Visualization of Co-Readership Patterns from an Online Reference Management System

Peter Kraker\(^a\)*, Christian Schlögl\(^b\), Kris Jack\(^c\), Stefanie Lindstaedt\(^d\)

\(^a\)Know-Center, Inffeldgasse 13, 8010 Graz, Austria
\(^b\)University of Graz, Universitätsstraße 15, 8010 Graz, Austria
\(^c\)Mendeley, 144a Clerkenwell Road, EC2 RF London, UK
\(^d\)Know-Center, Inffeldgasse 13, 8010 Graz, Austria

Abstract

In this paper, we present a methodology and prototype for creating knowledge domain visualizations based on readership statistics recorded in the social reference management system Mendeley. In particular, we use co-readership patterns to map the field of educational technology. The resulting visualization, based on the most read publications in this field on Mendeley, reveals 13 subject 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.

Keywords: relational scientometrics, knowledge domain visualization, mapping, altmetrics, visualization methodology, readership statistics

1. Introduction

In recent scientometric literature, usage data is being discussed as a valuable alternative to citations. With the advent of e-journals, digital libraries, and

*Corresponding author

Email addresses: pkraker@know-center.at (Peter Kraker),
christian.schloegl@uni-graz.at (Christian Schlögl), kris.jack@mendeley.com (Kris Jack), slind@know-center.at (Stefanie Lindstaedt)
web-based archives, click and download data have been suggested as a potential alternative to citations (Kurtz et al., 2005; Rowlands and Nicholas, 2007). Compared to citation data, usage data has the advantage of being available earlier, shortly after the paper has been published. In many instances, usage statistics are also easier to obtain and collect (Bollen et al., 2005; Brody et al., 2006; Haustein and Siebenlist, 2011). Furthermore, usage statistics allow for an analysis of publications and research outputs that do not receive citations or for which citations are not tracked (Priem and Hemminger, 2010).

Among alternative metrics, readership statistics sourced from social reference management tools like BibSonomy[1] and Mendeley[2] have been of high interest. It has been shown that readership statistics provide a good coverage of top publications (Bar-Ilan et al., 2012), and that there is a medium correlation between readership data and citations (Schlögl et al., 2013) and a medium to high correlation between the impact factor and journal readership (Kraker et al., 2012). Furthermore, Jiang et al. (2011) employ readership statistics from CiteULike to form clusters based on the occurrence and co-occurrence of articles in user libraries. They also correlate these clusters with ISI subject categories, and find them as effective as citation-based clusters when removing journals that cannot be found in CiteULike.

Therefore, we assume that co-readership can be used as a measure of subject similarity. Co-readership relation between two documents is established when at least one user has added the two documents to his or her user library (see Figure 1). The more often the same two documents have been added to user libraries, the more likely they are of the same or a similar subject. 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 has not been exploited before for knowledge domain visualization.

In this study, we employ co-readership patterns for knowledge domain visualization to explore the field of educational technology. Educational technology is multi-disciplinary and highly dynamic in nature, as it is influenced by changes in pedagogical concepts and emerging technologies (Siemens and Tittenberger 2009), as well as social change (Czerniewicz 2010). Therefore, it seemed to be especially appropriate for this kind of analysis.

2. Related Work

Traditionally, knowledge domain visualizations are based on citations. Small (1973) and Marshakova (1973) proposed co-citation as a measure of subject similarity and co-occurrence of ideas (see Figure 3 left side, for a graphical
representation of the relationship). This relationship can be employed to cluster documents, authors, or journals from a certain field and to map them in a two-dimensional space. Co-citation analysis has been used to map many fields, for instance information management (Schlögl, 2001, p. 48), hypertext (Chen and Carr, 1999), and educational technology (Chen and Lien, 2011) to name just a few. Furthermore, co-citation analysis has also been used to map out all of science (Small, 1999; Boyack et al., 2005).

There is, however, a significant problem with citations: they take a long time to appear. It takes around two to six years after an article is published before the citation count peaks (Amin and Mabe, 2000). Therefore, visualizations based on co-citations - and indeed all analyses that are based on incoming citations - have to deal with this time lag. Bibliographic coupling (Kessler, 1963) presents an alternative to co-citation analysis; it is formed when two documents cite the same source document (see Figure 3, right side). The more publications in the reference list the two documents have in common, the more related they are.

