Overview of the INEX 2010 Book Track: Scaling Up the Evaluation Using Crowdsourcing

Gabriella Kazai[1], Marijn Koolen[2], Jaap Kamps[2],
Antoine Doucet[3], and Monica Landoni[4]

[1] Microsoft Research, United Kingdom
v-gabkaz@microsoft.com
[2] University of Amsterdam, Netherlands
{m.h.a.koolen,kamps}@uva.nl
[3] University of Caen, France
doucet@info.unicaen.fr
[4] University of Lugano
monica.landoni@unisi.ch

Abstract. The goal of the INEX Book Track is to evaluate approaches for supporting users in searching, navigating and reading the full texts of digitized books. The investigation is focused around four tasks: 1) Best Books to Reference, 2) Prove It, 3) Structure Extraction, and 4) Active Reading. In this paper, we report on the setup and the results of these tasks in 2010. The main outcome of the track lies in the changes to the methodology for constructing the test collection for the evaluation of the Best Books and Prove It search tasks. In an effort to scale up the evaluation, we explored the use of crowdsourcing both to create the test topics and then to gather the relevance labels for the topics over a corpus of 50k digitized books. The resulting test collection construction methodology combines editorial judgments contributed by INEX participants with crowdsourced relevance labels. We provide an analysis of the crowdsourced data and conclude that – with appropriate task design – crowdsourcing does provide a suitable framework for the evaluation of book search approaches.

1 Introduction

Prompted by the availability of large collections of digitized books, e.g., the Million Book project[1] and the Google Books Library project,[2] the Book Track was launched in 2007 with the aim to promote research into techniques for supporting users in searching, navigating and reading the full texts of digitized books. Toward this goal, the track provides opportunities to explore research questions around three areas:

- Information retrieval (IR) methods for searching collections of digitized books,
- Mechanisms to increase accessibility to the contents of digitized books, and
- Users' interactions with eBooks and collections of digitized books.

[1] http://www.ulib.org/

[2] http://books.google.com/

S. Geva et al. (Eds.): INEX 2010, LNCS 6932, pp. 98–117, 2011.

Based around the three main themes above, the following four tasks were investigated in 2010:

1. The *Best Books* to Reference (BB) task, framed within the user task of building a reading list for a given topic of interest, aims at comparing traditional document retrieval methods with domain-specific techniques, exploiting book-specific features, e.g., back-of-book index, or associated metadata, e.g., library catalogue information;
2. The *Prove It* (PI) task aims to test focused retrieval approaches on collections of books, where users expect to be pointed directly at relevant book parts that may help to confirm or refute a factual claim;
3. The *Structure Extraction* (SE) task aims at evaluating automatic techniques for deriving structure from OCR and building hyperlinked table of contents;
4. The *Active Reading task* (ART) aims to explore suitable user interfaces to read, annotate, review, and summarize multiple books.

In this paper, we report on the setup and the results of each of these tasks at INEX 2010. However, the main focus of the paper is on the challenge of constructing a test collection for the evaluation of the BB and PI tasks. This challenge has so far remained the main bottleneck of the Book Track, which in the past three years has struggled to build a suitably large scale test collection relying on its participants' efforts alone. Indeed, the effort required to contribute to the building of such a test collection is daunting. For example, the estimated effort that would have been required of a participant of the INEX 2008 Book Track to judge a single topic was to spend 95 minutes a day for 33.3 days [11]. This level of demand is clearly unattainable. At the same time, as a requirement of participation, it poses a huge burden and is likely to be one of the causes of the low levels of active participation that follows the high number of registrations.

To address this issue, this year we explored the use of crowdsourcing methods to contribute both the topics and the relevance labels to the test collection. This follows the recent emergence of human computing or crowdsourcing [8] as a feasible alternative to editorial judgments [2,1,7,13]. Similarly to our case, such efforts are motivated by the need to scale up the Cranfield method for constructing test collections where the most significant effort and cost is associated with the collection of relevance judgments. By harnessing the collective work of the crowds, crowdsourcing offers an increasingly popular alternative for gathering large amounts of relevance data feasibly at a relatively low cost and in a relatively short time.

Our goal this year was to establish if crowdsourcing could indeed be relied upon for creating a suitable test collection for the Book Track. To this end, we combined editorial judgments contributed by 'trusted' INEX participants with crowdsourced data, using the editorial labels as a gold set to measure the quality of the crowdsourced labels. In addition, we also explored the possibility to crowdsource not only the relevance labels, but the test topics too. Our analysis shows that with the appropriate task design, crowdsourcing does indeed offer a solution to the scalability challenge of test collection building [10].

Table 1. Active participants of the INEX 2010 Book Track, contributing topics, runs, and/or relevance assessments (BB = Best Books, PI = Prove It, SE = Structure Extraction, ART = Active Reading Task)

ID	Institute	Created topics	Runs	Judged topics
6	University of Amsterdam	19-20, 22	2 BB, 4 PI	05, 10, 18-19, 42, 64, 82
7	Oslo University College	02-06	5 PI	02-03
14	Uni. of California, Berkeley	-	4 BB	-
41	University of Caen	-	SE	SE
54	Microsoft Research Cambridge	00-01, 07-09, 24-25	-	00-01, 05-09, 12, 15, 18, 23-25, 31, 33, 42-43, 63-63, 70, 78, 81-82
86	University of Lugano	15-18, 21, 23	-	
98	University of Avignon	-	9 BB, 1 PI	00, 24, 77
339	University of Firenze	-	-	SE
386	University of Tokyo	-	SE	-
663	IIIT-H	10-14	-	-
732	Wuhan University	-	SE	-

In the following, we first we give a brief summary of the actively participating organizations (Section 2). In Section 3, we describe the book corpus that forms the basis of the test collection. The following three sections discuss our test collection building efforts using crowdsourcing: Section 4 details the two search tasks: BB and PI; Section 5 details our topic creation efforts; and Section 6 details the gathering of relevance labels. Then, in Section 7 we present the results of the BB and PI tasks, while Sections 8 and 9 summarize the SE and ART tasks. We close in Section 10 with a summary and plans for INEX 2011.

2 Participating Organizations

A total of 93 organizations registered for the track (compared with 84 in 2009, 54 in 2008, and 27 in 2007). However, of those registered only 11 groups took an active role (compared with 16 in 2009, 15 in 2008, and 9 in 2007), see Table 1.

2.1 Summary of Participants' Approaches

The University of Avignon (ID=98, [3]) contributed runs to the BB and PI tasks. They experimented with a method for correcting hyphenations in the books and applying different query expansion techniques. For retrieval they used the language modeling approach of the Lemur toolkit. In total, they corrected over 37 million lines (about 6%) in the corpus that contained hyphenated words, leading to around 1% improvement in MAP. No improvements were observed as a result of query expansion.

