2012

A Framework for Manipulating and Searching Multiple Retrieval Types

Marc-Allen Cartright
Ethem F. Can
William Dabney
Jeff Dalton
Logan Giorda, et al.

A Framework for Manipulating and Searching Multiple Retrieval Types

Marc-Allen Cartright, Ethem F. Can, William Dabney, Jeff Dalton,
Logan Giorda, Kriste Krstovski, Xiaoye Wu, Ismet Zeki Yalniz,
James Allan, R. Manmatha, and David Smith

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

ABSTRACT

Conventional retrieval systems view documents as a unit and look at different retrieval types within a document. We introduce Proteus, a frame-work for seamlessly navigating books as dynamic collections which are defined on the fly. Proteus allows us to search various retrieval types. Navigable types include pages, books, named persons, locations, and pictures in a collection of books taken from the Internet Archive. The demonstration shows the value of multi-type browsing in dynamic collections to peruse new data.

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*

General Terms

Design, Human Factors, Standardization

Keywords

Proteus, object search, navigation

The idea of browsing over different types of information has been explored by various web sites and researchers for some time [1, 2]. Most efforts to date either collapse the types into one retrieval list or only use boolean-matching queries. These systems provide no way to maintain the separation between the types while performing boolean or ranked search while switching between types – for example, starting from a set of books discussing the trek of Marco Polo between Europe and Asia, then requesting ranked locations mentioned in that set of books and, from there, change to the people who are prominent at those locations, switching to books that discuss those people, and so on. We believe the next step in information systems is to afford seamless movement between boolean and ranked retrieval of such different retrieval types enabling users to easily negotiate this complex information space.

We introduce Proteus, a data model for representing, searching, and browsing objects extracted from book collections. The Proteus model defines an API for potential data sources to implement and use to provide and exchange information using that data model. Implementing this API allows Proteus to use the items from the system as part of the retrievable set of items, enabling horizontal movement between the "vertical" data types.

We demonstrate the value of Proteus with an implementation of the framework we have developed at the Center for Intelligent Information Retrieval. We have indexed a set of books from the Internet Archive. We use focused collections centered on topics such as William Shakespeare, the westward migration of American settlers, and the American Revolution. In each instance we provide the ability to retrieve pages, books, named persons, locations, and images from these collections, and navigate between them. We also integrate information from DBPedia to provide an existing ontology for the identified objects in the books.

We see the demonstration as an opportunity for users to browse a collection in a new way and discover information that would otherwise be difficult or impossible to deduce through previous navigation systems. The demonstration will be available at `http://books.cs.umass.edu` website.

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] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *Proc. SIGIR 2003*, pages 72–79, New York, NY, USA, 2003. ACM.

[2] C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries*, 6:124–138, 2006.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.