Semantic Bookworm: Mining Literary Resources Revisited

Annika Hinze, Michael Coleman, Sally Jo Cunningham, David Bainbridge
Computer Science Department, University of Waikato
Hamilton New Zealand
hinze@waikato.ac.nz, mjc62@students.waikato.ac.nz, {sallyjo, davidb}@waikato.ac.nz

ABSTRACT
In this paper, we describe Semantic Bookworm—a tool that supports scholarly text analysis. In contrast to the text-based Bookworm tool, the Semantic Bookworm identifies semantic concepts.

CCS Concepts
• **Information systems → Digital libraries and archives**
• **Computing methodologies → Lexical semantics**
• **Computing methodologies → Semantic networks**

Keywords
Semantic Analysis, Digital Humanities, Text Mining, Data Mining

1. INTRODUCTION & BACKGROUND
The large-scale digitization of print material makes it possible for scholars to expand the scope of their study to large—indeed, massive—collections of documents. Insights from traditional 'close reading' of small sets of texts now have the potential to be complemented by 'distant reading' [2] over these newly available digital resources, as data mining is brought to bear for the identification of unexpected patterns in corpora too large to be read.

This new style of humanistic research requires new tools to support analysis. One strand of tool development focuses on visual analysis, for example of the layout and structure of texts treated as images [5]; this approach obviously is particularly well-suited to documents that natively are primarily images. This present paper focused on a second, text-centered strand of research that treats a document as a sequence of *ngrams*, where decisions about the size of *n* and the choice of token as characters, syllables, or words are fundamental to this type of research. An ngram analysis tool is used to identify and quantify the occurrences of the specified token in the corpus (at a minimum providing a sum of the number of occurrences, and perhaps also offering statistical analysis of observed token frequencies as well), and then to visualize the occurrences (for example, by plotting frequency of token occurrence across time or against the structure of the underlying documents). Existing ngram analysis tools used in digital humanities research include the Ngram Statistics Package [1], Bookworm[1] [4], and the Google Ngram viewer[2] [5]. Recently a bookworm for the 4.6M public domain texts of the HathiTrust has become available.[3]

JCDL '16, June 19-23, 2016, Newark, NJ, USA
ACM 978-1-4503-4229-2/16/06.
http://dx.doi.org/10.1145/2910896.2925444

While these ngram analysis tools are demonstrably supporting new styles of humanities research that heretofore were not possible, the limitations of the tools are also becoming clear. Schmidt [7], for example, explores issues in using ngram analysis to tracking references to "the Enlightenment" as a historical period; the difficulties reported are caused to a large extent by the ngram analysis stripping away all cues (contextual, rhetorical, typographical) that allow a human reader to distinguish spurious uses of that phrase.

This paper introduces a *Semantic Bookworm* as a step towards combining the semantic, discriminative facilities of the human reader with the speed and coverage supported by automated mining tools. With Semantic Bookworm, not every occurrence of a word or phrase is counted, but rather the occurrence of a semantic concept. Section 2 introduces the architecture and implementation of the Semantic Bookworm, and Section 3 illustrates differences in results achievable with the Semantic Bookworm in comparison to the ngram-based (lexicographic) Bookworm.

2. SEMANTIC BOOKWORM
The original Bookworm software (Figure 1, blue and gray components) analyses a given document collection *lexically* using ngram analysis. The ngrams are then stored ready for search. The search function performs a look-up in the ngram index to identify the frequency of the ngram within the collection, and results are then visualized using, e.g., time graphs (examples in Sect. 3).

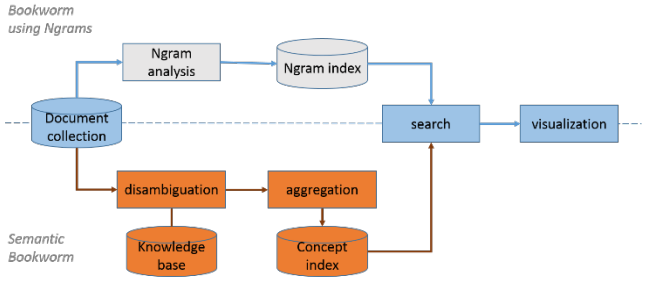

Figure 1: Combined architecture of Bookworm (original components blue and gray) and Semantic Bookworm (semantic components orange)

The Semantic Bookworm replaces those components that are shown in gray in Figure 1 by the components shown in orange. Here the document collection undergoes a disambiguation step using the knowledge base of semantic concepts we developed for Capisco [2]. In this process, all terms appearing in each document are mapped to their respective concept. This means that terms that are homonyms (same spelling, different meaning) may be linked to different concepts. Some terms may not be linked to any concept at all, as only those concepts are registered that are relevant to the document. The resulting concept information is then aggregated

[1] https://github.com/Bookworm-project/BookwormDB

[2] https://books.google.com/ngrams

[3] http://bookworm.htrc.illinois.edu/

and stored in a concept index. For search and visualization, we use the original Bookworm components.

3. EXAMPLE VISUALISATIONS

We present two examples comparing the results of lexicographic analysis (using the original Bookworm) and concept analysis (using the Semantic Bookworm).

3.1 Analyzing books' structure

In the first example, we analyze two books by Charles Dickens, *David Copperfield* (orange line in Figures 2 and 3) and *Great Expectations* (blue line). The sequence of chapters is treated as a chronology in the Bookworm timeline, and the graphs plot the number of occurrences of terms (Figure 2) and concepts (Figure 3).