Oslo University College (ID=7, [14]) took part in the PI task and explored semantics-aware retrieval techniques where the weights of verbs that reflect confirmation were increased in the index. Using language modeling as their retrieval approach, they show that the new index can improve precision at the top ranks.

The University of California, Berkeley (ID=14, [12]) experimented with page level retrieval in the BB task. They derived book level scores by summing the page level scores within the books. Page level scores were generated in two ways: using a probabilistic approach based on logistic regression, and using coordination-level match (CML). They found that simple methods, e.g., CML do not work for book retrieval. Their page level logistic regression based method yielded the best results overall.

The University of Amsterdam (ID=6, [9]) looked at the effects of pseudo relevance feedback (RF) in both the BB and PI tasks, and also investigated the impact of varying the units of retrieval, e.g., books, individual pages, and multiple pages as units in the PI task. In the BB task, they found that their book level retrieval method benefited from RF. In the PI task, they achieved best performance with individual page level index and using RF. With larger units, RF was found to hurt performance.

The University of Caen (ID=41, [6]) participated in the SE task, continuing their approach of last year that uses a top-down document representation with two levels, part and chapter, to build a model describing relationships for elements in the document structure. They found that their approach is simple, fast, and generic—using no lexicon or special language dependent heuristics—but is also outperformed by methods tuned to the corpus and task at hand.

3 The Book Corpus

The Book Track builds on a collection of 50,239 out-of-copyright books[3], digitized by Microsoft. The corpus contains books of different genre, including history books, biographies, literary studies, religious texts and teachings, reference works, encyclopedias, essays, proceedings, novels, and poetry. 50,099 of the books also come with an associated MAchine-Readable Cataloging (MARC) record, which contains publication (author, title, etc.) and classification information. Each book in the corpus is identified by a 16 character bookID, which is also the name of the directory that contains the book's OCR file, e.g., A1CD363253B0F403.

The OCR text of the books has been converted from the original DjVu format to an XML format referred to as BookML, developed by Microsoft Development Center Serbia. BookML provides additional structure information, including a set of labels (as attributes) and additional marker elements for more complex structures, like table of contents. For example, the first label attribute in the XML extract below signals the start of a new chapter on page 1 (label="PT_CHAPTER"). Other semantic units include headers (SEC_HEADER), footers (SEC_FOOTER), back-of-book index (SEC_INDEX), table of contents

[3] Also available from the Internet Archive (although in a different XML format).

(SEC_TOC). Marker elements provide detailed markup, e.g., indicating entry titles (TOC_TITLE) or page numbers (TOC_CH_PN) in a table of contents.

The basic XML structure of a typical book in BookML is a sequence of pages containing nested structures of regions, sections, lines, and words, most of them with associated coordinate information, defining the position of a bounding rectangle ([coords]):

```
<document>
 <page pageNumber="1" label="PT_CHAPTER" [coords] key="0" id="0">
  <region regionType="Text" [coords] key="0" id="0">
   <section label="SEC_BODY" key="408" id="0">
    <line [coords] key="0" id="0">
     <word [coords] key="0" id="0" val="Moby"/>
     <word [coords] key="1" id="1" val="Dick"/>
    </line>
    <line [...]><word [...] val="Melville"/>[...]</line>[...]
   </section>   [...]
  </region>     [...]
 </page>        [...]
</document>
```

The full corpus, totaling around 400GB, is available on USB HDDs. A reduced version (50GB, or 13GB compressed) is available via download. The reduced version was generated by removing the word tags and propagating the values of the `val` attributes as text content into the parent (i.e., line) elements.

4 Search Tasks

Focusing on IR challenges, two search tasks were investigated in 2010: 1) Best Books to Reference (BB), and 2) Prove It (PI). Both these tasks used the corpus described in Section 3, and shared the same set of topics (see Section 5).

4.1 Best Books to Reference (BB) Task

This task was set up with the goal to compare book-specific IR techniques with standard IR methods for the retrieval of books, where (whole) books are returned to the user. The scenario underlying this task is that of a user searching for books on a given topic with the intent to build a reading or reference list, similar to those often found in academic publications or Wikipedia articles. The reading list may be for educational purposes or for entertainment, etc.

The task was defined as: *"The task is to return, for each test topic, a ranked list of 100 (one hundred) books estimated relevant to the general subject area of the factual statement expressed within the test topics, ranked in order of estimated relevance."*

Participants were invited to submit either single or pairs of runs. Each run had to include the search results for all the 83 topics of the 2010 test collection (see Section 5). A single run could be the result of either a generic (non-book-specific) or a book-specific IR approach. A pair of runs had to contain a non-book-specific run as a baseline and a book-specific run that extended upon the baseline by exploiting book-specific features (e.g., back-of-book index, citation statistics, book reviews, etc.) or specifically tuned methods. A run could contain, for each topic, a maximum of only100 books, ranked in order of estimated relevance.

A total of 15 runs were submitted by 3 groups (2 runs by University of Amsterdam (ID=6); 4 runs by University of California, Berkeley (ID=14); and 9 runs by the University of Avignon (ID=98)), see Table 1.

4.2 Prove It (PI) Task

The goal of this task is to investigate the application of focused retrieval approaches to a collection of digitized books. The scenario underlying this task is that of a user searching for specific information in a library of books that can provide evidence to confirm or refute a given factual statement (topic). Users are assumed to view the ranked list of book parts, moving from the top of the list down, examining each result.

In the guidelines distributed to participants, the task was defined as: *"The task is to return a ranked list of 1000 (one thousand) book pages (given by their XPaths), containing relevant information that can be used to either confirm or reject the factual statement expressed in the topic, ranked in order of estimated relevance."*

Participants could submit up to 12 runs, each containing a maximum of 1,000 book pages per topic for each of the 83 topics (see Section 5), ranked in order of estimated relevance.

A total of 10 runs were submitted by 3 groups (4 runs by the University of Amsterdam (ID=6); 5 runs by Oslo University College (ID=7); and 1 run by the University of Avignon (ID=98)), see Table 1.

5 Test Topics for the Search Tasks

In an effort to focus the search intentions to more specific (narrow) topics, this year we defined the test topics around one-sentence factual statements. Unlike previous years, we also solicited test topics both from INEX participants and from workers on Amazon's Mechanical Turk (AMT) service,[4] a popular crowdsourcing platform and labour market. Our aim was to compare the two sources and to assess the feasibility of crowdsourcing the topics (and the relevance judgments later on) of the test collection.

