Overview of the INEX 2007 Book Search Track (BookSearch’07)

Gabriella Kazai\(^1\) and Antoine Doucet\(^2\)

\(^1\) Microsoft Research Cambridge, United Kingdom
gabkaz@microsoft.com
\(^2\) University of Caen, France
doucet@info.unicaen.fr

Abstract. This paper provides an overview of the newly launched Book Search Track at INEX 2007 (BookSearch’07), its participants, tasks, book corpus, test topics and relevance assessments, as well as some results.

1 Introduction

Libraries around the world and commercial companies like Amazon, Google and Microsoft are digitizing hundreds of thousands of books in an effort to enable online access to these collections. The Open Content Alliance (OCA\(^1\)), a library initiative formed after Google announced its library book digitization project, has brought library digitization efforts into the public eye, even though libraries have been digitizing books for decades before that. However, unlike most library digitization projects of the past, which centered around preservation and involved the careful and individual selection of materials to be digitized, the recent mass-digitization efforts aim at the conversion of materials on an industrial scale with minimum human intervention \(^2\).

The increasing availability of the full-text of digitized books on the Web and in digital libraries, both enables and prompts research into techniques that facilitate storage, access, presentation and use of the digitized content. Indeed, the unprecedented scale of the digitization efforts, the unique characteristics of the digitized material as well as the unexplored possibilities of user interactions make full-text book search an exciting area of research today.

Motivated by this need, the book search track was launched in 2007 as part of the INEX initiative. INEX was chosen as a suitable forum due to its roots in the evaluation of structured document retrieval (SDR) approaches and since searching for information in a collection of books can be seen as one of the natural application areas of SDR. For example, in focused retrieval a clear benefit to users is to gain direct access to parts of books (of potentially hundreds of pages) relevant to the information need.

The ultimate goal of the INEX book search track is to investigate book-specific relevance ranking strategies, UI issues and user behaviour, exploiting

\(^1\) http://www.opencontentalliance.org/
special features, such as back of book indexes provided by authors, and linking to associated metadata like catalogue information from libraries. However, searching over large collections of digitized books comes with many new challenges that need to be addressed first. For example, proper infrastructure has to be developed to allow for the scalable storage, indexing and retrieval of the digitized content. In addition, the setting up of a new track requires identifying suitable usage scenarios and tasks, establishing an evaluation framework complete with relevance criteria, judgement procedures and evaluation metrics, as well as the development of a support system infrastructure. In its first year, the track set to explore these issues with the aim to investigate the requirements for such an infrastructure.

This paper reports on the outcome of the BookSearch’07 track. It provides an overview of its participants, tasks, book corpus, test topics and relevance assessments, as well as some results and findings. Since, at the time of writing, the relevance assessments for one of the tasks (Page in Context) were still outstanding, the results for this task are not reported here.

This paper is organised as follows. Section 2 gives a brief summary of the participating organisations. In Section 3, we briefly describe the retrieval tasks defined at BookSearch’07. Section 4 details the book corpus, test topics, and relevance assessments. Section 5 presents the results of the evaluation. Finally, we close with a summary and plans for BookSearch’08.

2 Participating Organisations

In response to the call for participation, issued in April 2007, 27 organisations registered for BookSearch’07. Throughout the year, however, a number of groups dropped out and only about a third remained active by the end of the year. Most groups reported difficulties due insufficient resources, including lack of space to store the dataset or scalable approach to process it, as well as lack of time or human resources required to tackle the various tasks.

The 27 groups along with details of their participation are summarized in Table 1. As it can be seen, only 10 groups remained active throughout. 16 groups downloaded the book corpus, 7 groups contributed search topics, and only 2 groups managed to submit runs.

3 Retrieval Tasks

The track defined four tasks: 1) Book Retrieval, 2) Page in Context retrieval, 3) Classification and 4) User intent taxonomy building. A summary of these are given in the following sections. Further details and the various submission DTDs are available in the track’s Tasks and Submission Guidelines [7].

3.1 Book Retrieval Task

The goal of this task was to investigate the impact of book specific features on the effectiveness of book search systems, where the unit of retrieval is the