Mining and Analyzing Digital Archive Usage Data to Support Collection Development Decisions

Jewel Ward[†], Johan Bollen[‡], Jeffrey Pearson[†], Shing-Cheung Chan[†], Hui-Hsien Chi[†], Marie Chi[†], Kristine Guevara[†], Hsiao-han Huang[†], Genesan Kim[†], Maks Krivokon[†], Bo H. Lee[†], Pei-Han Li[†], Fenny Muliawan[†], Vu Nguyen[†], Barry W. Boehm[†], A. Winsor Brown[†], Edward Colbert[†], Alex Lam[†], Mayur Patel[†]

[†]University of Southern California
Los Angeles, CA 90089

{jewelw, jpearson, shingchc, hchi, mariechi, kguevara, hsiaohah, genesank, krivokon, bohlee, peili, muliawan, nguyenvu, boehm, awbrown, ecolbert, alexankl, mayurkup}@usc.edu

[‡]Old Dominion University
Norfolk, VA 23529

jbollen@cs.odu.edu

ABSTRACT

We demonstrate a "collection development decision support tool" that mines digital archive usage data. We want to better understand the University of Southern California (USC) Digital Archive's collection structure by analyzing the objects' characteristics, by analyzing the relationships between viewed objects, and by understanding usage trends over time. By relying on implicit patterns of usage data, such as co-retrievals, rather than explicit data, such as hit counts, we believe we can make more informed decisions about where to expend our resources.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Collection

General Terms: Algorithms, Management, Design

Keywords

Data mining, co-retrieval, usage analysis, collection development

1. INTRODUCTION

The USC Digital Archive currently holds around 122,000 intellectual objects of audio and image files divided amongst eleven collections searchable via one interface. Once eighteen remaining collections are moved from the legacy system to the current system by mid-year 2006, we will add completely new collections to the archive. We would like to augment our current process for determining what additional collections should be added by mining usage data from the existing collections.

2. IMPLEMENTATION

The UI is designed around the H3 viewer [1], which allows the user to cleanly conceptualize large quantities of data in context. Compared to a 2D graphic, the node-link 3D graphic better represents usage of a particular object and its relationship to other viewed objects within a common co-retrieval event cluster. Previous applications [2, 3] have examined co-authorship and usage trends based on web server log files, but the USC digital archive is a web application. Thus, the analysis tool pulls data from an Oracle database that contains key event data. The key event data includes the client IP address, the date and time the object was viewed, the object name, and the session ID number. The session ID number is used to determine unique user sessions. An algorithm [4] determines the importance of the co-retrieval events.

The demonstration of the tool will present a use case analysis in order to facilitate discussion of how data mining can support the collection development decision process.

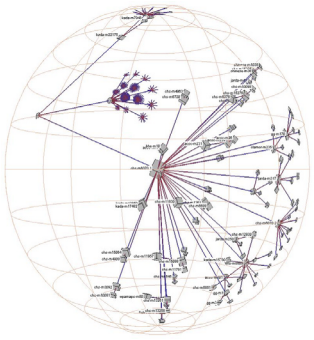

Figure 1. Collection structure based on object relationships

JCDL'05, June 7–11, 2005, Denver, Colorado, USA
ACM 1-58113-876-8/05/0006.

3. ACKNOWLEDGEMENTS

This project is based on work and supported in part by Herbert Van de Sompel and Rick Luce at LANL. We would like to thank Sara Tompson at USC for her contributions.

4. REFERENCES

[1] T. Munzner. H3 Viewer, 2002. http://graphics.stanford.edu/~munzner/h3/.

[2] J. Bollen and R. Luce. Evaluation of Digital Library Impact and User Communities by Analysis of Usage. *D-Lib Magazine*, 8(6), 2002.

[3] X. Liu, J. Bollen, M. Nelson, H. Van de Sompel, J. Hussell, R. Luce, and L. Marks. Toolkits for Visualizing Co-Authorship Graph. *Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries*, page 404, 2004.

[4] S. van Dongen. Markov Cluster Algorithm, 2000. http://micans.org/mcl/.