 1.1: Reference Description, 2004 (http://dublincore.org/documents/dces/).
6. Lagoze, C., Van de Sompel, H.: The Open Archives Initiative: Building a low-barrier interoperability framework. JCDL '01: Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries, 2001.
7. Lagoze, C., Hoehn, W., Arms, W., Allan, J. et al.: Core Services in the Architecture of the National Digital Library for Science Education (NDSL). Cornell University, Ithaca, arXiv Report, cs.DL/0201025, 2002.
8. Lynch, C.: The Z39.50 Information Retrieval Standard: Part I: A Strategic View of Its Past, Present, and Future. In D-Lib Magazine, April 1997 (http://www.dlib.org/dlib/april97/04lynch.html).
9. The Open Archives Initiative: The Open Archives Initiative Protocol for Metadata Harvesting (2.0), 2003 (http://www.openarchives.org/OAI/openarchivesprotocol.html).
10. Simeoni, F.: Servicing the Federation: the Case for Metadata Harvesting. In ECDL '04: Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science 3232, Springer, 2004.
11. Van de Sompel, H., Young, J., Hickey, T.:Using the OAI-PMH...Differently. In D-lib Magazine, July/August 2003.
12. Suleman, H., Fox, E.:Designing Protocols in Support of Digital Library Componentization. In ECDL'02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, pages 568-582, 2002.
13. Witten, I., Moffat, A., Bell, T. Managing Gigabytes: Compressing and indexing documents and images. Van Nostrand Reinhold. 1994.
14. Z39.50 Maintenance Agency: Information Retrieval (Z39.50): Application Service Definition and Protocol Specification, 2003.