5.1 INEX Topics

We asked the INEX Book Track participants to create 5 topics each, 2 of which had to contain factual statements that appear both in the book corpus and in Wikipedia. Participants were asked to fill in a web form for each of their topics, specifying the factual statement they found in a book, the query they would

[4] https://www.mturk.com/mturk/

use to search for this information, the URL of the book containing the fact, the exact page number, the URL of the related Wikipedia article, the version of the fact as it appears in the Wikipedia page, and a narrative detailing the task and what information is regarded by the topic author as relevant. A total of 25 topics were submitted by 5 groups. Of these, 16 facts appear both in books and in Wikipedia.

5.2 Crowdsourced Topics

To crowdsource topics, we created two different human intelligence tasks (HITs) on AMT which asked workers to find general knowledge facts either in the books of the INEX corpus or both in the books and in Wikipedia:

Fact Finding Both in Books and in Wikipedia (Wiki HIT): To gather factual statements that appear both in the book corpus and in Wikipedia, we created a HIT with the following instructions: "Your task is to find a general knowledge fact that appears BOTH in a Wikipedia article AND in a book that is available at http://www.booksearch.org.uk. You can start either by finding a fact on Wikipedia first then locating the same fact in a book, or you can start by finding a fact in a book and then in Wikipedia. Once you found a fact both in Wikipedia and in the book collection, fill in the form below. HITs with correct matching facts will be paid $0.25. Only facts that appear both in Wikipedia and in the booksearch.org's book collection will be paid." We provided an example fact and instructed workers to record the following data for the factual statement they found: the Wikipedia article's URL, the fact as it appeared in the Wikipedia article, the URL of the book that states the same fact and the exact page number.

We set payment at $0.25 per HIT and published 10 assignments. All 10 assignments were completed within 4 hours and 18 minutes.On average, workers spent 11 minutes on the task, resulting in an effective hourly rate of $1.31.

Fact Finding in Books (Book HIT): To gather factual statements that appear in the book corpus, we created a simple HIT with the following instructions: "Your task is to find a general knowledge fact that you believe is true in a book available at http://www.booksearch.org.uk. Both the fact and the book must be in English. The fact should not be longer than a sentence. Only facts that appear in the book collection at http://www.booksearch.org.uk will be paid." As with the Wiki HIT, we provided an example fact and instructed workers to fill in a form, recording the factual statement they found, the URL of the book containing the fact and the exact page number.

Given the response we got for the Wiki HIT and the simpler task of the Book HIT, we first set payment at $0.10 per HIT and published 50 assignments. However, only 32 of the 50 assignments were completed in 13 days. We then cancelled the batch and published a second set of 50 assignments at $0.20 per HIT, this time pre-selecting to workers by requiring at least 95% HIT approval rate. This time, all 50 assignments were completed in 14 days. The average time workers spent on the task was 8 minutes in the first batch and 7 minutes in the second batch (hourly rate of $0.73 and $1.63, respectively).

5.3 Topic Selection

All collected topics were carefully reviewed and those judged suitable were selected into the final test collection. All topics contributed by INEX participants were acceptable, while filtering was necessary for topics created by AMT workers. Out of the 10 Wiki HITs, only 4 topics were selected (40%). Of the 32 Book HITs in the first batch, 18 were acceptable (56%), while 36 were selected from the 50 Book HITs in the second batch (72%). Topics from AMT workers were rejected for a number of reasons:

- 19 topics were rejected as the information given was a (random) extract from a book, rather than a fact, e.g., "logical history of principle of natural selection", "This is a picture diagram of fall pipes with imperfect joints being carried through the basement of house into drain", "At the age of twenty five he married a widow forty years old; and for five-and-twenty years he was a faithful husband to her alone", "A comparison of all the known specimens shows the material to be of slate and exhibits general uniformity in shape, the most noticeable differences being in the handle and the connecting neck.";
- 5 topics were nonsensical, e.g., "As a result of this experience I became tremendously interested in the theater can be found on page 63 of the book titled Printing and book designing by Wilson, Adrian.", "dance", "I want a woman with a soul","the objective facts, in nature or in the life of man";
- 2 topics had missing data, i.e., book URL, fact or page number;
- 2 topics referred to facts outside the book corpus, e.g., CBS news;
- 5 topics had incorrect book URLs or page references;
- 1 topic was the example fact included in the HIT.

Comparing the average time workers took to complete their task in the acceptable and not-acceptable sets of topics, we only found a small difference of 522 seconds vs. 427 seconds, respectively (st.dev. 676 and 503 seconds, min. 13 and 22 seconds, and max. 3389 and 2437 seconds), thus proving of little use in our case to automate filtering in the future.

In addition, out of the total 58 selected AMT topics, 18 had to be modified, either to rephrase slightly or to correct a date or name, or to add additional information. The remaining 40 HITs were however high quality (even more diverse and creative than the topics created by the INEX participants) and seemingly reflecting real interest or information need.

From the above, it is clear that crowdsourcing provides an attractive and promising means to scale up test topic creation: AMT workers contributed 58 topics, while INEX participants created only 25 topics. However, the quality of crowdsourced topics varies greatly and thus requires extra effort to weed out unsuitable submissions. This may be improved upon by pre-selection workers through qualification tasks [2,1] or by adopting more defensive task design [15]. Indeed, we found that selecting workers based on their approval rate had a positive effect on quality: batch 2 of the Book HITs which required workers to have a HIT approval rate of 95% had the highest rate of acceptable topics (72%). In addition, paying workers more (per hour) also shows correlation with the resulting quality.

6 Relevance Assessments

Like the topics, the relevance assessments were also collected from two sources: INEX participants and workers on AMT.

6.1 Gathering Relevance Labels from INEX Participants

INEX participants used the assessment system module of the Book Search System,[5] developed at Microsoft Research Cambridge. This is an online tool that allows participants to search, browse, read, and annotate the books of the test corpus. Annotation includes the assignment of book and page level relevance labels and recording book and page level notes or comments. Screenshots of the relevance assessment module are shown in Figures 1 and 2.

The assessment pools were created by taking the top 100 books from the BB and PI runs and ordering them by minimum rank and by popularity, i.e., the number of runs in which a book was retrieved. The book ranking of PI runs was based on the top ranked page of each book.

Relevance labels were contributed to a total of 30 topics. Of these, 21 topics were selected, those with the most judged pages at the start of January 2011, which were then consequently used for the AMT experiments (see next section). Relevance data gathering from INEX participant was frozen on the 22nd of February 2011.

