Search and Exploration of Scanned Books

Marc-Allen Cartright, Jeff Dalton, and James Allan

Center for Intelligent Information Retrieval
Dept. of Computer Science
Univ. of Massachusetts
Amherst, MA 01003

ABSTRACT

In this demo, we present Proteus, a novel interface for interacting with multiple retrieval types extracted from scanned books provided by the Internet Archive. The primary purpose of Proteus is to provide a rich interactive experience for users to explore collections with automatically extracted and linked entity data. The system supports seamlessly *shifting perspectives* between books, entities, and topics. Proteus provides a starting point for a variety of exploratory search tasks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search Process*; H.5.4 [**Information Storage and Retrieval**]: Hypertext/Hypermedia—*Navigation*

Keywords

Proteus, object search, navigation

Introduction

In this demonstration we present an extension of the Proteus search system, a system designed for search and exploration over multiple retrieval types [1]. We extend the previous system in several directions. First, we improve the automatic entity extraction with models that incorporate historic context, which significantly reduces noisy entity links. Second, we introduce new navigation leveraging temporal and geographic data that allows browsing of links across time and place. Lastly, we update Proteus to provide more robust and flexible navigation across multiple types of extracted data. We designed Proteus to enable users to not only search over the text and entities of a collection, but to navigate over the connections between entities and topics by *shifting perspectives*. Consequently, Proteus provides an excellent platform for a variety of exploratory search tasks, including Evidence Finding [2].

The BooksOnline workshop provides an excellent forum to test the system against users to receive feedback and allow us to further refine the functionality and user experience of Proteus. We now describe Proteus in detail.

Related Work

BooksOnline'12, October 29, 2012, Maui, Hawaii, USA.
ACM 978-1-4503-1714-6/12/10.

Digital Libraries have been an active area of research for many years—at least two separate conferences are devoted solely to the advancement of digital library research[1] However for the sake of brevity, we describe two systems that we believe resemble Proteus in purpose, and explain how Proteus differs from these other implementations.

Daffodil, now called ezDL[2], is a desktop application designed to enable search in digital libraries [4]. While the source data is the same between ezDL and Proteus, ezDL primarily works with fully digital libraries and the associated metadata. In contrast, Proteus was built to not only navigate over the explicit elements in a digital library (books), it also aims to navigate over extracted retrieval types such as people, locations, and topics.

The Perseus Project[3] was an initiative started in 1985 to bring classics texts online in an effort to make them more accessible to a larger number of classicists [3]. The aims of Proteus and Perseus are similar, however Perseus is a mostly hand-curated collection, compiled over decades of work and editing. Proteus aims to perform as much as possible of this entity identification and resolution automatically. Additionally, Perseus targets several very specific collections, while we intend to make Proteus easily retargetable to a wide variety of digital archive collections.

Proteus in Action

Figure 1 shows a sample result page in Proteus. For any given query, a user can search over multiple types, as shown in the dotted blue box. We provide the ability to retrieve books, pages, people, locations, images, and unsupervised topic clusters. As shown in Figure 1 the entities are linked to real-world objects from Wikipedia and annotated with structured metadata (including pictures, and geographic data). The linked data provides a flexible way for users to start with an entity of interest from Wikipedia and find references to it directly across the collection. Beyond an entity, the flexible search interface can support the evidence finding task, which starts from a statement in text that has been highlighted by a user [2].

In addition to retrieval over multiple types, Proteus provides structured navigation from any combination of selected relevant objects. Selected objects create a group that defines

[1]The Joint Conference on Digital Libraries (JCDL), based at `http://http://jcdl.org/`, and The Theory and Practice of Digital Libraries (TPDL), found at `http://www.tpdl.eu/`.
[2]`http://www.ezdl.de/home`
[3]`http://www.perseus.tufts.edu/hopper/`

