BooksOnline'11: 4th Workshop on Online Books, Complementary Social Media, and Crowdsourcing

Gabriella Kazai
Microsoft Research
Cambridge, UK
v-gabkaz@microsoft.com

Carsten Eickhoff
Delft University of Technology
Delft, The Netherlands
c.eickhoff@tudelft.nl

Peter Brusilovsky
University of Pittsburgh
Pittsburgh, USA
peterb@pitt.edu

ABSTRACT

The BooksOnline Workshop series aims to foster the discussion and exchange of research ideas towards addressing challenges and exploring opportunities around large collections of digital books and complementary media. The fourth workshop in the series, BooksOnline'11 pays special attention to the role of social media and the phenomena of crowdsourcing in the context of online books, which is expected to be key in defining new user experiences in digital libraries and on the Web. The workshop boasts a high quality program, including keynote addresses by Ville Miettinnen, CEO of Microtask and Adam Farquhar, Head of Digital Library Technology at The British Library. From the accepted papers two main themes became salient: 1) Information retrieval and information extraction methods focused on enhancing digital libraries, and 2) Studies and analyses of reading experience and behavior. This paper provides an overview of the workshop and the accepted contributions.

Categories and Subject Descriptors: H.3.7 [**Digital Libraries**]

General Terms: Experimentation, Human Factors

1. INTRODUCTION

In recent years, the number of digitally available books has increased dramatically through the digitization of physical books and by electronic publishing. Large scale digital libraries hold significant value to humanity by preserving knowledge and making it widely accessible. Furthermore, they show great potential for cross-media integration and industrial exploration. Some of the leading initiatives include Project Gutenberg (http://www.gutenberg.org), the Million Book project (http://www.ulib.org), and the Open Content Alliance (http://www.opencontentalliance.org). At the same time eBooks and eReaders are gaining wide acceptance and popularity. This is paralleled with social networking and content sharing platforms that accommodate millions of users who often dedicate great efforts to data collection, creation and verification. The harnessing of these considerable forces has the potential to revolutionize online reading both technically and in terms of interaction paradigms.

To match the great momentum in creating on-line book repositories, the BooksOnline workshop series aims to foster research initiatives that are focused on innovation opportunities and challenges created by large collections of digital books and complementary media. This year, the workshop focuses more explicitly and deliberately on exploring the role that social media may hold in enhancing digital libraries and online repository services. Topics include new interaction paradigms that define new, immersive and social reading experiences and novel methods to leverage the wisdom of the crowd as well as the necessary infrastructure.

The remainder of the paper summarizes the accepted contributions, grouped by the common themes that emerged.

BooksOnline'11, October 24, 2011, Glasgow, Scotland, UK.
ACM 978-1-4503-0961-5/11/10.

2. ACCEPTED PAPERS

2.1 System Focus: Information Retrieval and Extraction

Among the contributions that focus on aspects of systems, the common theme is on identifying important nuggets of information from books, from author names and bibliographic references to factual statements or passages. Another recurring issue is that of OCR errors.

Cartright, Field and Allan tackle the novel and challenging task of Evidence Finding in a collection of digitized books [2]. Its goal is to find confirming or refuting evidence in books for natural language assertions. The authors highlight the differences between the introduced problem and related fields such as Question Answering (QA). The approach taken and evaluated in this work tries to generate search engine queries based on factual assertions. These queries are subsequently issued to an index of 50,000 digitally scanned books from the INEX Book Track test collection. A preliminary evaluation reports superior performance to straight-forward bag-of-words and Sequential Dependence models.

Kim, Bellot, Faath and Dacos propose an automatic information extraction scheme for bibliographic references from scientific literature in the humanities [5]. On a sizeable corpus, they employ Conditional Random Fields to extract reference fields such as author names or publication titles with high precision. Particularly encouraging results are reported for the distinction of author first and last names where significant improvements over the state of the art could be achieved.

Smith, Manmatha and Allan in [9] focus on collections of books, instead of individual books, and demonstrate the potential benefits of the 'relational structure' of collections. Examples of relational structure include the relationships among books that are duplicates or translations of each other, or the citation or quotation of one passage of a book by another. With the aim to give scholars more powerful

tools, the authors present technical solutions to inferring book relationships, aligning books for structure discovery, and metadata propagation. An advantage of the proposed solutions is the promise they hold for mitigating other problems, such as OCR errors, with scanned book collections.

Kamps in [4] investigates the effectiveness of various IR ranking methods on library catalogue data and compares these with results from the online public access catalogue (OPAC). His findings demonstrate the importance of author information for ranking books. He shows that ranking books based on author scores derived from expert finding methods leads to substantial and significant performance improvements over standard book ranking methods, which already outperform OPAC. On the other hand, ranking books based on the combination of the author scores and the book retrieval scores yields no further improvement. This provides further evidence that author information captures an important aspect of relevance. This then highlights the challenge of how to support users in judging the relevance of authors when exploring a new and unfamiliar subject area or topic.

