The Quest for Keeping an Overview: Knowledge Domain Visualizations based on Co-Readership Patterns

Article · January 2015

CITATIONS
0

READS
49

4 authors, including:

Christian Schloegl
Karl-Franzens-Universität Graz
49 PUBLICATIONS **455** CITATIONS
SEE PROFILE

Kris Jack
University of Dundee
24 PUBLICATIONS **141** CITATIONS
SEE PROFILE

Stefanie Lindstaedt
Graz University of Technology
141 PUBLICATIONS **1,164** CITATIONS
SEE PROFILE

The Quest for Keeping an Overview: Knowledge Domain Visualizations based on Co-Readership Patterns

Peter Kraker*, Christian Schlögl†, Kris Jack‡ and Stefanie Lindstaedt*
*Know-Center
8010 Graz, Austria
Email: {pkraker,slind}@know-center.at
†University of Graz
8010 Graz, Austria
Email: christian.schloegl@uni-graz.at
‡Mendeley
EC2 RF London, UK
E-Mail: kris.jack@mendeley.com

***Abstract*—Given the enormous amount of scientific knowledge that is produced each and every day, the need for better ways of gaining – and keeping – an overview of research fields is becoming more and more apparent. In a recent paper published in the Journal of Informetrics [1], we analyze the adequacy and applicability of readership statistics recorded in social reference management systems for creating such overviews. First, we investigated the distribution of subject areas in user libraries of educational technology researchers on Mendeley. The results show that around 69% of the publications in an average user library can be attributed to a single subject area. Then, we used co-readership patterns to map the field of educational technology. The resulting knowledge domain visualization, based on the most read publications in this field on Mendeley, reveals 13 topic areas of educational technology research. The visualization is a recent representation of the field: 80% of the publications included were published within ten years of data collection. The characteristics of the readers, however, introduce certain biases to the visualization. Knowledge domain visualizations based on readership statistics are therefore multifaceted and timely, but it is important that the characteristics of the underlying sample are made transparent.**

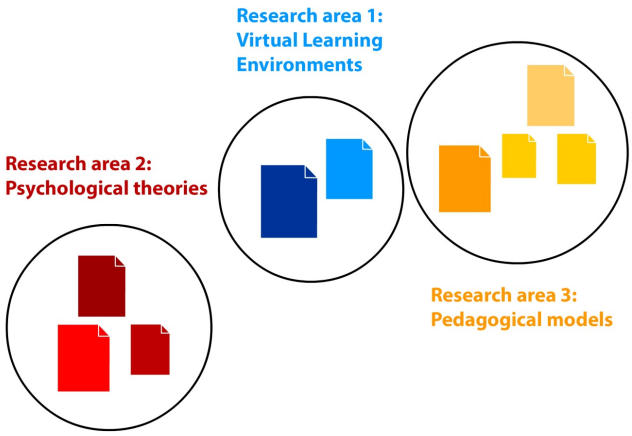

Fig. 1. Examplary knowledge domain visualization of the field of "educational technology", illustrating the main concepts of knowledge domain visualizations.

I. Introduction

Given the enormous amount of scientific knowledge that is produced each and every day, the need for better ways of gaining – and keeping – an overview is becoming more and more apparent. Knowledge domain visualizations are a means of getting such an overview (see Figure 1 for an examplary visualization). They show the main areas in a field, and assign relevant articles to these main areas. An additional characteristic of knowledge domain visualizations is that areas of a similar subject are positioned closer to each other than areas of an unrelated subject. Furthermore, knowledge domain visualizations may display relevance and other properties of individual areas or papers using size, color and placement. Hence, an interested researcher can see the intellectual structure of a field at a glance without performing countless searches with all different sorts of queries.

Even though the idea of knowledge domain visualizations has been around for quite some time, and despite their obvious usefulness, they are not yet widely available. Part of the reason may be that in the past, the data needed to construct these visualizations (citations) was only available from a few rather expensive choices. Part of the reason may be that there has been an emphasis on all-encompassing overviews. While they provide valuable insights into the structure of science as a whole, they are usually not interactive and provide little value in day-to-day work where you want to be able to zoom into specific publications. There are several applications that can be used to create one's own overview, but they can usually only be operated by users that are information visualization specialists.

In a recent paper published in the Journal of Informetrics [1], we describe an interactive visualization that can be used by anyone. The visualization is based on a novel data source – the online reference management software Mende-

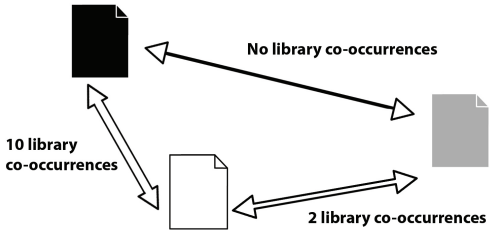

Fig. 2. Relationships between documents in a field based on co-readership. Co-occurrence in user libraries is employed as a measure of subject similarity.

ley[1]. Mendeley enables users to store their reference papers in a personal library and share them with other people. The number of times a paper has been added to user libraries is commonly referred to as the number of readers, or in short readership. The papers for the visualization were selected from Mendeley's research catalog which is crowd-sourced from over 2.5 million users from around the world and offers structured access to more than a 100 million papers.

One of the most important steps when creating a knowledge domain visualization is to decide which measure defines the similarity between two articles. The measure is used to determine where an article gets placed on the map and how it is related to other articles. Again, we used Mendeley data, specifically co-readership information, to tackle this issue. A co-readership relation between two documents is established when at least one user has added the two documents to his or her user library. When Alice adds Paper 1 and Paper 2 to her user library, the co-readership of these two documents is 1. When Bill adds the same two papers, the co-readership count goes up to 2, and so on. Our assumption was now that the higher the co-readership of two documents, the more likely they are of the same or a similar subject. It's not unlike two books that are often rented together from a library – there is a good chance that they address related topics.

The topical relationship established by co-readership can then be exploited for visualizations by clustering those papers that have high co-readership numbers (see Figure 2). To the best of our knowledge, this measure had not been exploited before for knowledge domain visualization.

In our study, we first investigated the distribution of subject areas in user libraries in order to test our assumption that co-readership implies subject similarity. Then, we employed co-readership patterns to create a knowledge domain visualization. As a use case, we chose the field of educational technology.

II. Distribution of Subject Areas in User Libraries

Subject homogeneity, meaning that a significant share of papers in a collection can be attributed to a single subject, is a necessary precondition that the results of co-readership analysis are valid; otherwise the assumption that co-occurrence of articles in user libraries implies subject similarity cannot be upheld. Therefore, we analyzed the subject distribution of articles included in Mendeley user libraries and compared it to the subject area distribution of reference lists of articles

[1]http://mendeley.com

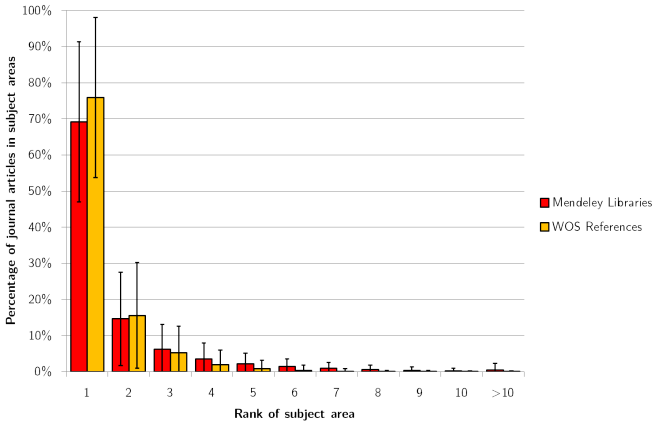

Fig. 3. Subject area frequency distribution of articles in user libraries from educational technology (n=72,721 journal articles in 1,107 user libraries) and cited references in WoS articles (n=13,841 cited references in 1,394 documents). Ranks 11-25 (Mendeley) and 11-12 (WOS articles) were summed up.

in Web of Science. The basis of this analysis is the user profiles and user libraries data set of researchers in educational technology (n=2,154 users). The categorization of users into sub-disciplines is determined by self-ascription of users on Mendeley.