Bibliographic coupling is based on outgoing citations available at the time of publication and can therefore be used to map the research front. One difference between bibliographic coupling and co-citation analysis is that the former is a retrospective method (Garfield, 2001), which means that the results cannot change over time unless new publications are added. Another downside of
bibliographic coupling is that prior knowledge is needed to define which publications are part of the research front. For an overview of the properties and the performance of the two citation-based mapping techniques refer to Egghe and Rousseau (1990, chap. III.4) and Boyack and Klavans (2010).

In contrast to citations, usage statistics have been almost exclusively used in evaluative scientometrics (see e.g. Darmoni et al., 2002; Bollen et al., 2007; Schloegl and Gorraiz, 2010). There are only a handful of examples in relational scientometrics and knowledge domain visualization. One of the first are Polanco et al., who propose to use co-occurrences of document requests for clustering and mapping. Bollen and van de Sompel (2006) use consecutive accesses to journal articles as a measure of journal relationships. They derive clusters of journals which are statistically significantly related to ISI subject categories. For a later paper, Bollen et al. (2009) create an overview map of all of science. The authors aggregate hundreds of millions of user interactions with digital libraries and bibliographic databases. Then, they re-create click-streams for each user, aggregated by journal, and apply network analysis to them.

In social reference management tools we can go beyond mere usage: we are able to inspect the users’ library data. This is an improvement in several regards; first, we are able to use library co-occurrence from a single service as a basis for mapping the intellectual structure of a scientific domain. Second, being able to precisely attribute papers to individual readers allows for a better understanding of the results. With the help of profile information, we can furthermore analyze the influence of different geographic regions or career stages.

3. Data Source

All data for this study was sourced from Mendeley on 10 August 2012. Mendeley provides users with software tools that support them in conducting research (Henning and Reichelt, 2008). One of the most popular of these tools is Mendeley Desktop, a cross-platform, freely downloadable PDF and reference management application. It allows users to organize their personal libraries
into folders and apply tags to them for later retrieval. The articles, added by users around the world, are then crowd-sourced into a single collection called the Mendeley research catalog (Hammerton et al., 2012). At the time of writing, this catalog contains more than one hundred million unique articles, crowd-sourced from over two and a half million users.

The users of Mendeley do not only help with building the catalog but also with structuring it. Users can identify themselves as belonging to a scientific discipline and optionally also to a sub-discipline. In August 2012, Mendeley offered 25 disciplines (see Table 1), and 473 sub-disciplines (see Table 2 for the sub-disciplines of “Education”). Each time, a user from a certain (sub-)discipline adds a document to his or her library, the document is automatically assigned to this (sub-)discipline in the catalog.

The following data sets have been sourced on 10 August 2012 and represent data for the sub-discipline educational technology that had been accumulated in the system up to that point:

- **Documents**: all documents in the field of educational technology (n=144,500 documents)
- **Co-occurrences**: co-occurrences of these documents in Mendeley user libraries (n=56,049,431 co-occurrences)

4. Methodology

We followed the knowledge domain visualization process as proposed by Börner et al. (2003). 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

---

3 As a result, a document can be assigned to more than one (sub-)disciplines.

4 Co-occurrence calculation is a computationally intensive process. Therefore, the number of documents per user library was capped to 500. If a user library contained more than 500 documents from educational technology, 500 documents were randomly selected. Then the co-occurrences were calculated.
<table>
<thead>
<tr>
<th>Arts and Literature</th>
<th>Astronomy / Astrophysics / Space Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological Sciences</td>
<td>Business Administration</td>
</tr>
<tr>
<td>Chemistry</td>
<td>Computer and Information Science</td>
</tr>
<tr>
<td>Design</td>
<td>Earth Sciences</td>
</tr>
<tr>
<td>Economics</td>
<td>Education</td>
</tr>
<tr>
<td>Electrical and Electronic Engineering Engineering</td>
<td></td>
</tr>
<tr>
<td>Environmental Sciences</td>
<td>Humanities</td>
</tr>
<tr>
<td>Law</td>
<td>Linguistics</td>
</tr>
<tr>
<td>Management Science / Operations Research</td>
<td>Materials Science</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Medicine</td>
</tr>
<tr>
<td>Philosophy</td>
<td>Physics</td>
</tr>
<tr>
<td>Psychology</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>Sports and Recreation</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: List of the 25 disciplines in the Mendeley catalog (Source: http://www.mendeley.com/research-papers/)