6.2 Crowdsourcing Relevance Labels

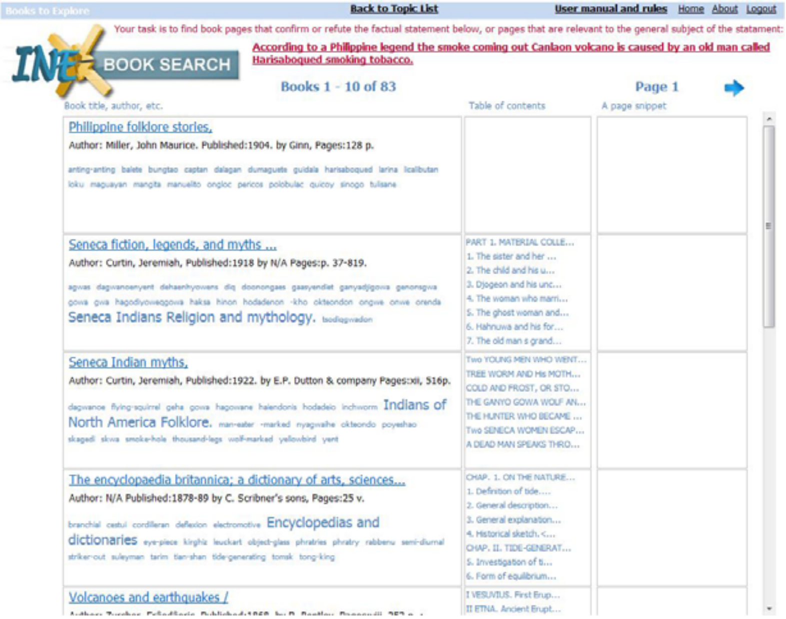
We collected further relevance labels from workers on AMT for the selected 21 topics. We created 21 separate HITs, one for each topic, so that the title of the HITs could reflect the subject area of the topic in the hope of attracting workers with interest in the subject.

Each HIT consisted of 10 pages to judge, where at least one page was already labeled as *confirm* or *refute* by an INEX participant. This was done to ensure that a worker encountered at least one relevant page and that we had at least one label per HIT to check the quality of the worker's work. In each batch of HITs, we published 10 HITs per topic and thus collected labels for 100 pages per topic from 3 workers, obtaining a total of 6,300 labels (3 labels per page).

Pooling Strategy. When constructing the AMT assessment pools (100 pages per topic), we combined three different pooling strategies with the aim to get i) a good coverage of the top results of the official PI runs, ii) a large overlap with the pages judged by INEX assessors (so that labels can be compared), and iii) to maximise the number of possibly relevant pages in the pool:

- Top-n pool: we pool the top n pages of the official PI runs using a round-robin strategy.

[5] `http://www.booksearch.org.uk/`



Your task is to find book pages that confirm or refute the factual statement below, or pages that are relevant to the general subject of the statament:

According to a Philippine legend the smoke coming out Canlaon volcano is caused by an old man called Harisaboqued smoking tobacco.

Philippine folklore stories,
Author: Miller, John Maurice. Published:1904. by Ginn, Pages:128 p.
anting-anting balete bungtao captan dalagan dumaguete guidala harisaboqued larina Icalibutan loku maguayan mangita manuelito ongloc pericos polobulac quicoy sinogo tulisane

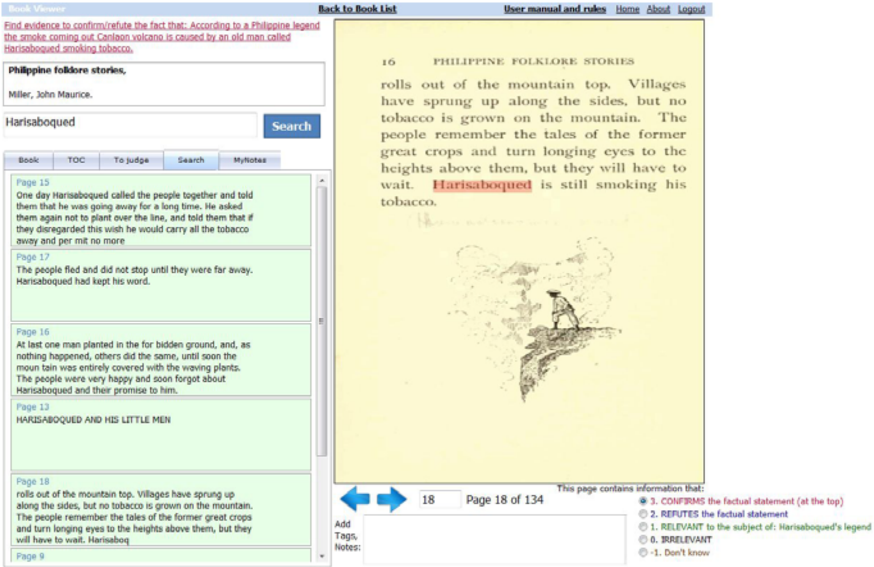
Seneca fiction, legends, and myths ...
Author: Curtin, Jeremiah, Published:1918 by N/A Pages:p. 37-819.
agwas dagwanoenyent dehaenhyowens diq doonongaes gaasyendiet ganyadjigowa genonsgwa gowa gwa hagodiyoweqgowa haksa hinon hodadenon -kho okteondon ongwe onwe orenda Seneca Indians Religion and mythology. tsodiqgwadon
PART 1. MATERIAL COLLE...
1. The sister and her ...
2. The child and his u...
3. Djogeon and his unc...
4. The woman who marri...
5. The ghost woman and...
6. Hahnuwa and his for...
7. The old man s grand...

Seneca Indian myths,
Author: Curtin, Jeremiah, Published:1922. by E.P. Dutton & company Pages:xii, 516p.
dagwanoe flying-squirrel geha gowa hagowane halendonis hodadeio inchworm Indians of North America Folklore. man-eater -marked nyagwaihe okteondo poyeshao skagedi skwa smoke-hole thousand-legs wolf-marked yellowbird yent
Two YOUNG MEN WHO WENT...
TREE WORM AND His MOTH...
COLD AND FROST, OR STO...
THE GANYO GOWA WOLF AN...
THE HUNTER WHO BECAME ...
Two SENECA WOMEN ESCAP...
A DEAD MAN SPEAKS THRO...

The encyclopaedia britannica; a dictionary of arts, sciences...
Author: N/A Published:1878-89 by C. Scribner's sons, Pages:25 v.
branchial cestui cordilleran deflexion electromotive Encyclopedias and dictionaries eye-piece kirghiz leuckart object-glass phratries phratry rabbenu semi-diurnal striker-out suleyman tarim tian-shan tide-generating tomsk tong-king
CHAP. 1. ON THE NATURE...
1. Definition of tide....
2. General description...
3. General explanation...
4. Historical sketch. <...
CHAP. II. TIDE-GENERAT...
5. Investigation of ti...
6. Form of equilibrium...

Volcanoes and earthquakes /
I VESUVIUS. First Erup...
II ETNA. Ancient Erupt...

