Work-in-progress
[because of the methodological part and the demo character best to be presented as presentation]

Rob Koopman[1], Shenghui Wang[1], , Andrea Scharnhorst[2]
[1] OCLC Research Leiden, The Netherlands
[2] Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences, The Hague, The Netherlands

Between information retrieval services and bibliometrics research – new ways of semantic browsing and visual analytics

We are confronted with growing information spaces for which we have little possibilities to gain an overview effectively and efficiently. Information about the size and composition for instance of a library collection automatically conveyed when visiting the library physically, is made invisible by on-line interfaces. In other words we trade remote access for intangibility. [2] In the last couple of years, this problem has been explicitly addressed by Information Retrieval specialists [3]. Incidentally, we see *macroscope* like structures [4] emerging, which allow the visitor to browse through an information space also remotely, and to experience serendipity [5].

We report in this paper about a recent initiative of OCLC to create an alternative access to its products, among them ArticleFirst, a database with 65 million articles. The interactive interface ARIADNE has been created at the beginning of 2015, and is continuously updated since [1]. (http://thoth.pica.nl/relate)

The main idea behind the ARIADNE interface is to create a browseable representation of related and interlined context for any query. By context we mean related terms, authors, journals, but also Dewey classifications, and other parts of the bibliographic record. Those are visualized in the form of a network where more related entities are positioned closer to each other and vice versa.

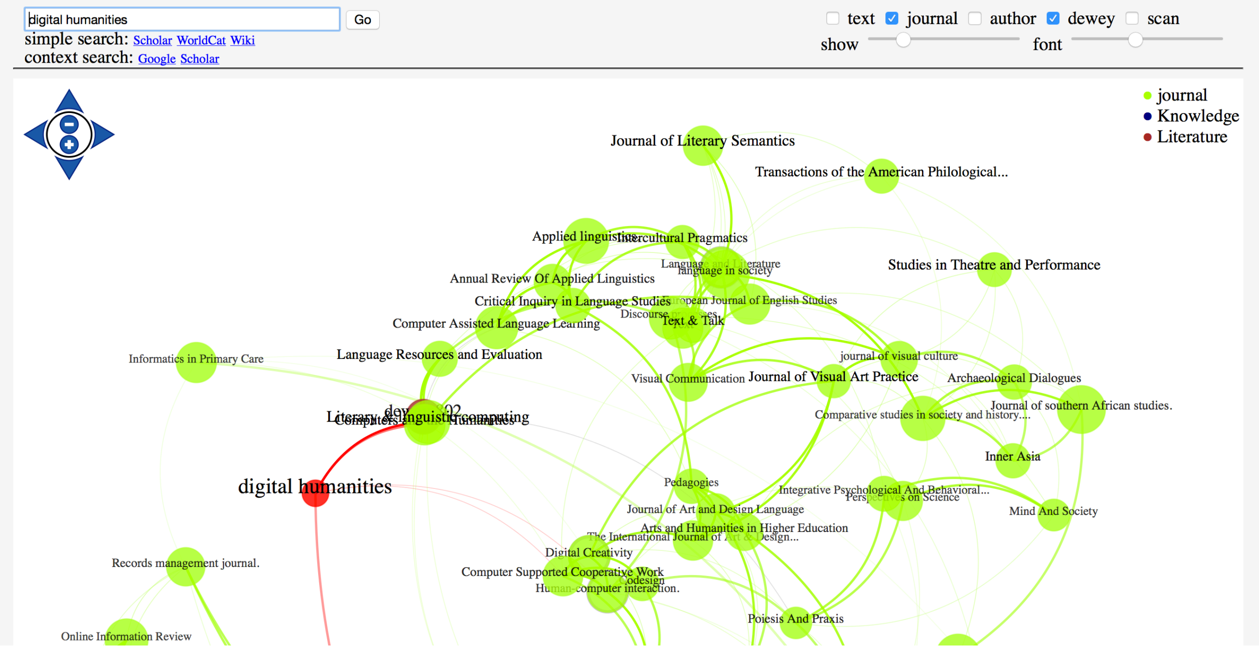
Relatedness or similarity is defined on the basis of a shared lexical profile in a high dimensional word space. To achieve a reactive, click-through interface the original sparse semantic matrix of a size of four million vectors of 1 million dimensions is pre-processed by means of Random Projection, a well-established technique for dimension reduction [8,9]. The resulting much smaller semantic matrix makes it possible to calculate relatedness or similarity between different types of entities on the fly.

