Building and Experimenting with a Heterogeneous Collection

Zoltán Szlávik and Thomas Rölleke

Queen Mary University of London, London, UK
{zolley, thor}@dcs.qmul.ac.uk

Abstract. Today's integrated retrieval applications retrieve documents from disparate data sources. Therefore, as part of INEX 2004, we ran a heterogeneous track to explore the experimentation with a heterogeneous collection of documents. We built a collection comprising various sub-collections, re-used topics (queries) from the sub-collections and created new topics, and participants submitted the results of retrieval runs. The assessment proved difficult, since pooling the results and browsing the collection posed new challenges and requested more resources than available. This reports summarises the motivation, activities, results and findings of the track.

1 Introduction

A heterogeneous track has been part of INEX 2004. The task of the track was to explore how to build and maintain a testbed, how to create topics, and how to perform retrieval runs, assessment and evaluation.

1.1 Motivation

Before 2004, the INEX collection has been a collection of XML documents with a single DTD. However, in practical environments, XML retrieval requires to deal with XML documents with different DTDs, because a collection comprises documents of different purpose, authors and sources. Further, information in practical environments is spread over XML documents, relational databases, and other data source formats. Therefore, we included in INEX 2004 a heterogeneous track (also known as het track) that addressed the heterogeneity of a collection.

A heterogeneous collection poses a number of challenges:

- For content-only (CO) queries, approaches for homogeneous and well-typed collections can make direct use of the DTD. The DTD can be used, for example, for identifying what element type is reasonable to present in the retrieval result. In a heterogeneous collection, we might have several or no DTD's, and retrieval methods independent of DTD are essential, and DTD mappings might be useful.
- For content-and-structure (CAS) queries, there is the problem of mapping structural conditions to different sub-collections. If we consider structural

N. Fuhr et al. (Eds.): INEX 2004, LNCS 3493, pp. 349–357, 2005.

conditions as useful, then a DTD-based mapping of structural conditions is essential for CAS queries. Methods known for federated databases could be applied here. We can distinguish between manual, semi-automatic or fully automatic methods for creating the schema mappings.
- When performing retrieval runs, the retrieval algorithms need to merge the results retrieved from different sub-collections. For an experimental point of view, we can compare global strategies that know the whole collection with local strategies that make only use of the knowledge that can be derived per sub-collection. The latter strategies are probably closer to what we meet in reality.
- The content of a relational database can be represented in an XML document (collection, respectively). The question is whether the retrieval of relational databases via XML is beneficial.

The goal of the INEX het track was to set up a test collection, and investigate the new challenges.

This track aims to answer, among others, the following research questions:

- For CO queries, what methods are feasible for determining elements that would be reasonable answers? Are pure statistical methods appropriate and sufficient, or are ontology-based approaches also helpful?
- What methods can be used to map structural criteria such that they can be applied (make sense) for a collection for which the DTD might be different or even not known!?
- Should mappings focus on element names (types) only, or also deal with element content?
- Should the data be organized (and indexed) as a single collection of heterogeneous documents, or is it better to treat het coll as a set of homogeneous sub-collections?
- Are evaluation criteria developed for homogeneous collections also suitable for heterogeneous collections, or should other criteria and metrics be applied?

Since this was the first year of the heterogeneity track, the focus of the activities was on making a test collection available to participants, create some topics and perform retrieval runs and assessment, and apply evaluation measures.

The emphasis was on investigating the How to do it, with a detailed look at individual topics and runs, and the technicalities involved. A statistical measure was not the aim of the first year of het track.

1.2 Activities

The participants of this track carried out the following activities:

- Construction of a heterogeneous test collection (sometimes called het coll): We used the current INEX corpus, and added various sub-collections including DBLP, HCIBIB, Berkeley lib, Duisburg bibdb, and QMUL bibdb (the latter an XML representation of a relational database). The collection is maintained at http://inex.is.informatik.uni-duisburg.de:2004/internal/hettrack/.