<table>
<thead>
<tr>
<th>Business Education</th>
<th>Comparative Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counselling</td>
<td>Curriculum Studies</td>
</tr>
<tr>
<td>Education Research</td>
<td>Educational Administration</td>
</tr>
<tr>
<td>Educational Change</td>
<td>Educational Technology</td>
</tr>
<tr>
<td>Language Education</td>
<td>Mathematics Education</td>
</tr>
<tr>
<td>Medical Education</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>Physical Education</td>
<td>Science Education</td>
</tr>
<tr>
<td>Sociology of Education</td>
<td>Special Education</td>
</tr>
<tr>
<td>Teacher Education</td>
<td>Testing and Evaluation</td>
</tr>
</tbody>
</table>

Table 2: List of the 18 sub-disciplines of “Education” (Source: http://www.mendeley.com/disciplines/education/)
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

4.1. Data Selection and Pre-processing

The documents included in the analysis were taken from the Mendeley sub-discipline of educational technology. As mentioned before, 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

[http://www.mendeley.com/disciplines/education/educational-technology/]
educational technology in their user profile.[6]

To retrieve the most important documents, the document list was sorted by the number of library occurrences within the sub-discipline. 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. The threshold was chosen upon manual inspection of the results. For smaller threshold values (and consequently, larger amounts of documents), the resulting clusters proved to be less coherent.[7] 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.

In a next step, a co-occurrence matrix was created. In line with McCain (1990), diagonal values were treated as missing values. In addition, document pairs with no combined readership were treated as missing values.[8]

Based on the co-occurrence matrix, we computed the Pearson correlation coefficient matrix with pairwise complete observations. These correlation coeffi-

---

[6] User libraries were sourced at a later point (23 January 2013). Only users that signed up before 10 August 2012 were considered to ensure congruency with the rest of the data set. However, there might be certain shifts in the user base: in the case that users changed their sub-discipline, Mendeley provided only the most recent one chosen.

[7] This is most likely due to the fact that the co-occurrence matrix (see next step) became relatively sparse when we added more documents. In the future, this problem might be mitigated by the rising number of users in Mendeley.

[8] Usually, these cases are put down as zero co-occurrences. We did not find much difference between the two variations, but in the case of missing values, the clusters were more appropriate and coherent. The clusters also proved to be more stable in a bootstrapping analysis. One reason for this could be that the matrix in co-readership analysis is less sparse than in co-citation analysis. Treating document pairs with no combined readership as missing values might therefore serve as a better indicator of discrimination between documents. Therefore, the missing values approach was chosen. Nevertheless, it remains to be determined whether this will hold true for future data sets.
coefficients were then used to calculate Euclidean distances between the documents.

4.2. Clustering and Mapping

The matrix of correlation coefficients was the basis for multidimensional scaling (MDS) and hierarchical agglomerative clustering (HAC). Multidimensional scaling was used to project the documents into a two-dimensional space, clustering to find research areas in the projection. For hierarchical agglomerative clustering, we employed Ward’s method (minimum variance) using the R command \texttt{hclust}. Ward’s method successively merges those two clusters that minimize the increase in the total within-cluster variance (Hair et al., 2010, p. 510). It is known to join smaller clusters and to produce clusters of approximately the same size (Tan et al., 2007, p. 523).

The number of clusters was determined by the elbow method using the R function \texttt{elbow.batch}. This function defines an elbow when the number of clusters \( k \) explains at least 80% of the variance in the model, and when the increment is lower than 1% for a bigger \( k \). This criterion was reached at an explained variance of 84% and lead to a total of 13 clusters.

In a second step, we plotted the results in a two-dimensional space with non-metric multidimensional scaling (NMDS). NMDS is often employed in scientific mapping efforts. Examples can be found in White and McCain (1998) and in Tsay et al. (2003). NMDS is an iterative approach, similar to regression analysis. Beginning with a random start configuration, it tries to minimize a given stress function in consecutive steps. Since NMDS is prone to reaching local minima, usually a number of random starts are used to find an optimum solution.