The interface is lightweight. More recently, different connections to other open access information spaces, such as GoogleScholar, Wikipedia, WorldCat, have been implemented. Via hyperlinks the query (simple search) or a selection of prominent terms around the query (context search) can be automatically transferred to those other information services. This new feature allows for an iterative, back and forward search for contextualization both inside of ArticleFirst and beyond. (Fig 1)

In this presentation we give a demo of the tool. We also discuss different possible use applications:

- as entity disambiguation tool for information providers
- as contextualization engine for researchers
- as bibliometric tool to find related documents in information spaces without citations

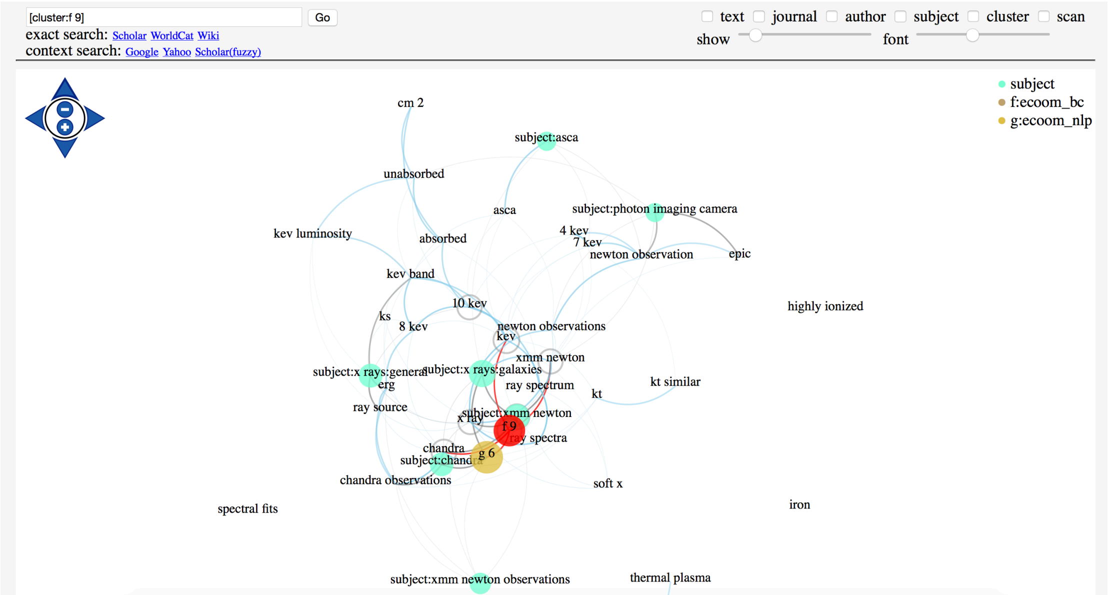
Concerning the latter, we report about an experiment with a specific dataset that contains records and citations – the so-called astrophysics dataset - used by a group of researchers to compare methods of topic delineation [6,7]. For this specific collaboration we created a separate ARIADNE instantiation, called Little Ariadne (http://thoth.pica.nl/astro/relate). Little Ariadne provides an interface to a set of about 100k documents. Since the ISSI2015 citations have been add as entities to the semantic matrix. The matrix further also includes cluster assignments produced by different research teams, and treats them as additional attributes to a document. (Fig 2) The OCLC team applied themselves standard clustering algorithms (K-Means and the Louvain community detection algorithm) on the semantic representation of articles built from the semantic matrix, both with and without citations. The clustering results are comparable with other results based on well-accepted citation-based approaches. Based on this experiment we discuss possibilities to use the semantic-matrix approach for bibliographic information spaces where citations are not available.


