New Tasks on Collections of Digitized Books

Gabriella Kazai[1], Antoine Doucet[2], and Monica Landoni[3]

[1] Microsoft Research Cambridge
[2] University of Caen
[3] University of Lugano

Abstract. Motivated by the plethora of book digitization projects around the world, the Initiative for the Evaluation of XML Retrieval (INEX) launched a Book Search track in 2007. The track focused on Information Retrieval (IR) tasks, exploring the utility of traditional and structured document retrieval techniques to books. In this paper, we propose four new tasks to be investigated at the Book Search track. The tasks aim to promote research in a wider context across IR, Human Computer Interaction, Digital Libraries and eBooks. We identify novel problem areas, define tasks around these and propose possible evaluation methods.

1 Introduction

Through mass-digitization projects, such as the Million Book project[4] and the Google Book Search Library project, thousands of digitized books are becoming available on the Web and in digital libraries. The unprecedented scale of these efforts, the unique characteristics of the digitized material, the unexplored possibilities of user interactions, combined with the specialised domain that books represent, raise a range of previously unexplored research questions. In this paper, we highlight four questions and define tasks based around these for the INEX 2008 Book Search track[5].

2 Structure extraction

Unlike digitally-born content, the logical structure of digitized books is not readily available. A digitized book is often only split into pages with possible paragraph, line and word markup. This is also the case for the 50,000 digitized books used at INEX. The use of more meaningful structure, e.g., chapters, table of contents, bibliography, or back-of-book index, to support focused retrieval has been explored for many years at INEX and has been shown to increase retrieval performance [1]. The goal of our proposed structure extraction task is thus to explore the challenge of identifying the structure of digitized books. The task will first focus on tables of contents, but can later be expanded to the identification of more exhaustive structure information. To set up the task, a selection of 100 digitized books, representing a variety of genre and structure types, is to be selected from the corpus and distributed to participants. Participants will be required to build the tables of contents for each of those books. The resulting tables will then be compared to a manually built ground truth, and evaluated using standard precision and recall metrics. Because the ground truth is not

[4] http://www.ulib.org/
[5] http://www.inex.otago.ac.nz/tracks/books/books.asp

necessarily optimal, we also intend to evaluate the quality of the generated tables independently. This can be achieved by asking users to grade all tables of contents (including the original ones) on a quality scale.

3 Fact finding from trusted sources

The much debated 'Google generation' report by the CIBER research team at University College London claims that young people rely on Internet search tools to conduct their research without assessing the reliability of the information they find. In an effort to confirm or reject these findings, we propose a fact finding task that aims to pit information found on the Web against information extracted from books, which may be seen as more authoritative sources. The task would require applying question answering techniques to books to return facts to users. These will then be evaluated against answers obtained from Web search engines by human judges on aspects of trustworthiness.

4 Virtual bookshelf

By displaying related books in proximity of each other on book shelves, libraries supports serendipitous browsing and discovery. Motivated by this, we propose a task to build virtual book shelves using content-based IR techniques, such as classification. Participants would be required to create and submit a fixed length list of books related to a given user query. The evaluation of the corresponding virtual book shelves will be conducted through user tasks based on the queries and observing users' browsing behaviour and collecting judgement on the usefulness of the related books presented to them. Analysis of the way users browse a virtual book shelf and how successful they are in completing their task will be used to provide quantitative evaluation.

5 Supporting users' active reading

Active reading [2] is the combination of reading with critical thinking and learning, and involves not just reading per se, but also underlining, highlighting and commenting. Techniques to support active reading have been explored in the past, e.g. [3], but never on a large scale with multiple books where multi-document summaries and reviews are necessary. This is increasingly important to support users in managing their personal digital libraries. In order to address this need, we propose an active reading task with the aim to explore suitable user interfaces able to support the various types of reading activities in a multiple book environment. The task will focus on a subset of books aimed at readers of different communities. The evaluation will be through user studies which will focus on different facets of usability and engage users with realistic tasks in order to assess overall levels of interest.

References

1. van Zwol, R., van Loosbroek, T. Effective Use of Semantic Structure in XML Retrieval. ECIR 2007: 621-628
2. Adler, M.J., van Doren, C. How to Read a Book. Simon and Schuster, New York, NY. 1972
3. Crossen, A., Budzik, J., Warner, M., Birnbaum, L., Hammond, K. J. XLibris: an automated library research assistant. IUI 2001: 49-52.