- Selection of 20 CO and CAS queries from the existing INEX body and creation of four new topics. The topics were selected and created with the aim to retrieve documents from several sub-collections.
- INRIA has developed and experimented with a tool, XSum, for graphically representing XML documents; one of the main purposes of the tool was to enable the user to grasp the structure and aspect of various XML datasets, with or without a DTD.[1]
- Retrieval runs on the heterogeneous collection for this set of queries (see appendix).
- The assessment has been not carried out yet, due to technical problems and restricted resources. The aim is to join the het coll with the relevance assessment tool used for the INEX IEEE collection.
- For the evaluation, we aim at a qualitative (query-and-run-oriented) analysis rather than a quantitative average-oriented analysis of results.

Based on the results and experience gained in 2004, a larger and quantitative het track can be carried out in following years.

2 Collection Creation

Table 1 shows the sub-collections that were used this year.

Table 1. Sub-collections used in 2004

Collection	MB(unpacked)	Number of elements
IEEE Computer Society	494	8.2M
Berkeley	33.1	1194863
CompuScience	313	7055003
bibdb Duisburg	2.08	40118
DBLP	207	5114033
hcibib	30.5	308554
qmul-dcs-pubdb	1.05	23436

From creating the sub-collections, we learned the following:

1. For a larger scale het track, methods and tools are needed for managing a set of sub-collections. With restricted resources, the management of 5-10 sub-collections is achievable but more sub-collections require tools and resources.
2. Sub-collections come with syntax errors (non-tidy XML). It is best to correct those errors centrally and "by hand", but keep a carefully maintained log of the changes made.

[1] Currently, XSum represents the XML elements and attributes structure within an XML document, statistics such as numbers of elements on a given path. The tool is developed in Java, and freely available.

3 Topic Creation

Given the objectives of the het track, four types of topics have been proposed in the topic creation guideline:

1. CO (Content Only Topics): Since CO queries do not take structural information into account. This type had not been found challenging, but any CO query used in the ad-hoc track could be used in the het track and gave similar results (because the test collection used for the ad-hoc track is part of the het track).
2. BCAS (Basic Content and Structure Topics): This type of topics focuses on the combination of singular structural constraints with a content-based constraint. The aim is synonym matches for structural constraints.
3. CCAS (Complex Content and Structure Topics): are the het track equivalent of the CAS topics of the ad-hoc track, specified used the NEXI language. The aim is to enable transformations and partial mappings of the topic path upon the different collections in het track, without losing the IR component of the topic.
4. ECCAS (Extended Content and Structure Topics): extended CCAS to enable the specification of the correctness path transformation and mapping probabilities.

3.1 Re-used Topics

Twenty topics were selected from the ad-hoc topics to re-use in het track. After examining the ad-hoc topics, 10 CO topics were selected that probably contain results not only in the IEEE (also referred to and used as inex-1.3 and inex-1.4) sub-collection. 10 CAS topics were also selected. The main criterion was that topics should possibly have relevant results in more sub-collections. Selected CAS topics were identified as CCAS het track topics.

3.2 New Topics

Four new topics (see B) were created by participants of which three topics are CCAS and one is BCAS.

4 Retrieval Runs

The main difference between a mono- and a heterogeneous track is that sub-collections are specified in the run submissions. In order to be able to examine results with respect to the considered sub-collections, a slightly modified version of the ad-hoc track's submission format has been proposed (see C).

Actually, the consideration of sub-collections poses some major research question, since we cannot assume that each run considers all sub-collections:

1. How do we pool results from runs if some runs considered a sub-collection X and other runs considered a sub-collection Y?
2. How does an evaluation measure deal with the incompleteness of runs?

Another issue is the assignment of topics to participants. Is it useful to assign topics under strict rules and supervision, trying to make sure that sub-collections are covered equally, and the same number of runs is performed per topic, etc? Or is it the nature of heterogeneous track that this effort is not justified and is rather to be replaced by a random assignment?