We here contrast the occurrences of the term "night" (counting each ngram) with the concept *Night* (counting each time the text deals with the concept rather than, for example, its occurrence in the farewell 'Good night'). Comparing these two sets of graphs, we note that the term "night" is overall mentioned more often throughout the books. This includes, for example, every mention of the word in phrases such as 'Good night'. We also note that even though in "Great expectations" the term "Night" is mentioned only somewhat less than in "David Copperfield", the concept of *Night* hardly ever appears in the former book and quite frequently in the latter.

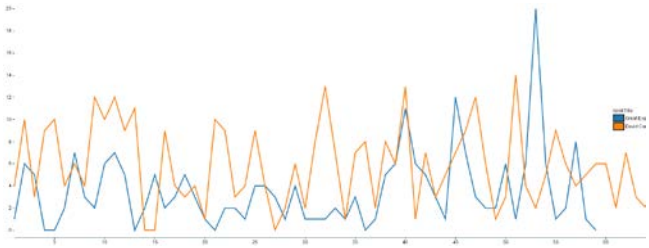

Figure 2: Term "night" across chapters of Dickens' books

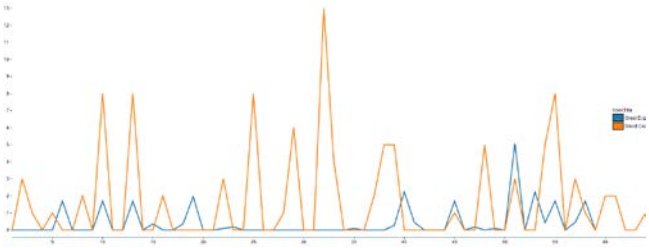

Figure 3: Concept *Night* across chapters of Dickens' books

Figures 4 and 5 show the distribution of the term and the concept of time, respectively. We note that overall the concept *time* occurs much more often than the term "time".

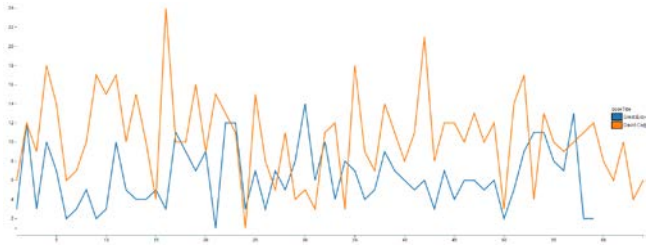

Figure 4: Term "time" across chapters of Dickens' books

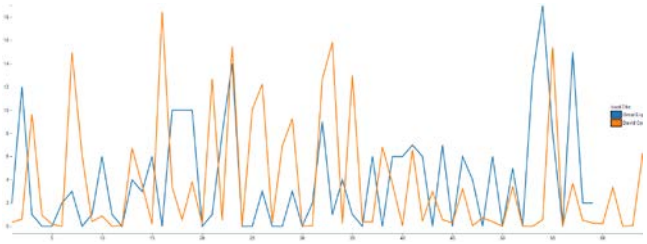

Figure 5: Concept *Time* across chapters of Dickens' books

3.2 Analyzing a collection structure

The second example analyses a collection of Charles Dicken's letters (here shown from 1835 to 53). We compare Dickens' use of the terms "god" (blue) and "philosophy" (orange) in Figure 6, left, with the occurrence of the concepts *God* and *Philosophy*, right. One can see that even though the term "Philosophy" is rarely used, topics of a philosophical nature appear quite often. On the other hand, the term "god" is used more often than the concept.

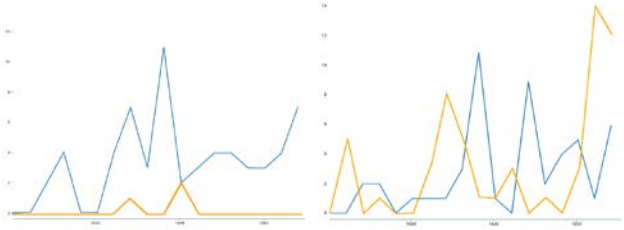

Figure 6: Bookworm (left) and Semantic Bookworm (right)

4. REFERENCES

[1] Banerjee, S., and Pedersen, T. 2003. The design, implementation, and use of the ngram statistics package. *Computational Linguistics and Intelligent Text Processing*, 370-381.

[2] Hinze, A., Taube-Schock, C., Bainbridge, D., Matamua, R., & Downie, J. S. 2015. Improving access to large-scale Digital libraries through Semantic-enhanced Search and Disambiguation. *JCDL'15*, 147-156.

[3] Kirschenbaum, M.G. 2007. The remaking of reading: Data mining and the digital humanities. *The NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation.* 2007

[4] Leonard, P. 2014. Mining large datasets for the humanities. *IFLA WLIC 2014 Libraries, Citizens, Societies: Confluence for Knowledge*, http://library.ifla.org/930/

[5] Lin, Y., Michel, J.B., Aiden, E.L., Orwant, J., Brockman, W. and Petrov, S. 2012. Syntactic annotations for the Google Books ngram corpus. *ACL system demonstrations*, 169-174.

[6] Rushmeier, H., Pintus, R., Yang, Y., Wong, C. and Li, D. 2015. Examples of challenges and opportunities in visual analysis in the digital humanities. *IS&T/SPIE Electronic Imaging*, 939414-939414.

[7] Schmidt, James. 2013. Tracking 'the Enlightenment' Across the Nineteenth Century. *International. Conference on the History of Concepts*, 30-39.