Fig. 1. Relevance assessment module of the Book Search System, showing the list of books in the assessment pool for a selected topic



Find evidence to confirm/refute the fact that: According to a Philippine legend the smoke coming out Canlaon volcano is caused by an old man called Harisaboqued smoking tobacco.

Philippine folklore stories,

Miller, John Maurice.

Harisaboqued

Page 15
One day Harisaboqued called the people together and told them that he was going away for a long time. He asked them again not to plant over the line, and told them that if they disregarded this wish he would carry all the tobacco away and per mit no more

Page 17
The people fled and did not stop until they were far away. Harisaboqued had kept his word.

Page 16
At last one man planted in the for bidden ground, and, as nothing happened, others did the same, until soon the moun tain was entirely covered with the waving plants. The people were very happy and soon forgot about Harisaboqued and their promise to him.

Page 13
HARISABOQUED AND HIS LITTLE MEN

Page 18
rolls out of the mountain top. Villages have sprung up along the sides, but no tobacco is grown on the mountain. The people remember the tales of the former great crops and turn longing eyes to the heights above them, but they will have to wait. Harisaboq

Page 9

16 PHILIPPINE FOLKLORE STORIES

rolls out of the mountain top. Villages have sprung up along the sides, but no tobacco is grown on the mountain. The people remember the tales of the former great crops and turn longing eyes to the heights above them, but they will have to wait. Harisaboqued is still smoking his tobacco.

18 Page 18 of 134

Add Tags, Notes:

This page contains information that:

- ● 3. CONFIRMS the factual statement (at the top)
- ○ 2. REFUTES the factual statement
- ○ 1. RELEVANT to the subject of: Harisaboqued's legend
- ○ 0. IRRELEVANT
- ○ -1. Don't know

Fig. 2. Relevance assessment module of the Book Search System, showing the Book Viewer window. Relevance options are listed below the book page image.

Table 2. Statistics on the INEX and AMT relevance labels for the selected 21 topics

Source	INEX	AMT	INEX+AMT
Unknown	2	805	807
Irrelevant (0)	4,792	3,913	8,705
Relevant (1)	814	148	962
Refute (2)	18	113	131
Confirm (3)	349	1,321	1,670
Total	5,975	6,300	12,275

- Rank-boosted pool: in this pool the pages from the PI runs are reranked using a favorable book ranking. This book ranking is based on both the official BB and PI runs, and was used to create the pools for the INEX assessors to judge. The resulting page ranking has potentially more relevant pages in the top ranks and has a large coverage of the pages judged by the INEX assessors.
- Answer-boosted pool: we use a heuristic similarity function to increase the number of potentially relevant pages in the pool. We take all keywords (removing stopwords) from the factual statement of the topic that does not appear in the query and subject part of the topic, and rank the pages submitted to the PI task using coordination level matching.

As a result of the mixed pooling methods, in each 100 page assessment pool we have roughly the top 30 pages per pooling method plus the known relevant pages. Pages can occur only once in each HIT, but the known relevant pages could occur in multiple HITs, leading to 1,918 query/page pairs.

6.3 Collected Relevance Data

Statistics of the collected relevance labels are presented in Table 2. The Unknown category is used for when assessors could not judge a page (because the page was not properly displayed, or the text was written a language the assessor could not read). This happened more often in the crowdsourcing phase than in the INEX assessment phase.

For the 21 topics, a total of 5,975 page-level relevance labels were collected from INEX participant and 6,300 labels from workers on AMT. However, the AMT set contains 3 judgments per page, while the INEX data contains only one label per page (mostly). Due to missing participant IDs in the user accounts of two INEX assessors, 430 pages ended up being judged by multiple assessors. As a rule, only one label was required from the set of INEX participants, so when a page was judged by an INEX participant, it was removed from the pool. On the other hand, three labels were required by non-INEX users of the Book Search System. Interestingly, out of the 430 pages with multiple judgments, there are only 39 pages with disagreements (agreement is 91%).

A noticeable difference between the INEX and AMT labels is the relative high volume of *relevant* labels at INEX and *confirm* labels in the AMT set. The latter

Table 3. Agreement and consensus among the AMT workers and agreement between AMT majority vote and INEX labels, over different classes of labels

	All	Binary	proof
AMT agreement	0.71	0.78	0.89
AMT consensus	0.90	0.92	0.91
AMT-INEX agreement	0.72	0.77	0.78
AMT-INEX consensus	0.87	0.89	0.91

is at least partly due to the fact that each HIT had at least one page labeled as *confirm* or *refute* (but mostly *confirm*). In the next section, we look at the agreement between INEX and AMT labels.

6.4 Analysis of Crowdsourced Relevance Labels

In this section, we look at agreement and consensus among the AMT workers. For agreement we look at average pairwise agreement per topic and page (so over pairs of judgments). We have, in principle, judgments from three AMT workers, resulting in three pairs of different workers, whose average pairwise agreement may range from 0 (all three pick a different label) to 1 (all three pick the same label). Consensus is the percentage of labels that form the majority vote. That is, the page label that gets the majority vote of m workers out of the total of n workers labelling that page leads to a consensus of $\frac{m}{n}$. A higher consensus means agreement is more concentrated among a single label. This is useful when there are more than 2 possible labels. Again we have, in principle, judgments from three AMT workers, whose consensus may range from 0.3333 (all three pick a different label) to 1 (all three pick the same label). We also look at agreement between AMT majority vote labels and INEX labels. If this agreement is high, AMT labels might reliably be used to complement or replace editorial judgments from INEX participants.

We look at agreement and consensus among the AMT labels using a number of label classes:

- All classes: no conflation of labels, giving four classes: *irrelevant*, *relevant*, *refute* and *confirm*
- Binary: the *relevant*, *refute* and *confirm* labels are conflated, leading to only two classes: *irrelevant* and *relevant/confirm/refute.*
- Proof: we ignore the irrelevant labels and conflate the refute and confirm labels, leading to two classes: *relevant* and *confirm/refute.*

In Table 3 we see the agreement and consensus among the AMT labels. If we differentiate between all 4 labels, agreement is 0.71. Consensus is 0.90, which means that, on average, the majority vote for a label forms 90% of all worker votes. If we consider only binary labels, the percentage agreement is higher. Also the agreement among the different degrees of relevance is high with 0.78. Due to the relatively strong percentage agreement, consensus is high among all sets.

Table 4. Statistics on the official Prove It relevance assessments based on the INEX and AMT labels

Sets	INEX	AMT	ip2c-set
Judgements	5,537	1,873	6,527
Irrelevant (0)	4,502	1,500	5,319
Relevant (1)	712	17	719
Confirm/Refute (2)	323	356	489

We also look at agreement between the relevance judgments derived from the majority vote of the AMT labels with gold set of INEX labels (bottom half of Table 3). Agreement over all 4 label classes is 0.72. AMT workers are more likely to label a page as *refute* or *confirm* than INEX participants. Without the *irrelevant* labels, the *relevant* labels dominate the INEX judgments and the *refute/confirm* labels dominate the AMT judgments, which leads to a somewhat lower agreement on these labels.