5 Assessment and Evaluation

During the preparation for assessment and evaluation, we identified the following two main challenges:

1. Browsing the results and the collection. The browsing tool X-Rai was initially developed for the IEEE collection only, and currently cannot handle larger sub-collection files, even the QMUL sub-collection with its 1.05MB, efficiently. Therefore, the two smallest sub-collections (bibdbpub and qmuldcs-dbpub) were converted into many small files, and made available for browsing.
2. Pooling. The aforementioned problem also affected the pooling procedure, as the format of submission runs could not be exactly used for pooling. The other challenge in pooling was that, unlike the ad hoc track runs, het track runs could consider various sets of sub-collections, and there has not been a straightforward method to create pools from this kind of source, e.g. "use the first 150 results in each run" method may create larger pools for sub-collections having more elements in the top-ranked results and small for those having less.

6 Summary and Conclusions

The first year of het track established a heterogeneous collection, reused and created topics, and performed retrieval runs. The assessment and evaluation is currently outstanding.

The discussion among the participants and the work carried out raised the following questions:

1. What makes the heterogeneity of a collection? The current het coll is viewed as little heterogeneous since it consists "only" of XML documents, and all documents are about computer science literature. Can we measure heterogeneity?
2. How can we manage many and large sub-collections? In particular creating the browsing facilities for the sub-collections and the assessment proved difficult. Can we easily split (and possibly merge files)?
3. Topics and retrieval runs relate only to some sub-collections. Topics might have been created and runs might have been performed without considering the whole collection. How is this incompleteness captured in an evaluation?

Het track has established a collection and experience about how to do it and where the difficulties are. INEX is now ready for the next phase of het track, and it can re-use and extend the existing collection and pay particular attention to the efficient inclusion of new sub-collections into the whole process.

A Topic Format

```
<!ELEMENT inex_topic (title,
   content_description,
   structure_description,
   narrative,keywords)>
<!ATTLIST inex_topic
  topic_id   CDATA  #REQUIRED
  query_type CDATA  #REQUIRED
>

<!ELEMENT title (#PCDATA)>
<!ELEMENT content_description   (#PCDATA)>
<!ELEMENT structure_description   (#PCDATA)>
<!ELEMENT narrative     (#PCDATA)>
<!ELEMENT keywords      (#PCDATA)>
```

B Het Track Topics

```
Topic created by IRIT:

<?xml version="1.0" encoding="ISO-8859-1"?>
<inex_topic topic_id="1" query_type="BCAS">
  <title>
    //bb[about(.,"PhD thesis amsterdam")]
  </title>
  <content_description>
    I'm looking for bibliography entries concerning
    PhD thesis obtained at the university of Amsterdam
  </content_description>
  <structure_description>
    I'm looking for full references of PhD thesis: it
    means that results elements should contain the author,
    the title,the year and the school/city where the PhD
    thesis was obtained.
  </structure_description>
  <narrative>
    I'm maybe interested in working in Amsterdam next year
    and I would like to know what are the research subjects
    in the city. I think that a way to obtained this information
    (in the collections we have) is to see what are the subjects
    of the PhD thesis obtained in Amsterdam.
  </narrative>
```

```
  <keywords>
    phD thesis, university, amsterdam
  </keywords>
</inex_topic>
```

