User-centered Content Provisioning
over Large Collections of eBooks

Sascha Tönnies[1] and Wolf-Tilo Balke[1,2]

[1] L3S Research Center, Appelstraße 9a, 30167 Hannover, Germany
[2] IFIS TU Braunschweig, Mühlenpfordstraße 23, 38106 Braunschweig, Germany
toennies@L3S.de, balke@ifis.cs.tu-bs.de

Abstract. Managing and distributing published information is traditionally the mission of libraries. But in times of digital information provisioning and personalized content delivery, the processes to fulfill this mission have to be reconsidered. Beyond simple keyword indexing using library categorization systems, digital corpora need to be preprocessed for later access directly by the end user. Thus, major functions of the classical librarian like assessing the actual information need and mediating between the library categorization and the end user are to some degree bypassed and have to be compensated for. Moreover, also the quality control of a digital library's metadata annotations used for subsequent querying of collections has to be guaranteed. In this paper we discuss the importance of metadata quality control for large eBook collections.

Keywords: eBook, Content Provisioning, Quality Assurance, Digital Libraries

1 Introduction

Today traditional libraries are still the role model in the field of successful content provisioning. In particular, large stacks of books are managed, indexed, and delivered to a heterogeneous community of customers, where each customer is individually served. Recently, user-centered alerting services and document delivery services like e.g., *subito* (http://www.subito-doc.com), a service of libraries in Germany, Austria, and Switzerland, or *British Library Direct* (http://direct.bl.uk/) for the UK, provide quick and easy-to-use access to large collections of books. Basically, the business model is to make on-demand scans of articles from periodicals or books, and then send them to the user, as well as to support the actual lending of books.

In times of digital information and the interconnectedness through the Web, it is of course only to be expected that libraries extend their service offering beyond lending or copying (parts of) physical books. Moreover, digitization projects for books are underway on a grand scale also driven by companies. Well known examples include Google Books (http://books.google.com/) with an estimated 7 million volumes in 2008, and the Open Content Alliance's Open Library (http://openlibrary.org) capturing about 1 million volumes, as well as bibliographical references for 22.6 million books as of 2008. In addition, vendors like Amazon.com are already delivering electronic versions of current books (with an estimated 300,000 volumes as of 2009).

Since the traditional librarian is to some degree bypassed in Web-based systems, electronic access to digitized corpora is managed in different ways. For instance, Amazon.com enables the customer to download an eBook from a limited choice of popular books (generally the New York Times bestsellers) onto the Kindle device. But customers already have problems with this kind of access, due to the lack of organization possibilities: there is, e.g., no possibility to organize books within a folder structure and only a very limited full-text search is possible. On the other hand, the Google Book search and the Open Library project enable users to perform a full-fledged Web-search style access to books basically indexing all the keywords in a book. Still, all these portals lack a *user-centered* access to the corpora that would be comparable to traditional library services.

2 Controlling Metadata Quality for Querying eBook Collections

Since eBooks are quite complex in terms of their content, the question is how do users find the 'right' book? Possible approaches to solve this problem range from strict professional classification using e.g., the Library of Congress subject headings (LCSH) or the Dewey Decimal System (DCC), to full-text searches over the entire content of the books. However, both approaches have their limitations. By strictly categorizing books using a certain system, experienced users are able to get high precision. But still, without some experience and a well-maintained classification system a digital library will not provide a user-centered access. In contrast, using full-text search will result in a very high recall, but usually rather low precision depending on the query terms.

2.1 The Query Process

A solution to the problem of effective information access can be provided by breaking large eBook collections down to specialized topic-centered digital libraries. Such focused collections do not index the whole portfolio of available books, but just those books interesting for their domain. Besides the topical focus, the major success factor is the respective *user interface* and the (generally metadata based) *query facility*.

In terms of suitable interfaces information visualization is becoming increasingly prevalent for understanding and explaining information. Currently, faceted navigation is a popular technique for supporting exploration and discovery of digital libraries and document collections. Facets refer to different kinds of categories used to characterize information items in each corpus. However, the large-subject-space problem is still unresolved and makes innovative, yet understandable extensions of the faceted model essential [6]. A promising example is GoPubMed (http://www.gopubmed.org) providing ontology-based literature search over around 19 million biomedical research journals in the Medline collection.

Besides presentational issues, a major concern is the creation of suitable metadata. When describing collections information providers can rely on simple bibliographic metadata (like authors, book title, or publication year), or consider the content of each book. In this case the respective vocabulary and annotations to be considered valua-

ble, are largely depending on the specific content's domain. For instance, for eBooks in the domain of chemistry the extraction of chemical entities or reactions is needed, whereas for cultural heritage collections, a reconciliation of historic terms may be necessary. However, the manual annotation of collections is a cumbersome and expensive task. Hence, automated ways of metadata creation for digital documents are currently heavily investigated usually relying on statistical techniques like word co-occurrences or Semantic Web techniques, see e.g., [10, 3].

