Head Start: Improving Academic Literature Search with Overview Visualizations based on Readership Statistics

Peter Kraker
Know-Center
Inffeldgasse 13
8010 Graz, Austria
pkraker@know-center.at

Christoph Trattner
Know-Center
Inffeldgasse 13
8010 Graz, Austria
ctrattner@know-center.at

Kris Jack
Mendeley
144a Clerkenwell Road
EC2 RF London, UK
kris.jack@mendeley.com

Stefanie Lindstaedt
Know-Center
Inffeldgasse 13
8010 Graz, Austria
slind@know-center.at

Christian Schlgl
University of Graz
Universittsstrasse 15
8010 Graz, Austria
christian.schloegl@uni-graz.at

WebSci13, May 24, 2013, Paris, France.
ACM 978-1-4503-1889-1

Abstract

At the beginning of a scientific study, it is usually quite hard to get an overview of a research field. We aim to address this problem of classic literature search using web data. In this extended abstract, we present work-in-progress on an interactive visualization of research fields based on readership statistics from the social reference management system Mendeley. To that end, we use library co-occurrences as a measure of subject similarity. In a first evaluation, we find that the visualization covers current research areas within educational technology but presents a view that is biased by the characteristics of readers. With our presentation, we hope to elicit feedback from the Websci'13 audience on (1) the usefulness of the prototype, and (2) how to overcome the aforementioned biases using collaborative construction techniques.

Author Keywords

literature search; co-readership analysis; library statistics; interactive map; educational technology; alternative metrics

ACM Classification Keywords

D.2.8 [Software Engineering]: Metrics; H.2.8 [Database Applications]: Scientific databases; H.5.4 [Hypertext/Hypermedia]: Navigation

Mendeley

Mendeley provides users with software tools that support them in conducting research [3]. One of the most popular of these tools is Mendeley Desktop, a cross-platform, freely downloadable PDF and reference management application. It helps users to organize their personal research libraries by storing them in relevant folders and applying tags to them for later retrieval. The articles, provided by users around the world, are then crowd-sourced into a single collection called the Mendeley research catalogue [2]. At the time of writing, this catalog contains more than 80 million unique articles, crowd-sourced from over 2 million users. The user's publications are also augmented by readership counts, allowing them to track the popularity of their individual papers within the Mendeley community. These readership counts indicate how many Mendeley users have added the author's article to their personal research library.

Introduction

The web increasingly shapes the way researchers connect and collaborate. Social reference management systems in particular, enable researchers to store their references in an online library, and to share and discuss them with other users. Apart from that, web activity also serves as a data provider for the analysis of science. Usage data allows for assessing the quality of a publication long before the first citations arrive [10]. Thanks to the web, it is now possible to view scholarly communication through the eyes of the reader [11].

Next to quality control, information overload is an important issue in science. Science has been growing exponentially since its inception 400 years ago [9], and it still continues to grow to this day [6]. The number of researchers in China, for example, doubles every 8 years [8]. Therefore, it is usually quite hard to get an overview of a research field at the beginning of a scientific study. One needs to work through long lists of search results and their references to build a mental model of the field. Recent publications are often buried far down the list, because they have not received many citations yet.

In our work, we want to address these problems of classic literature search using readership statistics from social reference management systems. We have developed an interactive visualization which gives an overview of the most important areas in a field, and shows popular publications related to each of these areas. In contrast to earlier visualizations based on co-citations, the map is based on co-occurrences of papers in user libraries. The visualization is intended to assist students and researchers in any field. As a use case, we have chosen to explore the field of educational technology. The first prototype is available online[1]. It was created with D3.js[2] and can be run in Chrome and Firefox without additional plugins.

Data & Method

Our visualization is based on data from the popular reference manager Mendeley[3]. In line with earlier findings such as [7], we use a structural measure rather than a content-based measure to determine subject similarity. Research terminology has proven to be too fluent to provide consistent results over time. In contrast to previous visualizations though, we employ readership statistics instead of citations to calculate subject similarities. Readership statistics have a distinct advantage over citations: they are available much earlier, shortly after the paper has been published. It has been shown that readership statistics provide a good coverage of top publications [1] and that they provide a reasonable correlation with the impact factor [5]. Furthermore, there are first empirical results indicating that library co-occurrences serve as a measure of subject similarity [4].

The visualization was created using multi-dimensional scaling (MDS) and hierarchical agglomerative clustering (HAC) on library co-occurrences. The basic method is described in full detail in [5]. We used HAC with Ward's method to establish the research areas. The number of clusters was determined by the elbow method. We then employed non-metric MDS to place the articles on the map. To unclutter the map, a force-directed layout with a collision detection algorithm was used. The areas were named by sending the paper titles to Zemanta[4] and Open Calais[5]. Both services crawl the semantic web and return a number of concepts that describe the content. The more words a concept has,

[1] http://labs.mendeley.com/headstart

[2] http://d3js.org

[3] http://mendeley.com

[4] http://zemanta.com

[5] http://opencalais.com

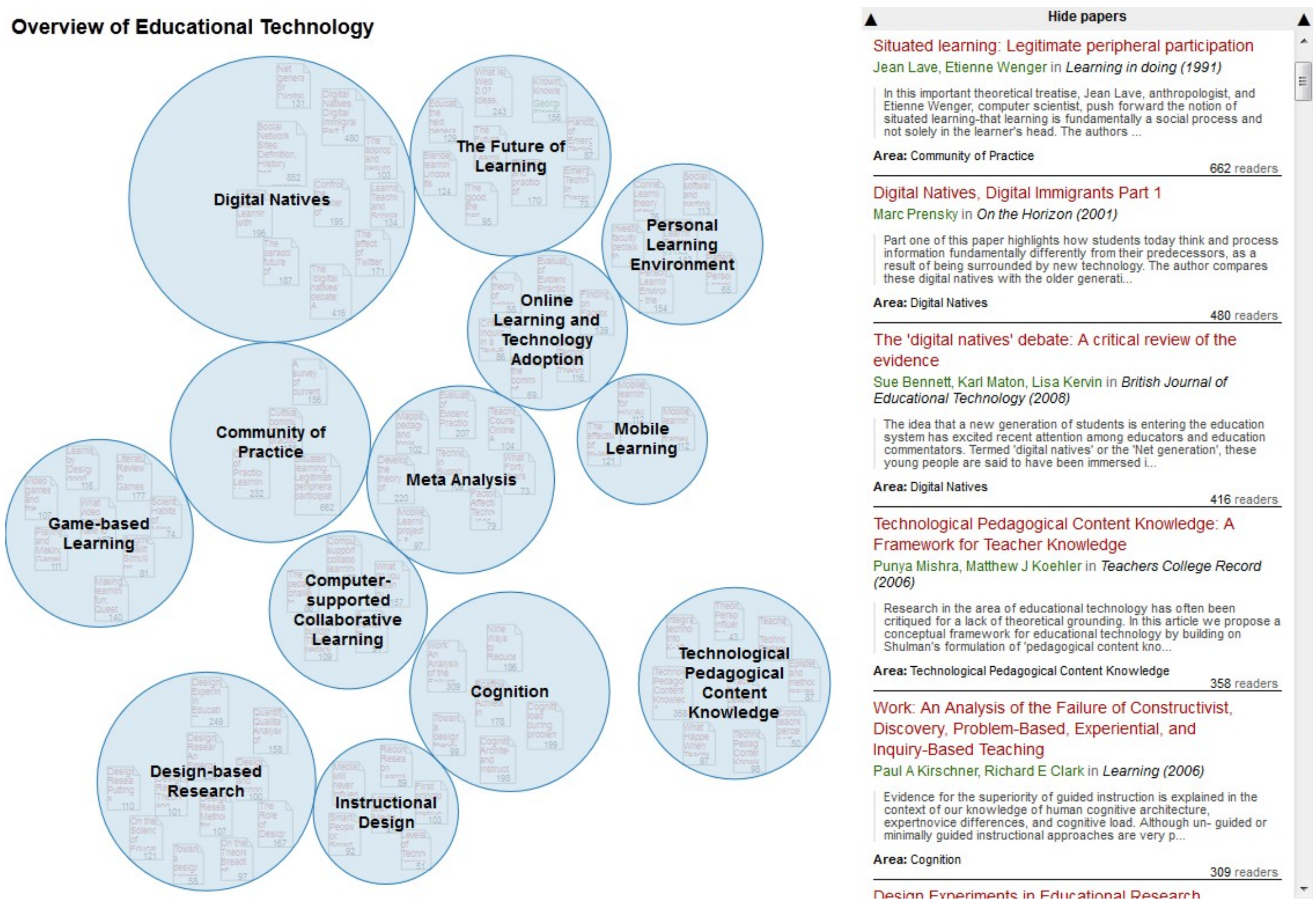

Figure 1: Overview of Educational Technology

and the more often it occurs within titles and abstracts of an area, the more likely it is to be the name of the area. Manual interventions were only necessary to adjust the naming of some areas and to select the number of publications included in the map. As mentioned above, we took the field of educational technology as a use case. A publication has to be read at least 16 times by Mendeley users from the field of educational technology, leading to a total of 91 papers. These papers appeared in 7,414 user libraries with a total of 19,402 co-occurrences.

Results

The overview visualization can be seen in Figure 1. Head Start follows the "Overview first, zoom and filter, then details-on-demand" concept popularized by Shneiderman [12] to support exploration. Initially, the main areas in the field are shown, represented by blue bubbles (see Figure 1). The size of the bubbles signifies the number of readers of publications in that area. The closer two areas are in the visualization, the closer they are subject-wise. Once you click on a bubble, you are presented with popular papers in that area. The dropdown on the right displays the same data in list form (see Figure 2).

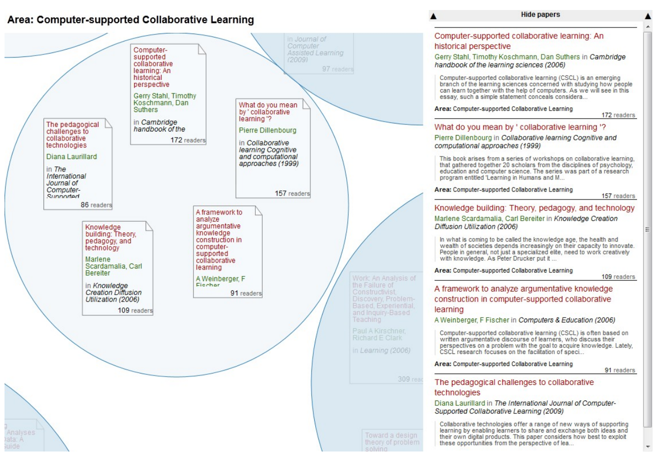

Figure 2: Zooming into the area "Technological Pedagogical Content Knowledge"

By clicking on one of the papers, one can access all metadata for that paper (see Figure 3). If a preview is available, the user can retrieve it by clicking on the thumbnail in the metadata panel. This allows the user to learn about the field on many levels of abstraction. The inclusion of full texts where possible ensures that the user can do most of the exploration in a single user interface.

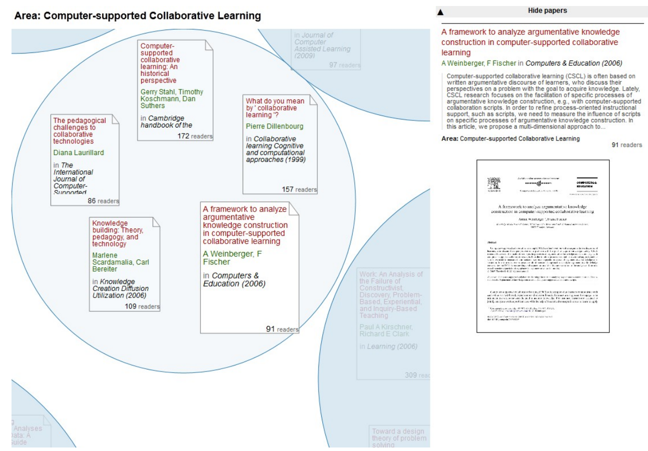

Figure 3: Showing the metadata of a paper

The map contains thirteen research areas that cover the main aspects of educational technology. The visualization can be roughly divided into three parts. The bottom third represents the psychological, pedagogical and methodical basis of the field. In the top third, social and technological developments are displayed. In between, popular learning methods and technologies are being discussed. In the center, there is an area of meta reviews. The publications of this area are most often read together with publications from other areas.

Evaluation

In a first evaluation, the visualization has been compared to 10 other analyses of educational technology literature. Among these analyses were four quantitative analyses based on citations and word frequencies, four qualitative content analyses, one study of larger streams in educational technology, and one Delphi study. The details can be found in the appendix.

In comparison to citation analyses, the proposed visualization is more diverse. It presents more research areas, and these areas are by and large more specific than those in citation analyses. Interestingly, most research areas from qualitative analyses are covered as well. While some of these topics might not be represented by their own cluster, they are often represented as part of another cluster. Furthermore, the visualization is a very recent representation of the field: 80% of the publications included are from the last 10 years.

