Ariadne's Thread — Interactive Navigation in a World of Networked Information

Rob Koopman
OCLC
Schipholweg 99,
Leiden, The Netherlands
rob.koopman@oclc.org

Shenghui Wang
OCLC
Schipholweg 99
Leiden, The Netherlands
shenghui.wang@oclc.org

Andrea Scharnhorst
Royal Netherlands Academy of
Arts and Sciences
andrea.scharnhorst@dans.knaw.nl

Gwenn Englebienne
University of Amsterdam
Science Park 904
Amsterdam, The Netherlands
G.Englebienne@uva.nl

Abstract

This work-in-progress paper introduces an interface for
the interactive visual exploration of the context of queries
using the ArticleFirst database, a product of OCLC. We
describe a workflow which allows the user to browse live
entities associated with 65 million articles. In the on-line
interface, each query leads to a specific network
representation of the most prevailing entities: topics
(words), authors, journals and Dewey decimal classes
linked to the set of terms in the query. This network
represents the context of a query. Each of the network
nodes is clickable: by clicking through, a user traverses a
large space of articles along dimensions of authors,
journals, Dewey classes and words simultaneously. We
present different use cases of such an interface. This
paper provides a link between the quest for maps of
science and on-going debates in HCI about the use of
interactive information visualisation to empower users in
their search.

Author Keywords

knowledge maps; interfaces to digital libraries; science
maps; random projection; interactive visualisation

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User
Interfaces; H.3.3 [Information Storage and Retrieval].

Motivation

With digitization and the world wide web, any information seems to be at our fingertips. And yet it is very cumbersome to investigate a topic, to understand its context and history, and to find authoritative sources for it. It has been stated that information retrieval based on pure text-statistical methods has reached a certain limit [13]. For current information retrieval, it is difficult to answer questions such as:

- What are the different aspects of this topic?
- Are there related aspects missing in my search terms?
- Who are the most prominent authors about this topic?
- Which journals publish most about this topic?
- How have others — e.g. librarians — described and classified this topic?

For scientific articles, the entities involved in these questions—authors, journals, subjects—are essentially interlinked and can be accessed via bibliographic databases such as the *Web of Science*, *Scopus*, *SpringerLink*, *ArticleFirst* or *Microsoft Academic Search*. Back in the history of information science, the *Web of Science* has fostered the emergence of a whole new field: scientometrics. With the introduction of the principle of citation indexing [7] and access to networks of scientific papers, there have been dreams to visualize this fabric of science [16]. The making of science maps has culminated in the exhibition *Places&Spaces* [2]. It also resulted in discussions on how to implement them into digital libraries [3]. To visualize the context of an scholarly argument has further inspired the maker of interactive interfaces [6, 12] Still, bibliographic databases are dominated by a single search term window and ranked lists of results. Maps are only occasionally implemented.

Examples are the AuthorMapper[1] for *SpringerLink*, the co-author graph for *Microsoft Academic Search*,[2] or *HistCite* for *Web of Science*.[3] Most of the time, visual exploration is possible in stand-alone tools such as the *Sci²* tool[4] or *CitNetExplorer*.[5] Visual navigation has been also implemented for specific projects such as the *MESUR* project for clickstream data[6] or the *GenderBrowser*[7] which operates on a dump of JSTOR. At the same time, information and computer scientists (HCI, InfVis, Visual Analytics) call for further experiments to navigate visually and interactively through large information spaces [5, 14, 15]. This paper contribute to further connect different discourses about analysing, visualizing and navigating large bodies of scholarly knowledge.

Obviously, it would be valuable if in the search for topics we could seamlessly travel between relevant subjects, authors and journals. Imagine, starting with a search term, we could find other subjects related to it; after choosing a subject, the most relevant authors or journals could be presented; for an author, we would be able to find subjects he/she publishes about and which journals are most relevant to those. In such an envisioned scenario the user can shift his/her interests as the exploration goes on. *Serendipity* is not only expected, but full-heartily embraced. In this work-in-progress paper, we present an interface which enables such journeys. To be able to *scroll*
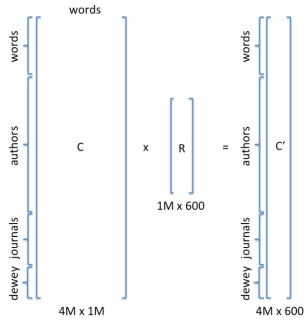
---

through information spaces has motivated many explorations of interactive interfaces [6]. This paper joins into those efforts but takes a slightly different approach, focussing on similarities in large vector spaces composed by terms, author, journal and Dewey classes.

Dataset   We develop a visual interactive interface to browse through contexts based on information from the article database *ArticleFirst* of OCLC. ArticleFirst contains more than 65 million article records from more than 30 thousand journals with which more than 3 million authors are associated. We treat author names, journal ISSNs, Dewey classes (assigned to journals) and topical terms (words) in the title and abstract of articles as entities. The method allows us to investigate the relations between roughly 1 million topical terms, 3 million authors, 30 thousand journals and 738 Dewey decimal classes.

Method

One issue for web-based interactive interfaces is responsiveness. This requires scalability of the underlying algorithms. For our interface, we solved this problem by combining an **off-line** preparation phase with an **on-line** process. Off-line, we build the semantic representation of each entity. Hereby, we use Random Projection to reduce dimensionality (Figure 1). In the **on-line** interface terms from a query are matched to entities in this reduced semantic matrix. The number of hits is further reduced to render a network layout easy to overview and navigate.

Off-line: Low-dimensional semantic representation using random projection   1).

The relatedness of entities is calculated based on the *context* they share, instead of direct co-occurrences in the data. The context is computed as follows. We use a

database of documents, in this case journal articles, from which, in this first implementation, we focus on titles and abstracts. After removing stop words, we select approximately one million most frequent terms (single words or two-word phrases) as topical terms in this corpus. These terms are used both to represent context (columns of C, Figure 1) and as searchable topical terms (first rows of C), although this is not a requirement. Each entity — in our case, topical term, author, journal, and Dewey decimal class, as shown in matrix C of Figure 1 — is therefore represented by the vector of co-occurrence frequencies between the entity and these terms, accumulated over articles. Authors, for instance, are represented in terms of the vocabulary they use across all their articles. The row vectors define the context for each entity. Cosine similarity calculated from these row vectors defines the semantic similarity between entities. As a consequence, it is possible to compute similarities between entities of different types, such as between a journal and an author, or between a topic and a Dewey class.

The high dimensionality of the vector representation is necessary to capture the long tail of word frequencies, but makes direct computation of similarities intractable. To make the algorithm scale, we reduce the dimensionality of the column vectors from 1M to 600 dimensions using Random Projection [1, 9]. The choice of keeping 600 dimensions was decided empirically for our dataset.

As a result of this approach, computing the low-dimensional representation and the following on-line matching of the search terms becomes very fast. Although this may seem to be a rather drastic approach, our experiments show that meaningful relationships are preserved. It is as if we would ignore some redundancy in the set of terms. R is a fixed matrix that is not related to
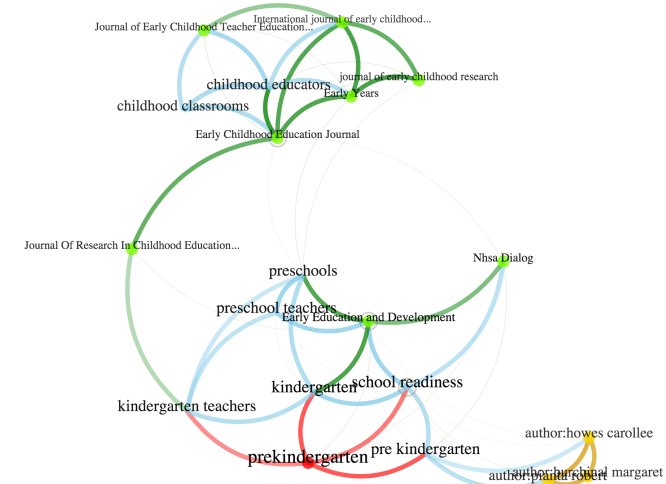

Figure 1: Dimensionality reduction using random projection, where R is a random matrix containing $-1, 1$.

C (contrary to other dimensionality reduction methods), so that creating it can be considered an unit-cost operation. Moreover, since each entry of C' is a just weighted sum of entries in C, instead of updating the term frequencies in C, we can adjust the relevant entries of C' directly without ever actually storing C. In the end, the manageable matrix of C' is stored, with each row representing an entity (either a topical term, an author, a journal or a Dewey decimal class in our case). Cosine similarity (implemented in the interface) or any other type of similarity measure can be computed fast between any pair of entities, which is essential for the interactive exploration. We call C' the **semantic matrix**.

On-line: Interactive exploration of networked entities
In the interactive web application — accessible at http://thoth.pica.nl/relate — any exploration starts with a query. The user can copy chunks of text into the search window, from single terms to paragraphs. The process looks up each of the entered terms in the semantic matrix C'. It retrieves the corresponding vector representation of the involved entities. Entities (word, author name or ISSN) which do not have an entry in the semantic matrix are ignored. Cosine similarities are then calculated between this averaged representation of the query and every entity in the matrix. The 500 most related entities are kept.

The next step is to filter out non-specific entities which are close to very many other entities. In order to do that, we select an entity, compute the Mahalanobis distance [11] between the selected entity and the similarity distribution of its neighbours, and only keep those neighbours that have the smallest Mahalanobis distance. The remaining entities are then ready to be projected to a two-dimensional visualisation using multi-dimensional

scaling. The result is a visualization of a network of related entities (Figure 2), of which each node is clickable. Once clicked, a new round of selection starts.

Figure 2: Let's start with *prekindergarten* (in red). The 20 most related entities including topics, authors (in yellow) and journals (in green) are presented.

User interface and first observations

We purposefully kept the interface plain. Our primary question is: Do the algorithm and resulting network of related entities present a meaningful context to our search quest? Does the network invite to further explore contexts? To enable cross-checking, the upper menu line contains buttons to look up the query in Wikipedia, WorldCat or ScholarGoogle.

We tested the interface for a couple of different search tasks.[8] In the *search for an author*, one can detect different spellings in which an author's name occurs in the dataset.[9] This could be a first visual screening test to identify issues in author disambiguation or other issues in metadata curation. We also observe that the specificity of a query term influences the search result. When *searching for a topic* such as "machine learning," the retrieved context contains related methods such as *neural network*, *svm*, *bayesian classifier* and *decision tree*, as well as related aspects such as *accuracy*, *feature selection*, *overfitting*, and *generalization performance*. The result is not a systematic classification but a very good reflection of the context of this topic in scholarly practice. Sometimes, unexpected terms catch our attention and invite to further investigation. They could also potentially lead to additions to existing classification schemes. More general or ambiguous terms are also interesting. Searching for "child care" for instance[10] reveals many aspects one would expect such as *family income*, *parenting skills*, *child health*, *disadvantaged children*. But we also find *immigrant families*, a term which invites to further exploration.

Because similarities are calculated between all different types of entities simultaneously, it depends very much on the query entry which types show-up in the resulting network. The minimalistic interface allows the user to restrict the display to one type, and this way similarities in only the journal or the author space become visible. The seamless transfer between them allows a first rough

delineation of a query in a topical space encompassing terms, authors, journals and Dewey decimal classes.

Future work and discussion

The interface and its underlying algorithmic processes realise principles of interactive visual browsing and navigating *in vivo* in a large bibliographic database. It builds on past and present attempts to integrate visual elements to on-line searching and browsing [4, 8, 10, 15]. Pursuing this work we will address the following tasks: (1) We will compare the applied algorithm with other algorithms used in bibliometrics for the delineation of topics and fields. (2) We will pursue more systematic user studies to investigate the role of context visualisation for information foraging. (3) Ultimately, we would like to enable the user to also retrieve related articles. The latter task entails a shift from an explorative interface to an implementation into the service of ArticleFirst.

There are also a couple of extensions of the method and its application we think are interesting. For example, could we incorporate temporal analysis — a timeline button? How can the visualisation be improved — algorithm-wise and design-wise? Could results be exported for secondary analysis? Could we apply sentiment analysis to add connotation to the links? Could elements of recommendations and knowledge discovery be added to the interface and if so in which way?

For the time being, the association we got in our experiments with the interface was to exploring a library through different kinds of catalogues: author catalogue and systematic catalogues. Sometimes, the triangulation between the landing points of search arrows into the space of author, journal, terms and subject headings would lead to an appropriate and comprehensive overview; sometimes

---

[8]A first exploration was done by the authors and members of the COST Action TD1210.

[9]See the different name variations of *Rienk van Grondel* at http://thoth.pica.nl/relate?input=%5Bauthor%3Avan+grondelle+r%5D

[10]See http://thoth.pica.nl/relate?input=child+care

it would lead to surprising associations revealing a meaning after closer inspection; and sometimes it would leave us with a labyrinth requiring a thread from Ariadne. It is possible that this is the best which we can expect from a navigable interface into large search spaces: that it complements other search options and enhances those by presenting overview and contextualisation in a tentative way, inviting further exploration, navigation, and eventually close inspection.

References

[1] Achlioptas, D. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. of Comput. and Syst. Sci. 66*, 4 (2003), 671–687.

[2] Börner, K. *Atlas of science: Visualizing what we know*. The MIT Press, Cambridge, MA., 2010.

[3] Börner, K., and Chen, C. Visual interfaces to digital libraries: Motivation, utilization, and socio-technical challenges. In *Visual Interfaces to Digital Libraries [JCDL 2002 Workshop]*, Springer-Verlag (London, UK, 2002), 1–12.

[4] Brooks et al. Hoptrees: Branching history navigation for hierarchies. In *Human-Computer Interaction INTERACT 2013*, Kotzé, Paula, Marsden, Gary, Lindgaard, Gitte, Wesson, Janet, and M. Winckler, Eds., vol. 8119 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, 316–333.

[5] Dörk, M., Williamson, C., and Carpendale, S. Navigating tomorrow's web: From searching and browsing to visual exploration. *ACM Trans. Web 6*, 3 (Oct. 2012), 13:1–13:28.

[6] Dörk, et al. Pivotpaths: Strolling through faceted information spaces. *Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2710–2719.

[7] Garfield, E. Citation indexes to science: A new dimension in documentation through association of ideas. *Science 122*, 3159 (1955), 108–111. doi: 10.1126/science.122.3159.108.

[8] Hall et al. The paths system for exploring digital cultural heritage. In *Proc. of the Digital Humanities Congress 2012*, M. P. Clare Mills and E. Ward, Eds., Studies in the Digital Humanities. Sheffield (2012).

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

[10] Kaizer, J., and Hodge, A. Aquabrowser library: Search, discover, refine. *Library Hi Tech News 22*, 10 (2005), 9–12.

[11] Mahalanobis, P. C. On the generalised distance in statistics. *Proceedings National Institute of Science, India 2*, 1 (1936), 49–55.

[12] Matejka, J., Grossman, T., and Fitzmaurice, G. Citeology: Visualizing paper genealogy. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (2012), 181–190.

[13] Mutschke, P., and Mayr, P. Science models for search: a study on combining scholarly information retrieval and scientometrics. *Scientometrics* (2014), 1–23.

[14] Salah et al. Significance of visual interfaces in institutional and user-generated databases with category structures. In *Proc. of the 2nd Int. ACM Workshop on Personalized Access to Cultural Heritage*, ACM (New York, USA, 2012), 7–10.

[15] Shneiderman et al. Visualizing digital library search results with categorical and hierarchical axes. In *Proc. 5th ACM Conference on Digital Libraries*, ACM (New York, NY, USA, 2000), 57–66.

[16] Waltman, L., van Eck, N. J., and Noyons, E. C. M.
A unified approach to mapping and clustering of
bibliometric networks. *J. of Informetrics 4*, 4 (2010),
629–635.