Because of the difficulties in automatically generating metadata, recently also the collaborative creation in form of tags and folksonomies has received considerable attention. Here, users annotate books or specific sections using a free vocabulary. The relevance (and thus to some degree correctness) of the resulting metadata is usually indicated by the number of individuals assigning a certain term. As an example consider LibraryThing (http://www.librarything.com), where users can assign tags, write reviews and rate books: looking at the commonly used tags also the limitations (like the frequent use of different spellings for the same concept, e.g., Science Fiction, science_fiction or sci-fi) of this approach for information filtering become clear.

2.2 Measuring the Quality of Metadata

Considering eBooks and quality assurance, the first possible source of errors lies within the process of digitizing books. Using OCR software the digitized result will always contain some OCR errors. Actually, this is not too problematic when building full-text indexes (since standard IR techniques like, e.g., TF/IDF are not really affected by unsystematic OCR errors). Still, for the process of tokenization or entity recognition in semantic classification techniques or in historic documents it definitely might already be an interesting factor affecting the overall retrieval quality [1].

The important part of introducing a convincing quality control for querying large digital collections lies in the metadata part. But while for traditional digital libraries the quality of (handcrafted) metadata may be measured in terms of completeness, correctness and relevance [8, 9], such metrics are difficult to obtain for semantically enriched digital collections. It is not sufficient to simply evaluate user satisfaction when employing semantically created metadata for querying, see e.g., [7], since users may look favorably upon the novelty of the interface rather than assess the retrieval effectiveness. The real benchmark of any information provider or library in the digital age lies in assessing the quality of the underlying metadata.

In the domain of collaborative tagging systems, some work already investigated tag quality, see e.g., [2]. Moreover, it has been shown that the distribution of different tags for each individual digital item tends to stabilize over time [4, 5]. Considering these properties of tagging systems, it seems likely that although not all tags are useful descriptors for resource sharing, social metadata can still be used as a reliable source of information.

On the other hand quality evaluations of semantic techniques for metadata generation are still rare. A good example is [11] where measures for the quality of automatically created taxonomies for classification are investigated. Major problems lie in the absence of 'gold standards' for benchmarking, as well as the loss of focus for individual instances when using purely statistical assumptions in semantic techniques.

3 Outlook

Given the extent and growth of digital content available in today's collections and libraries, it is essential to develop new methods of (automatically) generating suitable metadata. But this metadata can only play a significant role in information access, if ways to measure its quality with respect to the annotated resources, are researched in parallel. In this paper we have argued that given the scale of current digitization projects for eBooks, neither the purely bibliographical indexing of collections, nor a simple full-text search of the digital content can be a satisfying solution to this problem. However, first approaches in the area of measuring the reliability of metadata – either crafted by social processes (like tagging), or generated by intelligent semantic techniques (like automated taxonomy generation) – already show promising results. Now systematic research for developing suitable quality measures has to follow. Moreover, we believe that breaking down the choice and evaluation of measures to the specific domains of the respective digital content is a necessary prerequisite for successful application.

References

1. Abdulkader, A. and Casey, M.: Low Cost Correction of OCR Errors Using Learning in a Multi-Engine Environment. In *International Conference on Document Analysis and Recognition*, Barcelona, Spain, 2009.
2. Bischoff K., Firan C., Nejdl W., Paiu R.: Can all tags be used for search? In *ACM Conference on Information and Knowledge Management*. Napa Valley, CA, USA, 2009.
3. Cimiano P., Handschuh S., Staab S.: Towards the self-annotating web. In *International Conference on the World Wide Web*, New York, NY, USA, 2004.
4. Golder S.A., Huberman B.A.: Usage patterns of collaborative tagging systems. In *Journal of Information Science*, Vol. 32(2), 2006.
5. Halpin H., Robu V., Shepherd H.: The complex dynamics of collaborative tagging. In *International Conference on World Wide Web*, Banff, Alberta, Canada, 2007.
6. Hearst, M.: UIs for faceted navigation: recent advances and remaining open problems. In *Workshop on Human-Computer Interaction and Information Retrieval*, Redmond, WA, USA, 2008.
7. Kruk S.R., Kruk E., Stankiewicz K.: Evaluation of semantic and social technologies for digital libraries. In *Semantic Digital Libraries*, Springer, 2009.
8. Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., and Manitsaris, A.: A conceptual framework for metadata quality assessment. In *International Conference on Dublin Core and Metadata Applications*, Berlin, Germany, 2008.
9. Nichols, D. M., Chan, C., Bainbridge, D., McKay, D., and Twidale, M. B.: A lightweight metadata quality tool. In *ACM/IEEE Joint Conference on Digital Libraries*, Pittsburgh, PA, USA, 2008.
10. Sanderson M., Croft B.: Deriving concept hierarchies from text. In *ACM SIGIR Conference on Research and Development in Information Retrieval*. Berkeley, CA, USA, 1999.
11. Tönnies, S. and Balke, W-T.: Using semantic technologies in digital libraries - a roadmap to quality evaluation. In *European Conference on Research and Advanced Technology for Digital Libraries*, Corfu, Greece, 2009.