Andritsos, P., & Keilty, P. (2014). Level-Wise Exploration of Linked and Big Data Guided by Controlled Vocabularies and Folksonomies. Advances In Classification Research Online, 24(1). doi:10.7152/acro.v24i1.14670

Level-Wise Exploration of Linked and Big Data Guided by Controlled Vocabularies and Folksonomies

Periklis Andritsos
University of Toronto
Faculty of Information
140 St George Street
Toronto, ON M5S 3G6
periklis.andritsos@utoronto.ca

Patrick Keilty
University of Toronto
Faculty of Information
140 St George Street
Toronto, ON M5S 3G6
p.keilty@utoronto.ca

ABSTRACT

This paper proposes a level-wise exploration of linked and big data guided by controlled vocabularies and folksonomies. We leverage techniques from both Reconstructability Analysis and cataloging and classification research to provide solutions that will structure and store large amounts of metadata, identify links between data, and explore data structures to produce models that will facilitate effective information retrieval.

Keywords
Big Data, Subject Classification, Reconstructability Analysis.

INTRODUCTION

The amount of information stored in libraries is increasing in size. For example, the Scholars Portal service provided by the Ontario Council of University Libraries provides digital access to more than 500,000 books, more than 35 million journal articles. The rate at which such digital information is amassed does not allow the timely processing and analysis, as it requires domain expertise as well as increased computational resources.

In order to effectively retrieve these items, library catalogs have long relied on various forms of metadata, including controlled vocabularies, such as the Library of Congress Subject Headings, in order to facilitate users' access to items. Recently, many libraries have begun to augment controlled vocabularies with folksonomies, a system of classification derived from the practice of collectively managing and creating tags. As a result, library catalogs have given rise to an increasing amount of metadata.

While the amount of metadata has increased, a new mode of inquiry, problem solving, and decision-making has become pervasive in building large databases, consisting of applying computational and mathematical models to infer actionable insight and information from large quantities of data. Generally we think of big data analysis as applying computational and mathematical models to large quantities of data to answer sophisticated questions. This is a relatively recent mode of inquiry, resulting from advancements in the rates at which computers process large amounts of data. However, big data analysis, by definition, not only requires processing large quantities of data but also processing heterogeneous data, data that appears in a variety of formats (e.g. text, audio, video, structured databases, etc.), data that changes over time, and data that is generated at a rate much faster than it can be processed. Furthermore, data originate from distinct sources of varying quality and trustworthiness. There is, therefore, critical need to authenticate the provenance, quality, and veracity of data in any interpretation of it. As a result, big data creates new challenges for structuring data for effective information retrieval. The job of a data scientist is to manage such data and infer new insights.

In order to address these challenges with respect to controlled vocabularies and folksonomies, we turn to Reconstructability Analysis (RA) (Zwick 2004). RA is an approach for inducting modeling relationships and correlating variable data by taking a single source of data and clustering it into smaller subsets, each of which share a particular quality or trait (i.e. grouping like data with like data). Computer science research refers to this as a "level-wise exploration" of a data set, be it numerical or textual, where each level is a "decomposition" of models of data existing in levels of higher granularity. RA algorithmically assesses the content of each subset in relation to the larger, single source of data from which the data originates. This is important because it allows us to assess the context in which the data originally appears. RA is able to produce data set that are equivalent in their information content but differ in their original structure. For example, if we provide as an input the metadata of MARC records, RA will provide different models that include subsets of the MARC attributes with the same content.

To achieve its goal, RA uses set-theoretic or information-theoretic quantification of the information content in the input and attempts to output models with equal, or approximately equal quantities. In our case, we can consider the records stored in a library systems together with their controlled vocabularies and folksonomies. The

Advances in Classification Research, 2012, October 26, 2012, Baltimore, MD, USA.

ISSN: 2324-9773

different decompositions produced by RA will indicate which items in the library catalog best correlate with particular controlled vocabularies and folksonomies.

CONTROLLED VOCABULARIES AND FOLKSONOMIES

Library catalogs have long relied on controlled vocabularies to facilitate information retrieval. Controlled vocabularies are used in subject indexing schemes, subject headings, thesauri, taxonomies, and other knowledge organization systems. Controlled vocabularies mandate the use of predefined, authorized terms that have been preselected by the designer of a vocabulary, in contrast to natural language vocabularies, where there is no restriction on vocabulary. The Library of Congress Subject Headings is the predominant controlled vocabulary used by libraries throughout Canada and the United States. In recent decades, the role of controlled vocabulary has been called into question, as the standards and software underlying libraries did not anticipate performing controlled vocabulary on a widely disparate metadata from highly unreliable sources. Without controlled vocabulary, however, we are unable to meet our retrieval objectives of precision and recall (Svenonius 2003). Scholars have argued for the increasing importance of controlled vocabularies for maintaining our ability to give people a recognizable array of relevant choices (Mann 2003).