We selected the implementation provided by the R ecodist package (Goslee and Urban, 2007). It uses a modified stress function:

\[
\text{Stress} = \sqrt{\frac{\sum_{h,i} (d_{hi} - \hat{d}_{hi})^2}{\sum_{h,i} \hat{d}_{hi}^2}} \tag{1}
\]

\( d_{hi} \): dissimilarity between samples \( h \) and \( i \)
\( \hat{d}_{hi} \): distance predicted by regression

This implementation proved to produce the most clearly separable clusters in comparison to implementations that use the original stress function proposed by Kruskal (1964).

In the NMDS, a stress value of 0.2 was achieved which is the upper bound for an acceptable MDS result as described by Kruskal (1964). The R\(^2\) is reported as 0.86 by the NMDS. According to Hair et al. (2010), acceptable results for R\(^2\) start at 0.60.

To create labels for the clusters, titles and abstracts of the documents in each cluster were submitted to the APIs of Zemanta\(^9\) and OpenCalais\(^10\). 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 area. The results of this procedure were manually checked and corrected if needed.

A plot of the results from the procedure described above can be seen in Figure 5. Each symbol represents a document. The type of symbol signifies the research area it belongs to. These 13 areas are listed in the legend below the graph.

4.3. Web Visualization

In order to allow users to interact with this graph, we developed an interactive web visualization prototype. The visualization was created using D3.js\(^{11}\) 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. To avoid coding the documents with

\(^{9}\)http://zemanta.com
\(^{10}\)http://opencalais.com
\(^{11}\)http://d3js.org
Figure 5: Result of NMDS, HAC, and the naming algorithm. Each symbol represents a document, all of the documents with the same symbol constitute a research area.

symbols (as in Figure 5), research areas are 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 area.

Additionally, force-directed placement ([Fruchterman and Reingold 1991] was employed to unclutter the visualization and move documents into their
respective areas\textsuperscript{12}. To prevent overlapping documents, the collision detection algorithm by Mike Bostock\textsuperscript{13} was used.

It is important to note that - in contrast to the areas - the relative closeness of documents in the visualization does not necessarily imply relative subject similarity\textsuperscript{14}. To review the relationship between individual papers, one needs to go back to the original output of the MDS shown in Figure 5.

5. Results

The resulting visualization prototype, which can be accessed on Mendeley Labs\textsuperscript{15} is shown in Figures 6. 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. Regarding the interaction design, we followed the well-tested approach of “overview first, zoom and filter, then details-on-demand” (Shneiderman, 1996). Once a user clicks on a bubble, he or she is presented with relevant documents for that area (see Figure 7). The meta data of each document is displayed in the document representation itself. It consists of the most common meta data: title, author(s), year, and journal/conference name (if applicable).

The dropdown on the right displays the same data in list form with an added abstract. By clicking on one of the documents, a user can access all meta data for that document (see Figure 8). If a preview is available, it can be retrieved by clicking on the thumbnail in the meta data panel. In addition, a user can

\textsuperscript{12}The area centers were denoted as gravitational centers. Documents not within the limits of the area were instructed to move towards the gravitational center.

\textsuperscript{13}\url{http://bl.ocks.org/mbostock/3231298}

\textsuperscript{14}In the uncluttering effort using force-directed placement, the positions of documents are changed in a way that does not necessarily preserve the relative distances. Therefore, the distances between documents in the visualization do not represent the distances calculated with MDS anymore.

\textsuperscript{15}\url{http://labs.mendeley.com/headstart} The source code can be obtained from https://knowminer.at/svn/opensource/other-licenses/lgpl_v3/headstart/
Figure 6: Knowledge domain visualization of educational technology. The bubbles represent areas within the domain. The size of a bubble relates to the number of combined readers.

filter the publications by entering terms in the search field on top of the list (see Figures 6 and 7). Only publications that contain all of the search terms (Boolean AND) are displayed within the bubbles and the list. The list can be sorted by title, area, and number of readers to facilitate exploration via the list format.

5.1. Area Description and Distribution

As can be seen in Figure 6, there are 13 subject areas in the visualization with a combined readership of 13,630 at the time of data collection (10 August 2012). Table 3 gives an overview of the areas. It shows that they differ in terms of the number of documents and the number of readers. Digital Natives has the highest readership with over 20% of all readers. It has twice the readership of the second largest area: Design-based Research (DBR). DBR includes the most documents (11) of all areas. Community of Practice has only four documents,
but still sports the fourth most readers. The area with the least readers and the least number of documents is Mobile Learning with just 3 documents and a combined readership of 345.

