Behind the scenes of the digital museum of information retrieval research

Djoerd Hiemstra[1], Tristan Pothoven[1], Marijn van Vliet[1], and Donna Harman[2]

[2] University of Twente
The Netherlands
{hiemstra, pothovent, vlietwm}@cs.utwente.nl

[1] NIST
Gaithersburg, USA
donna.harman@nist.gov

1. INTRODUCTION

As more and more of the world becomes digital, and documents become easily available over the Internet, we are suddenly able to access all kinds of information. The downside of this however is that information that is not digital becomes less accessed, and is liable to be lost to us and to future generations. Whereas there are many scanning projects underway, such as Google books and the Open Library Alliance, these projects are not going to know about, much less find, the specialized scientific literature within various fields. This short paper describes the beginnings of a project to digitize some of the older literature in the information retrieval field [1]. The paper finishes with some thoughts for future work on making more of our IR literature available for searching.

2. INDEXING AND QUERYING

So far 14 of the older reports, such as Cyril Cleverdon's Cranfield reports and Gerard Salton's ISR reports have been scanned, along with a NIST Monograph containing a IR literature survey from the 1960's, a report on the MEDLARS evaluations, and Karen Sparck Jones's *Information Retrieval Experiment* book.

```
<museum>
(...)
 <item type="book" id="isr-10">
  <title>Document Retrieval System - Optimization and
      Evaluation</title>
  <author>Joseph John Rocchio</author>
  <publisher>Harvard University</publisher>
  <year>1966</year>
  <item id="isr-10-1" type="chapter">
   <title>Preface</title>
   <file source="isr-10/pdfs/frontmatter.pdf"/>
   <fulltext>
    <page nr="1">
     <p>THE COMPUTATION Harvard University LABORATORY
      Cambridge, Massachusetts Scientific Report No.
      ISR-10 INFORMATION STORAGE AND RETRIEVAL</p>
     <p>to The National Science Foundation Cambridge,
      Massachusetts March 1966 Gerard Salton Project
      Director</p>
    </page>
    <page nr="2">
     <p>&#xC2;&#xA9; Copyright, 1965 By
(...)
```

Figure 1: XML document resulting from OCR

The documents were scanned at 600 bpi greyscale and turned into PDF with hidden text via OCR. As the reports are almost all large documents with over 100 pages, scanning typically results in a separate pdf document per book chapter. When the pdfs of a report are added to the museum, its OCR recognized texts are derived (using `pdftotext`), and automatically marked up as shown in Figure 1. The elements *title*, *author*, *publisher*, and *year*, have to be inputed manually when uploading the document.

```
declare function museum:search(
  $query as xs:string, $page as xs:integer) as node()*
{
 let $tquery := tijah:tokenize($query)
 let $nexi := concat ("//item[about(.,", $tquery, ")]")
 let $opt := <TijahOptions returnNumber="{$page*10}"/>
 let $qid := tijah:queryall-id($nexi,$opt)
 let $result := tijah:nodes($qid)
 let $count := ceiling(tijah:resultsize($qid) div 10)
 return <result pages="{$count}"> {
  for $x in subsequence($result, $page*10 - 9, 10)
   let $nexi2 := concat ("//p[about(.,", $tquery, ")]")
   let $opt2 := <TijahOptions returnNumber="1"/>
   let $qr := tijah:query($x, $nexi2, $opt2)
   let $snippet := if ($qr) then $qr else ($x//p)[1]
   return <item id="{$x/@id}" type="{$x/@type}">
          { $x/title, $x/author, $x/file }
            <snippet> { $snippet/text() } </snippet>
          </item>
  }
 </result>
};
```

Figure 2: Example PF/Tijah XQuery

The XML data is inserted in PF/Tijah, an XQuery XML database system called Pathfinder (PF), which is integrated with an XML search system called Tijah [2]. PF/Tijah is developed at the University of Twente in cooperation with CWI Amsterdam and the University of Munich, and can be downloaded as part of the MonetDB/XQuery database system.[1] XML data that is inserted in PF/Tijah can be searched on any granularity. The system does not have the notion of "document": Any element can be retrieved using a keyword query, whether it is an *item*, a *page* or a *p* (paragraph) element. Note that both books and book chapters are tagged as *item* in the data, which allows us to do a simple query for items to retrieve both complete books and