In the past decade, many libraries have experimented with folksonomies, a system of classification derived from the practice of collectively managing and creating tags, mindful that metadata can be mined in all sorts of ways (Guy and Tonkin 2006). Many library and information researchers have come to see the benefits of combining controlled vocabularies with folksonomies, in order to achieve the benefits of both (Adler 2009), while others have countered arguments that tags will always occur idiosyncratically by showing that social tagging stabilizes over time (Halpin et al. 2007 and Golder & Huberman 2006). By providing relevant search terms, controlled vocabularies and folksonomies eliminate some of the guesswork of finding relevant search terms and, therefore, provide one means for effectively retrieving information within a library catalog.

RECONSTRACTUBILITY ANALYSIS

Databases are made up of *records*, (e.g. for each article, book or video/audio), each records contains a set of standardized *fields* (or *attributes*), (e.g. for title, subject, author etc.) while each field is made up of *words*. (or *values*). Such databases are collected by libraries and they may also exist in disparate sources that get linked using techniques from the Linked Open Data are of research, (Xin, Hassanzadeh, et al. 2012). We assume that such large tables are the input to the exploration system we are proposing.

An example of a *lattice* that RA builds to explore the input data in a level-wise fashion is given in Figure 1.

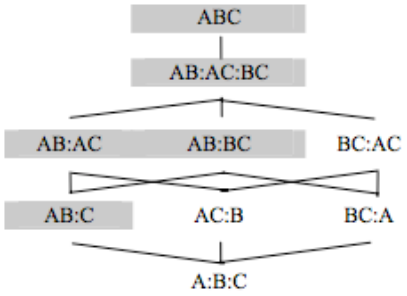

Figure 1: Reconstructability Analysis, (Zwick 2004)

Given a set of fields, A.B,C, RA constructs a set of models that indicate correlation of attributes. Sets of attributes separates by ":" are not correlated. For example, in Figure 1, the model "AB:C" indicates that A and B are correlated and are independent of C. Using optimization techniques, we may also skip some models, knowing that they can never have the same information content as the original data. In Figure 1, the analysis is done bottom-up and the greyed-out models will never be examined.

One advantage of RA is that it retains multiple models and, hence, we will be able to explore different ways in which the original fields can be correlated.

OUR APPROACH

In our approach we consider the fields provided by subject classifications and the tags provided by the users as fields that need to be explored together with the fields of the library database system. Our goal is to find out, by looking at the actual instance of library data, why some of the labels are attached to library items.

Considering Figure 1, if C is a tag then two outcomes are that

- from model "AC:B" we infer that it is correlated with field A,
- from model "BC:A" we infer that it is correlated with field B

By output several models equivalent to the original data, we empower librarians with automatic tools for the exploration of additional correlations as well as semantic information about the fields. Although we provide automatic techniques in order to sieve through vast amounts of textual data, we intend to complement the output correlations with user expertise to verify the correctness of the results.

CONCLUSIONS

We believe that our project comes at a moment in time where new systems for analyzing textual data sets are

ISSN: 2324-9773

needed. In fact, our plan is to use a real Big Data set from the Ontario Scholar's Portal, a service that provides over 140,000 academic publications in digital format. Such publications come from different research areas and fit the purpose of our project. Existing solutions explore the classification of individual documents. Digital Humanities can benefit from the exploration of several sources at the same time and the identification of commonalities and hidden knowledge the may exist when explored at the same time.

REFERENCES

Adler, M. (2009). Transcending library catalogs: a comparative study of controlled terms in Library of Congress Subject Headings and user-generated tags in LibraryThing for transgender books. *Journal of Web Librarianship* 3.4: 309-331.

Golder, S. and Huberman, B. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science* 32: 198-208

Guy, M. and Tonkin, E. (2006). Folksonomies: Tidying up tags? *D-Lib Magazine* 12. Retrieved from: http://www.dlib.org/dlib/january06/guy/01guy.html

Halpin, H.; Robu, V.; and Shephard, H. (2007). The complex dynamics of collaborative tagging. *WWW 2007: Proceedings of the 16th International Conference on World Wide Web*, May 8-12, 2007, Banff, Alberta, Canada: 211-220

Mann, T. (2003). Why LC subject headings are more important than ever. *American Libraries, 34,* 52-54.

Svenonius, E. (2003). Design of controlled vocabularies. In *Encyclopedia of library and information science* (pp. 822-838). New York: Marcel Dekker.

Xin, R. S.; Hassanzadeh, O.; Fritz, C.; Sohrabi, S.; and Miller, R. J. (2012). Publishing bibiliographic data on the Semantic Web using BibBase, Semantics Web – Interoperability, Usability, Applicability, IOS Press.

Zwick, M.. (2004). An overview of Reconstructability Analysis. Kybernets, vol.33, No. 5/6 (pp. 877-905).

ISSN: 2324-9773