digital humanities

□ text ☑ journal □ author ☑ dewey □ scan

Go
simple search: Scholar WorldCat Wiki
context search: Google Scholar
show
font
● journal
● Knowledge
● Literature
Journal of Literary Semantics
Transactions of the American Philological...
Applied linguistics
Intercultural Pragmatics
Language and Literature
language in society
Studies in Theatre and Performance
Annual Review Of Applied Linguistics
Critical Inquiry in Language Studies
European Journal of English Studies
Discourse processes
Text & Talk
Computer Assisted Language Learning
Informatics in Primary Care
Language Resources and Evaluation
journal of visual culture
Visual Communication
Journal of Visual Art Practice
Archaeological Dialogues
Journal of southern African studies.
Comparative studies in society and history....
Literary & linguistic computing
dewey 402
Computers and the Humanities
Inner Asia
digital humanities
Pedagogies
Integrative Psychological And Behavioral...
Perspectives on Science
Journal of Art and Design Language
Mind And Society
Arts and Humanities in Higher Education
The International Journal of Art & Design...
Digital Creativity
Records management journal.
Computer Supported Cooperative Work
Codesign
Human-computer interaction.
Poiesis And Praxis
Online Information Review

Fig1: Snapshot from ARIADNE for the query “digital humanities”, link: http://thoth.pica.nl/relate?input=digital+humanities+&type=1&type=3&dots=&hubness=5&show=100&fsize=100



[cluster:f 9] Go

exact search: Scholar WorldCat Wiki

context search: Google Yahoo Scholar(fuzzy)

☐ text ☐ journal ☐ author ☐ subject ☐ cluster ☐ scan

Fig 2: Snapshot from Little Ariadne showing the context of one specific cluster solution in the astrophysics dataset [link: http://thoth.pica.nl/astro/relate?input=[cluster:f+9]]

References:

[1] Koopman, R., Wang, S., Scharnhorst, A., & Englebienne, G. (2015). Ariadne's Thread - Interactive Navigation in a World of Networked Information. Arxiv Digital Libraries (cs.DL); 1503.04358v1; CHI 2015. Digital Libraries. doi:10.1145/2702613.2732781 (preprint, accepted for proceedings)

[2] Scharnhorst, A. (2015). Walking through a library remotely. Why we need maps for collections and how KnoweScape can help us to make them? *Les cahiers du numérique*, *11*(1), 103-127. 10.3166/lcn.11.1.103-127 Preprint OA available at http://arxiv.org/abs/1503.06776

[3] Mutschke, Peter, Mayr, Philipp, and Andrea Scharnhorst (Eds.) (2014), KMIR 2014 - Knowledge Maps and Information Retrieval: Proceedings of the First Workshop on Knowledge Maps and Information Retrieval co-located with International Conference on Digital Libraries 2014 - ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014). vol. Vol-1311, CEUR-WS.org. http://ceur-ws.org/Vol-1311/

[4] Börner, K. (2011). Plug-and-play macroscopes. Communications of the ACM, 54(3), 60. doi:10.1145/1897852.1897871

[5] Whitelaw, M. (2015). Generous Interfaces for Digital Cultural Collections. DHQ : Digital Humanities Quarterly, 9(1), 1–16. Retrieved from http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html

[6] Koopman, R., Wang, S., & Scharnhorst, A. (2015). Contextualization of Topics - Browsing through Terms, Authors, Journals and Cluster Allocations. In A. A. Salah, Y. Tonta, A. A. A. Salah, C. Sugimoto, & U. Al (Eds.), Proceedings of ISSI 2015 Istanbul. 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29th June to 4th July 2015 (pp. 1042–1053). Istanbul: Boğaziçi University Printhouse.

[7] Wolfgang Glänzel, Jochen Gläser and Andrea Scharnhorst (2015) Same data - different results? The performative nature of algorithms for topic detection in

science. Special session organized at the ISSI2015, June 29 – July 4, Istanbul, http://www.issi2015.org/files/downloads/SS02.pdf

[8] Johnson, W., Lindenstrauss, J. (1984): Extensions of Lipschitz mappings into a Hilbert space. Contemporary Math. 26, 189–206

[9] Achlioptas, D. (2003): Database-friendly random projections: Johnson-Lindenstrauss with binary coins. Journal of Computer and System Sciences 66(4), 671–687.