2.2 User Aspects: Experience and Behaviour

The contributions in this category focus on studying or designing rich, interactive and social reading experiences and tools and concepts to support or to evaluate these.

De Ribaupierre and Falquet describe a preliminary survey concerning the reading behavior of research professionals [8]. They argue that document retrieval will benefit from a more facet-oriented indexing to enable precise querying for desired information. These facets are envisioned to follow discourse elements such as definitions, hypotheses, algorithms, etc., rather than structural entities (e.g., conclusions or chapters). According to this scheme the authors foresee significant benefits for scientific literature study.

Colombo and Landoni in [3] describe progress on the HEBE project which aims at studying how to design eBook interfaces (or eBook architecture) for children in order to provide them with a highly engaging reading experience. They detail the results of a preliminary observational study of 45 school children as they interact with books and argue that reading is a complex experience with an important social dimension. They relate their findings to Csikszentmihalyi's flow theory, where flow is defined as a mental state of deep enjoyment and intense engagement in a certain activity, where most of a person's attention is devoted to accomplish that activity. The paper also presents interesting questions on how to address the challenge of defining and embracing a methodology where children can take an active role in the design process.

Sofronijevic introduces a novel interaction paradigm between reader and text, *Reading 2.0* [10]. Given this paradigm, a substantially higher degree of interactivity is assumed. As opposed to the traditional reading process, the *Solitary Reading*, Reading 2.0 is described as a more interactive and iterative process in which the boundaries between content producer and content consumer are less clear-cut, as individuals annotate, highlight or even extend and alter books. The open question remains how digital library research and engineering can facilitate the merits of Reading 2.0.

The paper by Regan [7] combines art and science in developing a book text visualization tool for analyzing the works in Philip Pullman's children's literature trilogy, His Dark Materials. The tools allow users to explore the rhythm of the characters' occurrences in the books, the relationships between the characters and the way in which they are depicted. The paper also gives brief insights of the experiences of the author and fans of the book with the tool. The author closes with a question that challenges current publishing practices and the laws that govern those practices.

Müller and Maurer in [6] introduce the concept of an Interactive Internet Book, which augments digitized books with four main layers: 1) the Facsimile layer delivers high quality device-dependent format for each page of a book, 2) the OCR layer results in the textual content presented to the user, 3) the Enhancement layer provides functionality for personal and group annotation, such as digital page markers, links, notes, highlights and even nano-publications, and 4) the Communication layer provides for group interactions and social media networks, such as sharing, social tagging, discussions, reading history, and search agents. The authors claim that the quality and the enhancements of an Interactive Internet Book go far beyond what is traditionally assumed. They demonstrate this by means of a use case study and a working prototype system.

Becker describes a HIT-driven quality assurance system for document conversions in digital libraries [1]. The author motivates the need for a hybrid system that exploits the benefits of large scale document archives and human computing. The central challenge addressed in this paper is the evaluation of OCR output. The author showcases how this task can be effectively and efficiently represented as a crowdsourcing HIT. The discussion gives useful insights into caveats and lessons learned for this domain.

3. ACKNOWLEDGMENTS

We would like to thank all authors who submitted contributions and the PC members for their excellent work.

4. REFERENCES

[1] C. Becker. Quality Assurance in Document Conversion: A HIT? In *Proc. CIKM BooksOnline Workshop*, 2011.

[2] M.-A. Cartright, H. A. Field, and J. Allan. Evidence Finding using a Collection of Books. In *Proc. CIKM BooksOnline Workshop*, 2011.

[3] L. Colombo and M. Landoni. Towards an engaging e-Reading experience. In *Proc. CIKM BooksOnline Workshop*, 2011.

[4] J. Kamps. The Impact of Author Ranking in a Library Catalogue. In *Proc. CIKM BooksOnline Workshop*, 2011.

[5] Y.-M. Kim, P. Bellot, E. Faath, and M. Dacos. Automatic Annotation of Bibliographical References in Digital Humanities Books, Articles and Blogs. In *Proc. CIKM BooksOnline Workshop*, 2011.

[6] H. Müller and H. Maurer. How to carry over historic books into social networks. In *Proc. CIKM BooksOnline Workshop*, 2011.

[7] T. Regan. Tools for Whom: Readers, Fans, or Authors? In *Proc. CIKM BooksOnline Workshop*, 2011.

[8] H. D. Ribaupierre and G. Falquet. New trends for reading scientific documents. In *Proc. CIKM BooksOnline Workshop*, 2011.

[9] D. A. Smith, R. Manmatha, and J. Allan. Mining Relational Structure from Millions of Books. In *Proc. CIKM BooksOnline Workshop*, 2011.

[10] A. Sofronijevic. Changes in Reading Research Proposition: Some Psychological Aspects of Reading 2.0. In *Proc. CIKM BooksOnline Workshop*, 2011.