In a first step, we analyzed the distribution of journal articles in user libraries. We used SCImago, which is a bibliometric service based on the bibliographic database Scopus, as an external validation source. SCImago categorizes each journal into one of 28 subject areas. The documents from the field of educational technology were matched to these subject areas through the journals they appear in. After this procedure, 1,107 user libraries, which contained at least one article in a journal that is indexed by SCImago, were left. A Mendeley user library in educational technology has on average 155.7 documents (SD=460, Median=17); slightly more than a third (56.7) of these documents are on average journal articles that appeared in journals indexed by SCImago (SD=202.2, Median=15).

We also created a data set of cited references from Web of Science. We searched for articles and reviews with the topic "educational technology" in the WOS Core Collection. This resulted in 1,394 documents. We retrieved the cited references for these documents; each document has on average 29.2 cited references (SD=23.8, Median=25). We then applied the procedure outlined above to match references to subject areas via their journals. This resulted in 1221 reference lists which contained at least one document that is indexed by SCImago; 38% of these (11.1 documents) are on average journal articles that appeared in journals indexed by SCImago (SD=12.7, Median=7).

Finally, we calculated the distribution of SCImago categories for each Mendeley user library from educational technology and each cited reference list for the article set retrieved from Web of Science. Afterwards, we ranked the results by subject area. For each library, the percentage of articles that are categorized into a common subject area was calculated. Then, the areas were ranked according to their frequency. The

average subject area distribution for all educational technology user libraries can be seen in Figure 3.

These results show that, as was expected, cited references in journal articles are very homogeneous with regards to their subject area distribution. Mendeley user libraries are less homogeneous, and they spread out over more subject areas. The top subject area, however, still accounts for 69.2% of articles in an average user libraries (compared to 76.0% in cited references), even though the number of journal articles in an average user library (56.7) is 5 times higher than the number of cited references in an average journal article (11.2). Therefore, although co-readership probably offers a weaker indication of subject similarity than co-citation, it can still be expected to serve as a useful indication of subject similarity. This is in line with an earlier study by [2] which finds that clusters based on the occurrence and co-occurrence of articles in user libraries of CiteULike are as effective as citation-based clusters.

III. Visualization of Co-Readership Patterns

A. Data

The following data sets have been sourced from Mendeley in 2012 and 2013 and represent data for the sub-discipline educational technology that had been accumulated in the system up to that point:

- User profiles and user libraries: all user profiles and their accompanying user libraries in the sub-discipline of educational technology (n=2,154 users)
- Documents: metadata of all documents in the field of educational technology (n=144,500 documents)
- Co-occurrences: co-occurrences of these documents in all Mendeley user libraries (n=56,049,431 co-occurrences).

B. Method

For the visualization of co-readership patterns, we followed the knowledge domain visualization process as proposed by [3]. It consists of four steps: (1) selection of an appropriate data source, (2) determination of the unit of analysis, (3) analysis of the data using dimensionality reduction techniques, and (4) visualization and interaction design. Each of these steps is detailed below. The whole procedure can be seen in Figure 4.

The documents included in the analysis were taken from the Mendeley sub-discipline of educational technology[2]. A document is added to a sub-discipline, if it has at least one reader from this sub-discipline. At the point of data collection, there were approximately 2,150 users that had indicated educational technology in their user profile.