Topic created by UMONTES:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<inex_topic topic_id="2" query_type="CCAS">
  <title>
    //article[about(.//author, nivio ziviani)]
  </title>
  <content_description>
    We are seeking for works with Nivio Ziviani as one of its authors
  </content_description>
  <structure_description>
    Title is a tag identifying works title and author is a
    tag identifying who wrote those works. They are usually part of
    front matter of a document, or part of  bottom matter  in a
    bibliography reference or can be an item in a volume index.
  </structure_description>
  <narrative>
    We are seeking for works with Nivio Ziviani as one of its
    authors. We want to catalogue all Nvio Ziviani works, so any
    reference, index entry , abstract or complete article will be
    relevant, but biography works  will not.
  </narrative>
  <keywords>
    Nivio Ziviani
  </keywords>
</inex_topic>
```

Topic created by RMIT:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<inex_topic topic_id="3" query_type="CCAS">
  <title>
    //article[about(.//abs, Web usage mining) or
    about(.//sec, "Web mining" traversal navigation patterns)]
  </title>
  <content_description>
    We are looking for documents that describe capturing and mining
    Web usage, in particular the traversal and navigation patterns;
    motivations include Web site redesign and maintenance.
  </content_description>
  <structure_description>
    Article is a tag identifying a document, which can also be
    represented as a book tag, an inproceedings (or incollection)
```

```
    tag, an entry tag, etc. Abs is a tag identifying abstract of
    a document, which can be represented as an abstract tag, an abs
    tag, etc. Sec is a tag identifying an informative document
    component, such as section or paragraph. It can also be represented
    as sec, ss1, ss2, p, ip1 or other similar tags.
  </structure_description>
  <narrative>
    To be relevant, a document must describe methods for capturing
    and analysing web usage, in particular traversal and navigation
    patterns. The motivation is using Web usage mining for site
    reconfiguration and maintenance, as well as providing recommendations
    to the user. Methods that are not explicitly applied to the Web
    but could apply are still relevant.
    Capturing browsing actions for pre-fetching is not relevant.
  </narrative>
  <keywords>
    Web usage mining, Web log analysis, browsing pattern,
    navigation pattern, traversal pattern, Web statistics, Web design,
    Web maintenance, user recommendations.
  </keywords>
</inex_topic>


Topic created by LIP6:

<?xml version="1.0" encoding="ISO-8859-1"?>
<inex_topic topic_id="4" query_type="CCAS">
  <title>
    //article[about(.,"text categorization") and
    (about(.//fm//au, "David D. Lewis")
    or about(.//bib//au, "David D. Lewis"))]
  </title>
  <content_description>
    I am looking for documents about text categorization which
    have been written by David D. Lewis, or related work from other authors.
  </content_description>
  <structure_description>
    The tags which are used in this topic come from the DTD of the
    ad hoc task collection. Article is a tag identifying a document,
    which can also be represented as a book tag, an inproceedings
    (or incollection) tag, an entry tag, etc. Fm is a tag identifying
    the header of a document which usually contains title, authors...
    Bib is a tag identifying the bibliography of a document.
    Au is a tag identyfying an author name.
  </structure_description>
  <narrative>
    To be relevant, a document must describe text categorization methods.
    It must have been written by David D. Lewis or must contain
    a bibliography entry with David D. Lewis.
  </narrative>
```

```
  <keywords>
    Text categorization, Text classifier
  </keywords>
</inex_topic>
```

C Run Format

```
<!ELEMENT inex_het_track_submission (description, topic+)>
<!ATTLIST inex_het_track_submission
participant-id CDATA #REQUIRED
run-id CDATA #REQUIRED
query (automatic | manual) #REQUIRED
topic-part  (T|D|K|TD|TK|DK|TDK)    #IMPLIED
task  CDATA  #IMPLIED
>

<!ELEMENT description (#PCDATA)>

<!ELEMENT topic (subcollections, result*)>
<!ATTLIST topic
topic-id CDATA #REQUIRED
>
<!ELEMENT subcollections (subcollection+)>
<!ELEMENT result (subcollection, file, path, rank?, rsv?)>
<!ELEMENT subcollection EMPTY>
<!ATTLIST subcollection name CDATA #REQUIRED>

<!ELEMENT file (#PCDATA)>
<!ELEMENT path (#PCDATA)>
<!ELEMENT rank (#PCDATA)>
<!ELEMENT rsv (#PCDATA)>
```

D Submitted Runs

- IRIT submitted 3 runs. One run is for CCAS, one for CO and one for BCAS topics. Files contain results for all the 24 het track topics. Various groups of sub-collections were considered for topics.
- RMIT submitted results of three different approaches, all approaches were applied to all topic types and topics (9 files - file groups of 3 - one file is for a specific approach, specific topic type (CCAS,CO,BCAS)). Various groups of sub-collections were considered for topics, often all sub-collections were used.
- UBERKELEY submitted 2 runs, used all CO topics, 12 (i.e. all but one) CCAS topics. All sub-collections were considered.
- UMONTES submitted 6 runs, 3 runs for all CO topics, 3 for all 'VCAS' (CCAS and BCAS together) topics, considered 5 sub-collections.
- UNIDU submitted 3 runs, considered only topic no. 1 (as CO) and used 3 sub-collections.