The map is mostly topical, with two exceptions: Meta Analysis is a collection of reviews/state-of-the-art analyses, and Design-based Research represents a specific method. The Future of Learning is also somewhat orthogonal as it describes technological developments.

The areas can again be assigned to meta-areas. These meta-areas are formed by areas that are close to each other, as is assumed by multidimensional scaling. On the top of the map (see Figure 6), 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 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, 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.

Another characteristic of multidimensional scaling is that it shows central and peripheral areas due to their placement on the map (McCain, 1990). Right in the middle, the area Meta Analysis contains reviews of the field. Its central position stems most likely from the fact that these reviews relate to many of the surrounding areas (cp. Figure 5), and that they appear in many user libraries together because of their comprehensive nature. Other central areas are Computer-supported Collaborative Learning, Communities of Practice, Online Learning and Technology Adoption, and Online Learning and Technology Adoption. The remaining areas are more peripheral in the visualization: Game-based Learning, Design-based Research, and Technological Pedagogical Content Knowledge in particular appear rather disconnected from the rest of the other subject areas.

5.2. Publication Types and Age of Publications

The 91 documents in the visualization can be divided into five different types of publications. The majority are journal articles (71 items, or 78%), followed by reports (7), books (6) and book chapters (5), and conference papers (2). The
<table>
<thead>
<tr>
<th>Area</th>
<th>No. Documents</th>
<th>No. Readers</th>
<th>% Readership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Natives</td>
<td>10</td>
<td>2,865</td>
<td>21.0%</td>
</tr>
<tr>
<td>Design-based Research</td>
<td>11</td>
<td>1,477</td>
<td>10.8%</td>
</tr>
<tr>
<td>The Future of Learning</td>
<td>9</td>
<td>1,183</td>
<td>8.7%</td>
</tr>
<tr>
<td>Community of Practice</td>
<td>4</td>
<td>1,175</td>
<td>8.6%</td>
</tr>
<tr>
<td>Cognitive Models</td>
<td>6</td>
<td>1,169</td>
<td>8.6%</td>
</tr>
<tr>
<td>Technological Pedagogical Content Knowledge</td>
<td>9</td>
<td>1,049</td>
<td>7.7%</td>
</tr>
<tr>
<td>Game-based Learning</td>
<td>8</td>
<td>993</td>
<td>7.3%</td>
</tr>
<tr>
<td>Meta Analysis</td>
<td>8</td>
<td>991</td>
<td>7.3%</td>
</tr>
<tr>
<td>Personal Learning Environment</td>
<td>6</td>
<td>648</td>
<td>4.8%</td>
</tr>
<tr>
<td>Online Learning and Technology Adoption</td>
<td>6</td>
<td>637</td>
<td>4.7%</td>
</tr>
<tr>
<td>Computer-supported Collaborative Learning</td>
<td>5</td>
<td>615</td>
<td>4.5%</td>
</tr>
<tr>
<td>Instructional Design</td>
<td>6</td>
<td>483</td>
<td>3.5%</td>
</tr>
<tr>
<td>Mobile Learning</td>
<td>3</td>
<td>345</td>
<td>2.5%</td>
</tr>
<tr>
<td>Sum</td>
<td>91</td>
<td>13,630</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Table 3: Areas in the visualization
71 journal articles were published in a variety of journals. The highest number of articles was published in “Computers & Education” (8), followed by “Educational Technology Research & Development” and “The Internet and Higher Education” (both 6) and “Review of Educational Research”, “Educational Researcher” and “Educational Psychologist” (all 5). These publication outlets are among the highest impact journals in the Journal Citation Reports (Thomson Reuters, 2013). All of the documents in the visualization are in English.

Figure 9 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). The relative small amount of publications from 2010 and 2011 can be explained by the circumstance that it is more difficult for recent publications to reach the threshold value than for older ones.

Classics within the field are still contained in this visualization; for the most part they inform research that is still prevalent today. Examples are “Situated learning: Legitimate peripheral action” (Lave and Wenger, 1991) or “Cognitive load during problem solving: Effects on learning” (Sweller, 1988). An exception is the area “Instructional Design” which contains only documents that were published before 2003. Here, the classic media debate between Clark and Kozma is represented, as well as other older papers relating to instructional design.

6. