Report on the INEX 2006 Heterogeneous Collection Track

Ingo Frommholz
Faculty of Engineering Sciences
University of Duisburg-Essen
D-47048 Duisburg, Germany
ingo.frommholz@uni-due.de

Ray Larson
School of Information
University of California
Berkeley, California 94720-4600
ray@ischool.berkeley.edu

Abstract

XML collections that consist of subcollections from different sources pose a challenge with respect to syntactic, semantic and genre heterogeneity, because they are based on different DTDs or schemas, cover various topics and consist of diverse document types. This led to the establishment of the Heterogeneous Track at INEX 2006. The goal of the track was to set up a testbed consisting of several different and diverse collections and defining retrieval tasks and appropriate topics. These are the foundations for the Heterogeneous Track at INEX 2007, where the focus is on run submissions, relevance assessments and proper evaluation of the proposed methods dealing with a heterogeneous collection.

1 Introduction

The various tracks within the Initiative for the Evaluation of XML Retrieval (INEX) have always dealt with homogeneous collections based on a single DTD or XML schema. With the advent of distributed systems (federations or peer-to-peer systems), we also need to cope with the situation where every node manages its own distinct subcollection. So most realistic XML collections will consist of documents from different sources, and thus be heterogeneous w.r.t.

- *syntactic diversity*: documents are validated against different DTDs or schemas;
- *semantic diversity*: subcollections cover different topics;
- *genre diversity*: we may find various document types in subcollections, e.g. original articles, comments, metadata records, scientific papers, etc.

Dealing with a set of heterogeneous collections poses a number of challenges for structured document retrieval with XML. For content-oriented queries, where no structural conditions on the returned elements are requested, most current approaches use the DTD or XML Schema for defining elements that would form reasonable answers. In a heterogeneous environment, DTD- and Schema-independent methods need to be developed. For content and

structure queries, there is the added problem of mapping structural conditions from one DTD or Schema onto other (possibly unknown) DTDs and Schemas (for instance, if documents from a specific author are requested and the author of a document might be called "author" in one subcollection and "creator" in another one). Overcoming this syntactic diversity means finding equivalent structural elements. Methods from federated databases could be applied here, where schema mappings between the different DTDs are defined manually. However, for a larger number of DTDs, automatic methods must be developed, e.g. based on ontologies. Due to semantic and genre diversity, not every subcollection will be relevant for a given user's information need. Since querying each subcollection separately is expensive w.r.t. communication costs and result post-processing, it has been suggested in the distributed IR literature that preselection of appropriate collections should be performed.

The considerations above led to the establishment of a Heterogeneous Track[1] at INEX. On the long run, the Heterogeneous Track at INEX aims to answer, among others, the following research questions:

- For content-oriented queries, what methods are possible for determining which elements contain reasonable answers? Are pure statistical methods appropriate, or are ontology-based approaches also helpful?
- What methods can be used to map structural criteria onto other DTDs?
- Should mappings focus on element names only, or also deal with element content or semantics?
- How can suitable collections be preselected in order to improve retrieval efficiency and without corrupting retrieval effectiveness?
- What are appropriate evaluation criteria for heterogeneous collections?

In order to cope with above questions, we need test collections which are heterogeneous syntactically, semantically and by genre. Since such collections had not been set up previously, the main focus of effort for the track in 2006 was on the construction of an appropriate testbed, consisting of different individual collections, and on the identification of reasonable tasks and topics. The resulting testbed provides a basis for the future exploration of the research questions outlined above.

2 Collection and Topic Creation

We set up subcollections which constitute our heterogeneous collection by partly reusing collections offered in previous INEX runs and by preparing new collections. A specific DTD was defined for every subcollection, if not already available, ensuring syntactic heterogeneity. Table 1 shows some statistics about the subcollections.

The subcollections serve different domains, ranging from computer science (e.g. bibdb Duisburg, IEEE, DBLP) through technology news (ZDNet) to travel advice (Lonely Planet) and general purpose information (Wikipedia). We find several document genres ranging from simple metadata records to the full texts of scientific papers, articles and web sites as well as textual annotations forming discussion threads. Therefore, we have subcollections which differ with respect to their syntax (DTD), semantic (domains served) and document genre.