6.5 Official Qrels

Page Level Judgments. From the multiple labels per page, we derived a single judgment for evaluation. First, we discarded judgments in the *unknown* category and conflate the *refute* and *confirm* labels to a single relevance value (=2). We give confirm and refute pages the same relevance value because the PI task requires a system to find pages that either confirm or refute the factual statement of the topic. Thus, both types of pages satisfy this task. We then use majority rule among the AMT labels and keep the lower relevance value in case of ties. For the 39 pages with disagreeing INEX labels, we chose the label with the higher relevance value. We merge the two sets by always keeping the INEX labels over and above an AMT label for the same page. We refer to the resulting set as the `ip2c-set` qrel set (INEX page level judgments and crowdsourced label set) and use this set as the official set for the evaluation of the PI task. Statistics for this set are given in Table 4. In total, we have 489 pages that confirm or refute a factual statement (23 per topic on average) and 719 pages that are relevant to the topic of the factual statement (34 per topic).

Book Level Judgments. In addition to the page level judgments, it was necessary to gather book level judgments to evaluate the BB runs. These labels were provided by the task organizers for the pool of books constructed from the top 10 books of all BB runs. Books were judged on a four-point scale: 0) irrelevant, 1) marginally relevant (i.e., the book contains only a handful of pages related to the subject of the topic, 2) relevant (i.e., the topic is a minor theme), and 3) perfect (the book is dedicated to the topic).

Statistics on the relevance judgements are given in Table 5. A total of 990 books have been judged for 21 topics (47 per topic). Of these, 210 were marginally relevant, 117 relevant and 36 were perfect. The 36 perfect books are spread across 11 topics. That is, for 10 topics no perfect books were pooled. There is 1 topic (2010070) with no relevant or perfect books.

Table 5. Statistics on the official Best Books relevance assessments

Judgements	990
Irrelevant (0)	627
Marginally relevant (1)	210
Relevant (2)	117
Perfect (3)	36

7 Evaluation Measures and Results for the Search Tasks

7.1 Best Books Task Evaluation

For the evaluation of the Best Books task, we use the book level relevance labels given in the `ib-org-set` qrel set and report standard trec-eval measures: Mean Average Precision (MAP), Precision at 10 (P@10) and Normalized Cumulative Gain at 10 (NDCG@10). NDCG@10 uses the graded relevance scores, while for the binary measures the four-point relevance scale was collapsed to binary labels (Irrelevant (0), all other relevant degrees (1)).

Table 6 shows the effectiveness scores for the Best Book runs, where NDCG@10 is regarded as the official measure.

The best BB run (NDCG@10=0.6579) was submitted by the University of California, Berkeley (p14-BOOKS2010_T2_PAGE_SUM_300) who employed page level retrieval methods and derived book level scores by summing the page level scores within the books. Page level scores were generated using a probabilistic approach based on logistic regression. A run by the University of Avignon followed close second with NDCG@10=0.6500. They experimented with a method for correcting hyphenations in the books and used the language modeling approach of the Lemur toolkit.

Table 6. Evaluation results for the INEX 2010 Best Books task

Run ID	MAP	P@10	**NDCG@10**
p14-BOOKS2010_CLM_PAGE_SUM	0.1507	0.2714	0.2017
p14-BOOKS2010_CLM_PAGE_SUM_300	0.1640	0.2810	0.2156
p14-BOOKS2010_T2FB_BASE_BST	0.3981	0.5048	0.5456
p14-BOOKS2010_T2_PAGE_SUM_300	0.5050	0.6667	**0.6579**
p6-inex10.book.fb.10.50	0.3087	0.4286	0.3869
p6-inex10.book	0.3286	0.4429	0.4151
p98-baseline_1	0.4374	0.5810	0.5764
p98-baseline_1_wikifact	0.4565	0.5905	0.5960
p98-baseline_2	0.4806	0.6143	0.6302
p98-baseline_2_wikifact	0.5044	0.6381	0.6500
p98-fact_query_10wikibests	0.4328	0.5714	0.5638
p98-fact_query_entropy	0.4250	0.5476	0.5442
p98-fact_query_tfidfwiki	0.3442	0.4667	0.4677
p98-fact_query_tfwiki	0.4706	0.5571	0.5919
p98-fact_stanford_deps	0.4573	0.5857	0.5976

7.2 Prove It Task Evaluation

For the evaluation of the PI task, we use the qrel set of `ip2c-set`, which contains page level judgements contributed both by INEX participants and by the workers on AMT. As detailed in Section 6, the set was created by first applying majority rule to the AMT labels after spam labels have been removed, where in case of ties we kept the lower relevance degree, then merging this set with the INEX labels always taking the INEX label above an AMT label.

As with the BB task, we report standard trec-eval measures: MAP, P@10 and NDCG@10. For NDCG, we used two different weighting options:

- 0-1-2 weighting, which simply reflects the original relevance grades, where pages that confirm/refute the topic statement are twice as important as pages that simply contain related information.
- 01-10 weighting that emphasizes pages that confirm/refute the topic statement, treating them 10 times as important as other relevant pages:
  - Irrelevant (0) → 0
  - Relevant (1) → 1
  - Confirm/Refute (2) → 10

For the binary measures, all classes of relevance were mapped to 1, while irrelevant to 0. We regard the NDCG@10 as the official measure. Table 7 shows the effectiveness scores for the Prove It runs, where only exact page matches are counted as hits.

The best PI run (NDCG@10=0.2946) was submitted by the University of Amsterdam (p6-inex10.page.fb.10.50), who investigated the impact of varying the units of retrieval, e.g., books, individual pages, and multiple pages as units in the PI task. They achieved best performance with their individual page level index and using pseudo relevance feedback.

Accounting for Near-Misses in the PI Task. Figure 3 shows the effectiveness scores for the Prove It runs, calculated over the `ip2c-set`, where near-misses

Table 7. Evaluation results for the INEX 2010 Prove It task (exact match only)

Run ID	MAP	P@10	NDCG@10 (0-1-2 weighting)	NDCG@10 (0-1-10 weighting)
p6-inex10.5page.fb.10.50	0.1163	0.2143	0.1703	0.1371
p6-inex10.5page	0.1209	0.2619	0.2182	0.1714
p6-inex10.page.fb.10.50	0.1521	0.3524	**0.2946**	0.2322
p6-inex10.page	0.1216	0.3238	0.2795	**0.2338**
p7-to_g.res	0.0453	0.1714	0.1276	0.0876
p7-to_g_2xover1.res	0.0342	0.1476	0.1225	0.0882
p7-to_g_2xover3.res	0.0288	0.1286	0.1124	0.0827
p7-to_g_5xover1.res	0.0340	0.1476	0.1195	0.0841
p7-to_g_5xover3.res	0.0262	0.1333	0.1119	0.0826
p98-fact_query_10wikibests_focus	0.0097q	0.0429	0.0321	0.0222

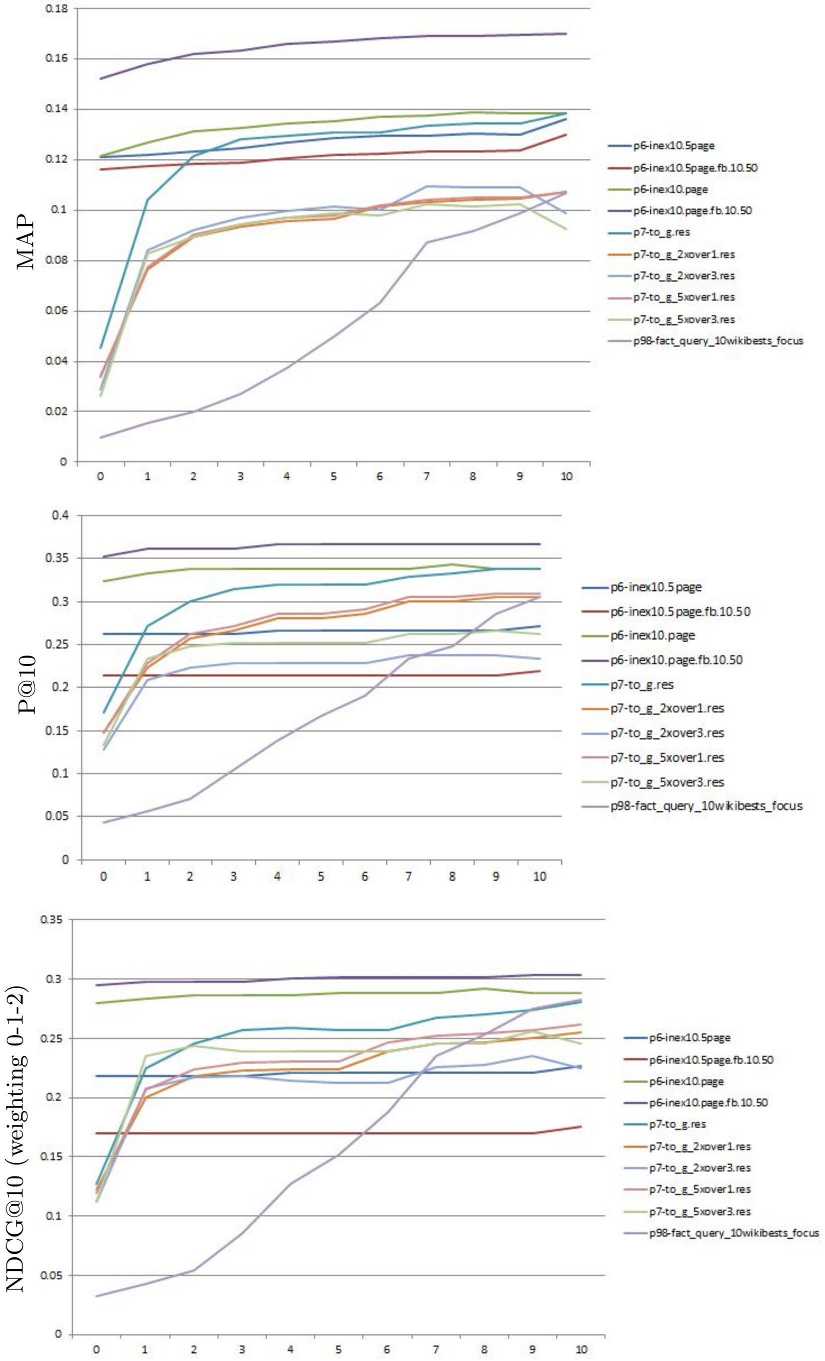


Fig. 3. Evaluation results for the INEX 2010 Prove It task with near-misses of n page distance

are taken into account. This was done by 'replacing' an irrelevant page in a run with a relevant page that is within *n* distance, starting with n=0 and increasing to 10 (where a relevant page could only be 'claimed' once). Some of the lower scoring submission pick up quickly, showing that they do retrieve pages in books with relevance, even retrieve pages that are in close proximity to the desired relevant page. The better scoring runs are fairly stable, demonstrating clearly that they are effective in locating the precise relevant pages inside the books.

8 The Structure Extraction (SE) Task

The goal of the SE task is to test and compare automatic techniques for extracting structure information from digitized books and building a hyperlinked table of contents (ToC). In 2010, the task was run only as a follow-up of the conjoint INEX and ICDAR 2009 competition [4,5], enabling participants to refine their approaches with the help of the ground-truth built in 2009.

Only one institution, the University of Caen, participated in this rerun of the 2009 task. Both the University of Caen and a new group, the University of Firenze, contributed to the building of the SE ground-truth data, adding 114 new books with annotated ToCs, increasing the total to 641 books.

The performance of the 2010 run is given in Table 8. A summary of the performance of the 2009 runs with the extended 2010 ground-truth data is given in Table 9.

Table 8. Score sheet of the run submitted by the University of Caen during the 2010 rerun of the SE competition 2009

	Precision	Recall	F-measure
Titles	18.03%	12.53%	12.33%
Levels	13.29%	9.60%	9.34%
Links	14.89%	7.84%	7.86%
Complete except depth	14.89%	10.17%	10.37%
Complete entries	10.89%	7.84%	**4.86%**

Table 9. Summary of performance scores for the 2009 runs with the extended 2010 ground-truth data; results are for complete entries

RunID	Participant	F-measure (2010)	F-measure (2009)
MDCS	MDCS	43.39%	41.51%
XRCE-run2	XRCE	28.15%	28.47%
XRCE-run1	XRCE	27.52%	27.72%
XRCE-run3	XRCE	26.89%	27.33%
Noopsis	Noopsis	8.31%	8.32%
GREYC-run1	University of Caen	0.09%	0.08%
GREYC-run2	University of Caen	0.09%	0.08%
GREYC-run3	University of Caen	0.09%	0.08%

9 The Active Reading Task (ART)

The main aim of ART is to explore how hardware or software tools for reading eBooks can provide support to users engaged with a variety of reading related activities, such as fact finding, memory tasks, or learning. The goal of the investigation is to derive user requirements and consequently design recommendations for more usable tools to support active reading practices for eBooks.