[1] `http://dbappl.cs.utwente.nl/pftijah`

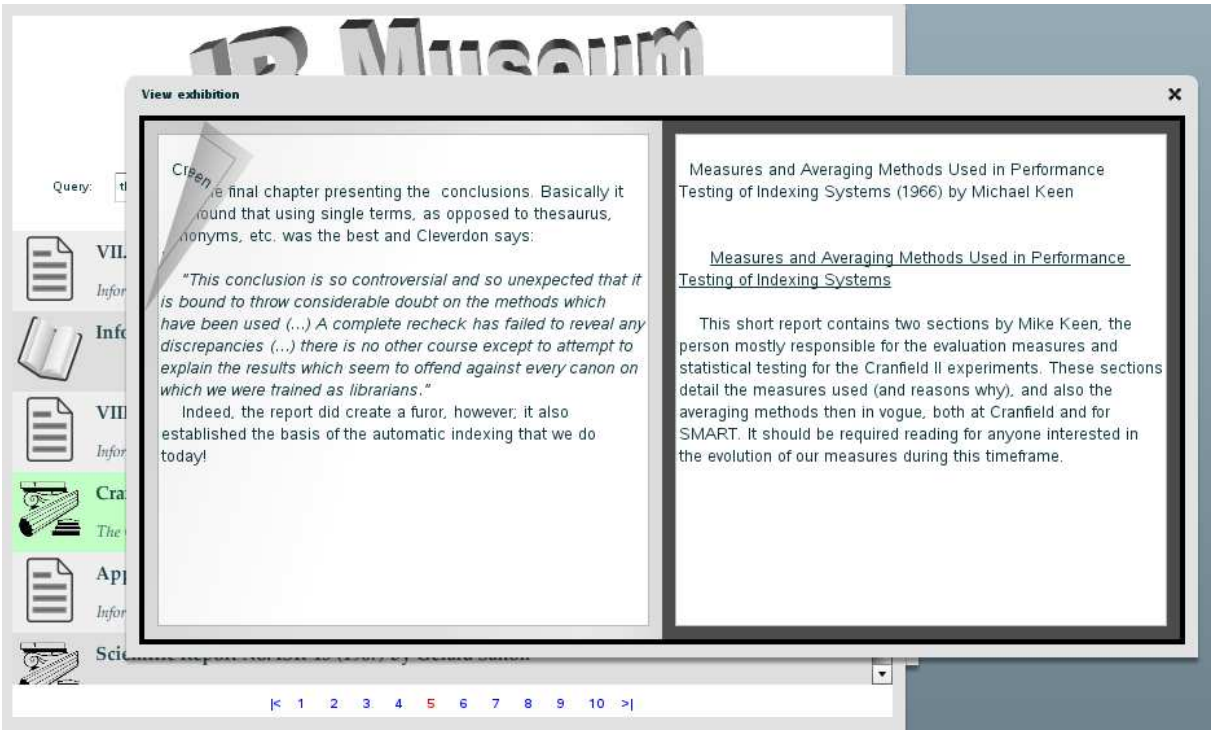

Figure 3: Screen shot showing an exhibition with search results in the back

individual chapters. The example query in Figure 2 shows how PF/Tijah supports powerful searching and result presentation by its query language. The function takes a text query and a result page and searches for *item* elements about the query. It then presents for the 10 best *item* elements its *id*, *type*, *title*, *author*, *file*, and for each element it searches for the best matching paragraph to be presented as a text snippet. All of this is done in one query.

3. EXHIBITIONS

At startup, the museum shows a number of *exhibitions*. The exhibitions contain for every book in the database a small background story. Exhibitions are presented to the user as traditional books with "real" pages that need to be turned by dragging them from right to left in order to go from one story to another. Exhibitions provide a guided tour along the reports and papers that fall under the exhibition's theme. Figure 3 shows a screen shot with a page from the Cranfield exhibition; and the search results partly visible in the back. Users can click directly to the original pdf documents from the exhibitions and from the search results.

All pdfs of the digital museum of information retrieval research are copyright free. They can be made available for download on request. Requests can be directed to the SIGIR by sending an email to `infodir_sigir@acm.org`. The code of the museum itself can be downloaded by choosing "view source" when clicking with the right mouse button.

4. CONCLUSION

What next? Well, we will be continuing the scanning project. The next step is to tackle the reports from the British Library; we are currently compiling a list and contacting them for permission. Several others have offered their books (copyright free) and at some point we would like to seriously go after the Salton books.

In the meantime, please take a look at what we have at: `http://www.sigir.org/museum/`, and learn about who introduced recall and precision, see how the Cranfield and MEDLARS tests were meticulously designed, be amazed by the fact that already in these early papers, full text search outperformed search using manually assigned keywords and thesaurus terms. Also, think of new ways of of accessing the documents. This project can provide some interesting challenges, particularly to the digital library community, including how to better access structured documents, how to deal with vocabulary shift over the years, and studies on how a user community would like to access this type of information.

5. ACKNOWLEDGMENTS

This work was funded in part by the ACM Special Interest Group on Information Retrieval (SIGIR) and by the Dutch Research Program MultimediaN.

6. REFERENCES

[1] Donna Harman and Djoerd Hiemstra. Saving and accessing the old information retrieval literature. *SIGIR Forum*, 42(2):16–21, 2008.

[2] Djoerd Hiemstra, Henning Rode, Roel van Os, and Jan Flokstra. PF/Tijah: Text search in an XML database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR)*, pages 12–17, 2006.