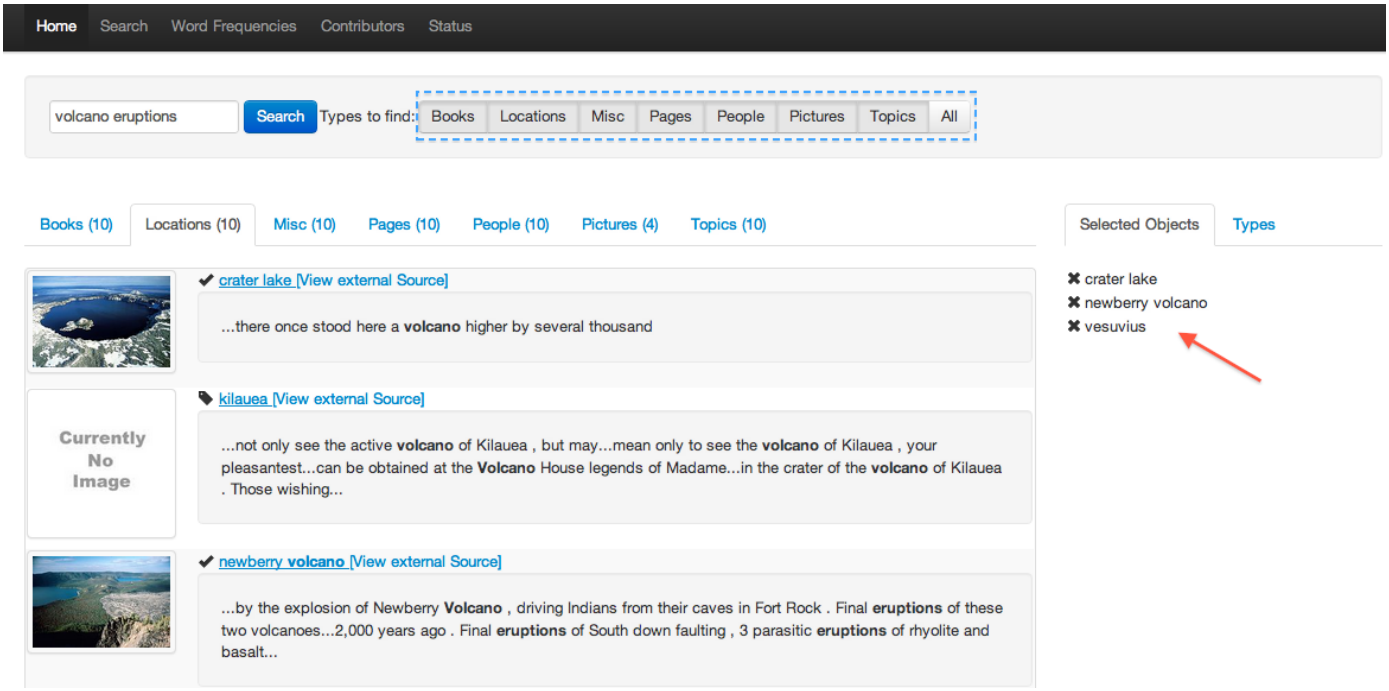

Figure 1: A sample result page in Proteus. (The dashed blue box and red arrow are for reference only).

the context of the subsequent search. Selected objects appear in the listing on the right, indicated by the red arrow in Figure 1. We currently return objects that are the most frequently mentioned with the selected objects. Many questions remain concerning how to best implement cross-type ranking and exploration. Should the search only involve selected types? Should the relevance be "anchored" to a topic using the original query? How should novelty and serendipity be incorporated? How can selected objects be used as a form of relevance feedback? Proteus provides a platform for us to experiment with these and similar questions.

Proteus System Architecture

Proteus is a system for for annotating, indexing, and searching across objects extracted from digital text collections. The source code is open source and freely available.[4] The input is OCR output from scanned books provided by the Internet Archive. The second step is NLP annotation, which is performed by the Factorie [5] NLP framework and includes sentence detection, true-casing, part of speech tagging, syntactic parsing, and named entity recognition. Third, the identified entities are linked externally to Wikipedia and internally. Lastly, the books and entities are indexed using MapReduce. During indexing, entity language models (ELMs) are constructed using the surrounding text.

The text indices and cross-document link data are searched using the Galago[6] search engine. A search integration service across the multiple data types is provided by a middleware data API layer, Aura, which is built upon the Finagle[7] network stack, a framework for high performance asynchronous RPC clients and servers. The data API is defined using the Thrift interface language, which allows support for data clients across a wide variety of languages. The separation of the data model provides a mechanism for creating external programmatic APIs, including the user interface. The user interface is based on the Scalatra web framework and employs the Bootstrap CSS/Javascript library to render results. The UI framework provides seamless scaling of Proteus across both desktop and mobile platforms.

[4] https://github.com/CIIR/Proteus
[5] http://code.google.com/p/factorie/
[6] http://www.lemurproject.org/galago.php
[7] http://twitter.github.com/finagle/

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0910884. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

REFERENCES

[1] M.-A. Cartright, E. F. Can, W. Dabney, J. Dalton, L. Giorda, K. Krstovski, X. Wu, I. Z. Yalniz, J. Allan, R. Manmatha, and D. A. Smith. A Framework for Manipulating and Searching Multiple Retrieval Types. In *Proc. SIGIR 2012*, pages 1001–1001, New York, NY, USA, 2012. ACM.

[2] M.-A. Cartright, H. A. Feild, and J. Allan. Evidence Finding Using a Collection of Books. In *Proc. BooksOnline 2011*, pages 11–18, New York, NY, USA, 2011. ACM.

[3] G. Crane. The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology. Technical report, Tufts University, 1998.

[4] N. Fuhr, C.-P. Klas, A. Schaefer, and P. Mutschke. Daffodil: An Integrated Desktop for Supporting High-Level Search Activities in Federated Digital Libraries. In *Proc. ECDL 2002*, pages 597–612, London, UK, UK, 2002. Springer-Verlag.