To retrieve the most important documents, the document list was sorted by the number of library occurrences within the sub-discipline. We introduced a threshold of 16 occurrences was introduced as selection criterion. This means, a document needs to have been added to at least 16 libraries owned by users who identified themselves as being in the field of educational technology to be included in the analysis, leading to a total of 91 documents. We introduced this threshold to cancel out noise in the data, and to present users with a manageable amount of documents. Since sub-discipline is an optional field in Mendeley, only a minority of users have filled out this field. In order to include more users in Mendeley, the co-occurrence calculation was extended to all user libraries. The 91 documents appeared in 7,414 user libraries with a total of 19,402 co-occurrences.

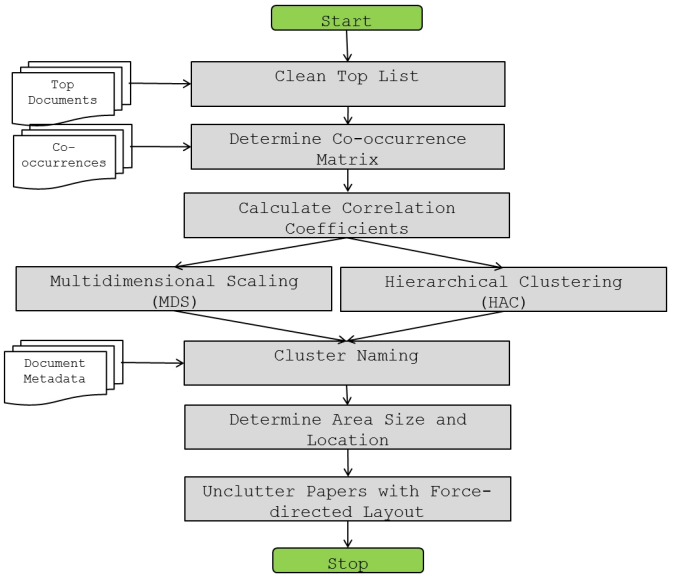

Fig. 4. Overview of the procedure used to create the co-readership visualization.

In a next step, a co-occurrence matrix was created. Based on the co-occurrence matrix, we computed the Pearson correlation coefficient matrix with pairwise complete observations. These correlation coefficients were then used to calculate Euclidean distances between the documents. The matrix of correlation coefficients was the basis for non-metric multidimensional scaling (NMDS) and hierarchical agglomerative clustering (HAC). Multidimensional scaling was used to project the documents into a two-dimensional space, clustering to find topic areas in the projection.

To create labels for the clusters, titles and abstracts of the documents in each cluster were submitted to the APIs of Zemanta[3] and OpenCalais[4]. Both services crawl the semantic web and return a number of concepts that describe the content. The returned concepts were compared to word n-grams generated from titles and abstracts. The more words a concept has (and therefore, the more information it contains), and the more often it occurs within the text, the more likely it is to be the label of the cluster. The results of this procedure were manually checked and corrected if needed.

In order to allow users to interact with the results, we developed an interactive web visualization prototype. The visualization was created using D3.js[5]. In the prototype, documents are represented as rectangles with dogears, a common metaphor, used in many icons and graphics. The size of the document signifies the number of readers it has. Topic areas are

[2]http://www.mendeley.com/disciplines/education/educational-technology/

[3]http://zemanta.com

[4]http://opencalais.com

[5]http://d3js.org

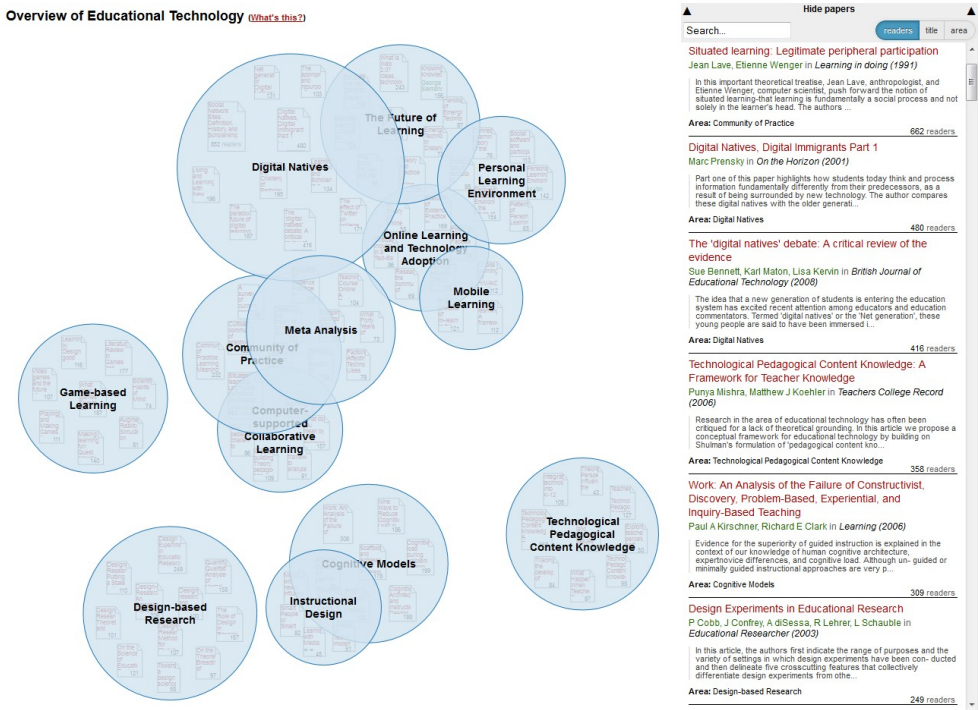

Fig. 5. Knowledge domain visualization of educational technology. The bubbles represent topic areas within the domain. The size of a bubble relates to the number of combined readers.

represented as bubbles. The center of each bubble is calculated as the mean of the coordinates of the publications based on the NMDS result. The size of the bubble is determined by the number of combined readers of the publications in the topic area.

Additionally, force-directed placement was employed on the documents to unclutter the visualization and move documents into their respective topic areas. To prevent overlapping documents, the collision detection algorithm by Mike Bostock[6] was used.

C. Results

The resulting knowledge domain visualization prototype, which can be accessed on Mendeley Labs[7], is shown in Figures 5. In the first few seconds of the visualization, the force-directed placement algorithm is executed. The papers are untangled and pulled into their respective areas, represented by the blue bubbles. After the force-directed algorithm has finished, users can interact with the visualization. The interaction design follows the well-tested approach of "overview first, zoom and filter, then details-on-demand" [4]. Once a user clicks on a bubble, he or she is presented with relevant documents for that area. By clicking on one of the documents, a user can access all meta data for that document. If a preview is available, it can be retrieved by clicking on the thumbnail in the meta data panel. By clicking on the white background, one can then zoom out and inspect another area.

1) Topic Area Description and Distribution: There are 13 topic areas in the visualization with a combined readership of 13,630 at the time of data collection. The topic areas can again be assigned to meta-areas. On the top of the map (see Figure 5), social and technological developments are being discussed (in *Digital Natives* and *The Future of Learning*). Beneath, there is a large cluster of learning methods and technologies, spanning *Mobile Learning*, *Personal Learning Environment*, *Online Learning and Technology Adoption*, *Community of Practice*, and *Game-based Learning*. On the bottom of the visualization, there is a cluster of topic areas that form the psychological, pedagogical, and methodological foundations of the field. The areas *Computer-supported Collaborative Learning*, *Instructional Design* and *Cognition* relate to psychology, while *Technological Pedagogical Content Knowledge* relates to pedagogy. Research methods are represented by *Design-based Research*.