ART is based on the evaluation experience of EBONI [16], and adopts its evaluation framework with the aim to guide participants in organising and running user studies whose results could then be compared. The task is to run one or more user studies in order to test the usability of established products (e.g., Amazon's Kindle, iRex's Ilaid Reader and Sony's Readers models 550 and 700) or novel e-readers by following the provided EBONI-based procedure and focusing on INEX content. Participants may then gather and analyse results according to the EBONI approach and submit these for overall comparison and evaluation. The evaluation is task-oriented in nature.

Our aim is to run a comparable but individualized set of studies, all contributing to elicit user and usability issues related to eBooks and e-reading. However, the task has so far only attracted 2 groups, none of whom submitted any results at the time of writing.

10 Conclusions and Plans

The INEX Book Track promotes the evaluation of modern access methods that support users in searching, navigating and reading the full texts of digitized books, and investigated four tasks: 1) Best Books to Reference, 2) Prove It, 3) Structure Extraction, and 4) Active Reading. In this paper, we reported on the setup and the results of these tasks in 2010.

The main track activity was in the two search tasks, Best Books and Prove It. A total of 15 BB runs were submitted by 3 groups, and a total of 10 PI runs by 3 groups. Best Book submissions were shown to be highly effective, the best BB run obtaining an NDCG@10 score of 0.6579 (University of California, Berkeley, who combine book level and page level scores), and the runner up run a score of 0.6500 (University of Avignon, who used a dedicated tokenizer within the language modeling approach). The Prove It submissions were surprisingly effective, given that they try to solve the genuine needle-in-a-haystack problem of book page retrieval. This was probably aided by the topics being verbose and specific statements of facts to be confirmed or refuted. The best PI run obtained an NDCG@10 score of 0.2946 (University of Amsterdam, using an individual page level index with pseudo relevance feedback). The SE task was run (though not advertised), using the same data set as last year. One institution participated and contributed additional annotations. The final task, ART, attracted the interest of two participants, but no comprehensive experiment was conducted.

The main outcome of the track this year lies in the changes to the methodology for constructing the test collection for the evaluation of the two search tasks. In an effort to scale up the evaluation, we explored the use of crowdsourcing both to create the test topics and then to gather the relevance labels for the topics over a corpus of 50k digitized books. The resulting test collection construction methodology combines editorial judgments contributed by INEX participants with crowdsourced relevance labels. With our quality control rich crowdsourcing design, we obtained high quality labels showing 78% agreement with INEX gold set data [10]. This has paved the way to completely removing the burden of relevance assessments from the participants in 2011.

In 2011, the track will shift focus onto more social and semantic search scenarios, while also continuing with the ART and SE tasks. The track will build on its current book corpus as well as a new collection from Amazon Books and LibraryThing. The PI task will run with minor changes, also asking systems to differentiate positive and negative evidence for a given factual claim. The BB task will be replaced by the new Social Search for Best Books (SSBB) task which will build on the corpus of 1.5 million records from Amazon Books and LibraryThing. SSBB will investigate the value of user-generated metadata, such as reviews and tags, in addition to publisher-supplied and library catalogue metadata, to aid retrieval systems in finding the best, most relevant books for a set of topics of interest.

Acknowledgments. The crowdsourcing experiments of the track were generously supported by Microsoft Research Cambridge. Marijn Koolen and Jaap Kamps were supported by the Netherlands Organization for Scientific Research (NWO, grants 612.066.513, 639.072.601, and 640.001.501).

References

1. Alonso, O., Mizzaro, S.: Can we get rid of TREC assessors? using Mechanical Turk for relevance assessment. In: Geva, S., Kamps, J., Peters, C., Sakai, T., Trotman, A., Voorhees, E. (eds.) Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation, pp. 15–16 (2009)
2. Alonso, O., Rose, D.E., Stewart, B.: Crowdsourcing for relevance evaluation. SIGIR Forum 42, 9–15 (2008)
3. Deveaud, R., Boudin, F., Bellot, P.: LIA at INEX 2010 Book Track. In: Geva, S., et al. (eds.) INEX 2010. LNCS, vol. 6932, pp. 118–127. Springer, Heidelberg (2010)
4. Doucet, A., Kazai, G., Dresevic, B., Uzelac, A., Radakovic, B., Todic, N.: ICDAR 2009 Book Structure Extraction Competition. In: Proceedings of the Tenth International Conference on Document Analysis and Recognition (ICDAR 2009), Barcelona, Spain, pp. 1408–1412 (2009)
5. Doucet, A., Kazai, G., Dresevic, B., Uzelac, A., Radakovic, B., Todic, N.: Setting up a competition framework for the evaluation of structure extraction from OCR-ed books. International Journal on Document Analysis and Recognition, 1–8 (2010)
6. Giguet, E., Lucas, N.: The Book Structure Extraction Competition with the Resurgence software for part and chapter detection at Caen University. In: Geva, S., et al. (eds.) INEX 2010. LNCS, vol. 6932, pp. 128–139. Springer, Heidelberg (2010)

7. Grady, C., Lease, M.: Crowdsourcing document relevance assessment with mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT 2010, pp. 172–179, Association for Computational Linguistics (2010)
8. Howe, J.: Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business, 1st edn. Crown Publishing Group (2008)
9. Kamps, J., Koolen, M.: Focus and Element Length in Book and Wikipedia Retrieval. In: Geva, S., et al. (eds.) INEX 2010. LNCS, vol. 6932, pp. 140–153. Springer, Heidelberg (2010)
10. Kazai, G., Kamps, J., Koolen, M., Milic-Frayling, N.: Crowdsourcing for book search evaluation: Impact of quality on comparative system ranking. In: SIGIR 2011: Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York (2011)
11. Kazai, G., Milic-Frayling, N., Costello, J.: Towards methods for the collective gathering and quality control of relevance assessments. In: SIGIR 2009: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York (2009)
12. Larson, R.R.: Combining Page Scores for XML Book Retrieval. In: Geva, S., et al. (eds.) INEX 2010. LNCS, vol. 6932, pp. 154–163. Springer, Heidelberg (2010)
13. Le, J., Edmonds, A., Hester, V., Biewald, L.: Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In: SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation, pp. 21–26 (2010)
14. Preminger, M., Nordlie, R.: OUCs participation in the 2010 INEX Book Track. In: Geva, S., et al. (eds.) INEX 2010. LNCS, vol. 6932, pp. 164–170. Springer, Heidelberg (2010)
15. Quinn, A.J., Bederson, B.B.: Human computation: A survey and taxonomy of a growing field. In: Proceedings of CHI 2011 (2011)
16. Wilson, R., Landoni, M., Gibb, F.: The web experiments in electronic textbook design. Journal of Documentation 59(4), 454–477 (2003)