Nevertheless, the co-readership analysis also has certain limitations. Being based on the readers, their characteristics may introduce biases to the visualization. Educational technology is an interdisciplinary field, but in Mendeley's discipline taxonomy it appears as a sub-discipline of education. Therefore, the map represents an education-dominated view. Areas that are mostly influenced by computer science such as adaptive hypermedia are missing from the visualization.

Conclusions & Future Work

While readership statistics have previously been shown to be a good indicator of research impact, our visualization now shows that they can also be used to map scientific fields. In comparison with quantitative literature analyses, it becomes apparent that our visualization covers many areas in the field of educational technology.

For a further evaluation, we are currently conducting expert interviews of the visualization with researchers from educational technology. We expect to gain deeper insights about representativeness and recency of the map. Furthermore, we are looking to gain insights into size, distribution and connections of the research. Last but not least, we want to know more about the usability of the map for research fields, and its extensibility. To overcome these biases, we want to empower users to adapt and extend the maps. Thus, researchers would be able to use the visualization as an overview of their personal library, and to collaboratively build a view of a field.

This is a discussion that we want to continue at WebSci'13 to get opinions of experts from other research fields. We plan to include maps for the research areas of Artificial Intelligence, Human-Computer Interaction, and Information Science until then. We hope to elicit feedback on (1) the usefulness of the prototype, and (2) how to overcome the aforementioned biases using collaborative construction techniques.

Acknowledgements

The authors would like to thank Andi Rutherford for his comments on the draft. The research presented in this work is in part funded by the European Commission as part of the FP7 Marie Curie IAPP project TEAM (grant no. 251514).The Know-Center is funded within the Austrian COMET program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth, and the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

[1] J. Bar-Ilan, S. Haustein, I. Peters, J. Priem, H. Shema, and J. Terliesner. Beyond citations: Scholars' visibility on the social Web. In *17th International Conference on*

Science and Technology Indicators, pages 1–14, 2012.

[2] J. A. Hammerton, M. Granitzer, D. Harvey, M. Hristakeva, and K. Jack. On generating large-scale ground truth datasets for the deduplication of bibliographic records. In *International Conference on Web Intelligence, Mining and Semantics 2012*, number January, 2012.

[3] V. Henning and J. Reichelt. Mendeley - A Last.fm For Research? *IEEE Fourth International Conference on eScience*, pages 327–328, 2008.

[4] J. Jiang, D. He, and C. Ni. Social reference: aggregating online usage of scientific literature in CiteULike for clustering academic resources. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, number 1, pages 401–402. ACM, 2011.

[5] P. Kraker, C. Körner, K. Jack, and M. Granitzer. Harnessing User Library Statistics for Research Evaluation and Knowledge Domain Visualization. In *Proceedings of the 21st International Conference Companion on World Wide Web*, pages 1017–1024, Lyon, 2012. ACM.

[6] P. O. Larsen and M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3):575–603, Sept. 2010.

[7] L. Leydesdorff. Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science*, 48(5):418–427, May 1997.

[8] National Science Board. *Science and Engineering Labor Force*, volume 22 Suppl 1. National Science Foundation, Arlington, Nov. 2010.

[9] D. J. D. S. Price. *Little science, big science*. Columbia Univ. Press, 1963.

[10] J. Priem and B. M. B. Hemminger. Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday*, 15(7), 2010.

[11] I. Rowlands and D. Nicholas. The missing link: journal usage metrics. *Aslib Proceedings*, 59(3):222–228, 2007.

[12] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343, 1996.

Appendix

Cho et al. (2012): "The landscape of educational technology viewed from the ETR&D journal"
Chen and Lien (2011): "Using author co-citation analysis to examine the intellectual structure of e-learning: A MIS perspective"
Wild et al. (2010): "Shifting Interests: Changes in the Lexical Semantics of ED-MEDIA"
Fisichella et al. (2009): "Who are you working with? - Visualizing TEL Research Communities"
Klein (1997) : "ETR&D-Development: An Analysis of Content and Survey of Future Direction"
Masood (2004): "A ten year analysis: Trends in traditional educational technology literature"
Maurer and Khan (2010): "Research trends in the field of e-learning from 2003 to 2008: A scientometric and content analysis for selected journals and conferences using visualization"
Hsu et al. (2012) : "Research Trends in Technology-based Learning from 2000 to 2009: A content Analysis of Publications in Selected Journals"
Spada et al. (2012): "Final report on the STELLAR Delphi study"
Czerniewicz (2010): "Educational technology - mapping the terrain with Bernstein as cartographer"