From what was mentioned above, it follows that pedagogical and psychological topics are covered very well in the visualization. However, topic areas that are largely influenced by computer science such as *Adaptive Hypermedia* or knowledge management (e.g. *Work-integrated Learning*) are missing from the overview. The bias towards disciplines strongly related to education can be explained by Mendeley's discipline taxonomy which was used to determine the paper pre-selection in this study. Even though educational technology is an interdisciplinary field, it appears solely as a sub-discipline of education. The sign-up process in Mendeley requires a user to first select a discipline such as education, social science, or computer and information science. In a second step, a user can select a sub-discipline, such as educational technology. Therefore, a scholar in educational technology with a background in computer science will conclude after

[6]http://bl.ocks.org/mbostock/3231298

[7]http://labs.mendeley.com/headstart. The source code can be obtained from https://github.com/pkraker/Headstart

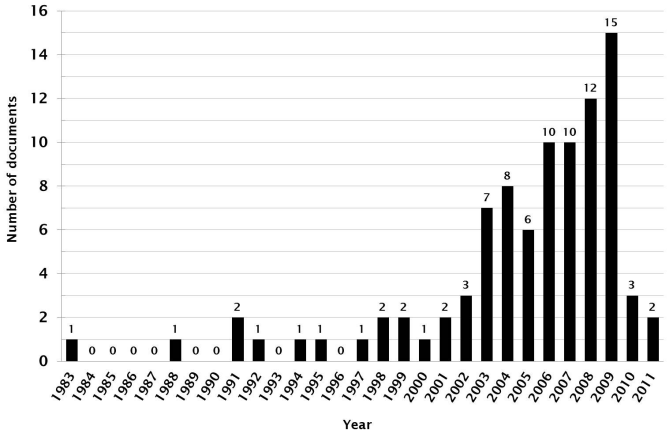

Fig. 6. Distribution of publication years of documents in the visualization (n=91)

the first step that his or her sub-discipline is not represented in Mendeley and choose another one.

At this point it should be mentioned that all scientometric analyses are subject to bias; in a study of downloads in an institutional repository [5], the authors found great differences in the correlation of usage impact factor and journal impact factor depending on the user base.

2) Age of Publications: Figure 6 shows the age distribution of the 91 publications covered in the visualization. 80% of publications were published from 2003 onwards, meaning that they were younger than ten years at the time of data collection (10 August 2012). Most documents were published in 2009. The median age of publications is 6.0 years (Mean = 7.3 years).

While this constitutes a contemporary selection of publications, the relative low proportion of articles younger than two years indicates that not all of the latest developments might be represented in the visualization. However, in a comparable co-citation mapping effort in educational technology [6], the mean age of papers was 14.1 years (Median = 14 years) which is almost double the age of publications in the co-readership analysis. In addition, only 18% of the 28 papers included in the co-citation analysis were less than 10 years of age.

IV. Evaluation

The visualization was evaluated with (1) a qualitative comparison to knowledge domain visualizations based on citations [6] [7], and (2) semi-structured interviews involving the use of the system with experts from the domain of educational technology. The paper accompanying paper is currently under review [8].

The qualitative comparison showed that topics covered in more recent literature such as participatory learning and technological pedagogical content knowledge are better represented in the co-readership visualization. The expert interviews continued this notion but they also revealed that some of the most recent developments such as MOOCs are not included.

The qualitative comparison furthermore showed that the co-readership analysis covers more areas than the co-citation analyses. There is still room for improvement though, as the experts pointed out that in some instances important papers were missing.

An analysis of the spatial features of the maps showed that there were many similarities among the maps created using co-citation and the co-readership visualization. The topical similarity also worked well, with only a few exceptions. Experts were torn, however, on the question of what the centrality of a bubble implies. The same is true for the size of the bubbles. Therefore, it will be important to conduct further research into the meaning of these concepts and provide users of the visualization with an adequate explanation.

V. Conclusions and Future Work

In our paper [1], we analyzed the adequacy and applicability of readership statistics recorded in social reference management systems for creating knowledge domain visualizations. We propose co-readership as a measure of subject similarity. An analysis of the distribution of subject areas in user libraries of educational technology researchers on Mendeley shows that 69.2% of the journal articles in an average user library can be attributed to a single subject area. This is in line with an earlier study [2] which finds that clusters based on the occurrence and co-occurrence of articles in user libraries of CiteULike are as effective as citation-based clusters.

The prototypical visualization based on co-readership patterns of the field of educational technology comprises of 13 topic areas, which can be aggregated to meta-clusters, therefore strengthening the assumption that co-readership indicates subject similarity. The visualization is a recent representation of the field: 80% of the publications included are from within ten years of data collection. However, not all of the latest developments were represented in the visualization due to the fact that it is harder to reach threshold values for the most recent publications. Nevertheless, the papers included in the co-readership analysis are on average almost half as young as the papers included in a comparable co-citation analysis by [6]. This suggests that co-readership analysis may be able to represent more recent aspects than co-citation. In order to generalize this statement and to better understand the differences between co-citation analysis, bibliographic coupling, and co-readership analysis, however, comparison studies between the different similarity measures must be carried out.

The characteristics of the readers introduce certain biases to the visualization. All scientometric analyses are subject to bias; it is therefore important that the characteristics of the underlying sample are made transparent. In the co-readership analysis, information encoded in the user profiles can be used to explain these characteristics. In the present study, a majority of readers were self-ascribed to the field of education and they came from an English-speaking country. This resulted in a map that represents an education science-dominated view from mainly an Anglo-American perspective.

One of the limitations of this work is that the methodology has only been tested for a single field of research. In the future, this study must therefore be repeated in other fields of research. This could be especially interesting for those fields that are dynamic in nature, and those that have not been scientometrically analyzed before due to the lack of citation data.

When applied to larger collections of documents, the procedure used in this paper may be problematic. Both hierarchical clustering and multidimensional scaling have a high computational complexity. Therefore, it will be important to investigate algorithms that can deal with large data sets such as force-directed layout for ordination, and community detection for the establishment of topic areas.

Finally, it seems promising to harness information encoded in the user profiles, such as location, discipline, and career stage, not only for a better understanding of the results (see above), but also for filtering the visualization. This would make it possible to compare visualizations, for instance between countries or career stages. Furthermore, with the availability of timestamps, it becomes possible to show the evolution of a research field over time at a granular level of detail.

ACKNOWLEDGMENT

Figures 1 and 2 were created by Maxi Schramm. 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] P. Kraker, C. Schlögl, K. Jack, and S. Lindstaedt, "Visualization of co-readership patterns from an online reference management system," *Journal of Informetrics*, vol. 9, no. 1, pp. 169–182, Jan. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1751157714001151http://arxiv.org/abs/1409.0348

[2] 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*. ACM, 2011, pp. 401–402.

[3] K. Börner, C. Chen, and K. W. Boyack, "Visualizing knowledge domains," *Annual Review of Information Science and Technology*, vol. 37, no. 1, pp. 179–255, Jan. 2003.

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

[5] J. Bollen and H. V. D. Sompel, "Usage Impact Factor: The Effects of Sample Characteristics on Usage-Based Impact Metrics," *Journal of the American Society for Information Science*, vol. 59, no. 1998, pp. 136–149, 2008.

[6] Y. Cho, S. Park, S. J. Jo, and S. Suh, "The landscape of educational technology viewed from the ETR&D journal," *British Journal of Educational Technology*, Aug. 2012.

[7] L.-C. Chen and Y.-H. Lien, "Using author co-citation analysis to examine the intellectual structure of e-learning: A MIS perspective," *Scientometrics*, vol. 89, no. 3, pp. 867–886, Jul. 2011.

[8] P. Kraker, "Educational Technology as Seen Through the Eyes of the Readers," *arXiv Preprint*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6462