[1] See also `http://inex.is.inf.uni-due.de/2006/het.html`

Collection	Size	SubColl.	Documents	Elements	Mean Elements per Document
Berkeley	52M		12800	1182062	92.3
bibdb Duisburg	14M		3465	36652	10.6
CompuScience	993M		250986	6803978	27.1
DBLP	2.0G		501102	4509918	9.0
hcibib	107M		26390	282112	10.7
IEEE (2.2)	764M		16820	11394362	677.4
IDEAlliance	58M	eml	156	66591	426.9
		xml1	301	45559	151.4
		xml2	264	58367	221.1
		xmle	193	32901	170.5
		xtech	71	14183	199.8
Lonely Planet	16M		462	203270	440.0
qmulcsdbpub	8.8M		2024	23435	11.6
Wikipedia	4.9G		659385	1193488685	1810.0
ZDNet	339M	Articles	4704	242753	51.6
		Comments	91590	1433429	15.7
Totals	9.25G		1570713	1219818257	776.6

Table 1: Components of the heterogeneous collection. *Element counts estimated for large collections.*

We developed 61 topics during the topic creation phase, some of which were derived from previous and current topics used in the main INEX Adhoc retrieval tasks, and some created specifically for collections in the Heterogeneous track. Topics are formatted in XML using a topic DTD which allows specification of both content-only queries and also permits specification of structural criteria, such as requesting only abstracts, titles or comments. This makes the topics suitable for content-and structure (CAS) tasks used in the INEX Adhoc tasks. Some topics also contain a hint about the collection used to identify the topic (called "scope" here). Figure 1 shows an example of a topic definition.

3 Tasks and Run Submissions

The following tasks were proposed for the Heterogeneous Track at INEX 2006:

Adhoc Content-only Task Here, content-oriented queries are applied without any structural constraints. The systems return a ranked list of documents from all collections.

Content-and-Structure Task 1 The system should return only elements specified in structural constraints. For example, if only an abstract or a title of a document is explicitly requested.

Content-and-Structure Task 2 The system should basically return the elements specified in certain structural constraints, but also similar elements. As an example, `<doctitle>` in one collection and `<title>` in other collections are most probably equivalent.

Resource Selection The goal here is to select the most relevant resources (i.e., collections) for a given topic. The system should return a ranked list of collections for this task

For future run submissions we defined a DTD which covers rankings of elements as well as rankings of subcollections. This DTD allows those submitting runs to specify the collections

```
<!DOCTYPE inex_het_topic SYSTEM "het-topic.dtd">
<inex_het_topic  topic_id="21">
<title>desktop search</title>
    <castitle>
        //doctitle[about(.,desktop search)]
        //description[about(.,desktop search)]
        //comment[about(.,desktop search)]
    </castitle>
<description>
    Desktop search is concerned with searching the items on your local computer
    (in contrast to web search like Google, where you search information on the
    web). Several companies offer tools for desktop search.
</description>
<narrative>
    I want to inform myself about desktop search. What tools exist
    (standalone, browser-based) and what are their advantages and disadvantages?
</narrative>
<ontopic_keywords>desktop search</ontopic_keywords>
<scope>
    <collection>zdnetart</collection>
    <collection>zdnetcom</collection>
</scope>
</inex_het_topic>
```

Figure 1: Example topic definition

actually used in resolving the topics. Thus it permits users to submit runs for only a subset of the collections, and in principle such runs could be scored without counting the ignored collections.

4 The Heterogeneous Track at INEX 2007

In the 2006 Heterogeneous Track, we managed to set up a collection whose subcollections are heterogeneous w.r.t. syntax, semantics and document genre. We also defined a number of tasks and created test topics for evaluation. Thus, the foundations have been laid for the new Heterogeneous Track at INEX 2007. For this year we will concentrate on receiving run submissions, creating a pooled test set and providing relevance assessments which are in turn used for evaluation. But there are further questions to be examined and discussed for 2007. If we want to use structural constraints, should we be defining some notion of structural relevance to measure how well a system has met the structural constraints, regardless of matching topicality? Another question is whether pooling for evaluation should be by collection, or should we perform a cross-collection pooling? Many other issues and questions are sure to arise as the Heterogeneous Track for INEX 2007 gets under way.

We invite researchers interested in distributed IR to participate in the Heterogeneous Track at INEX 2007.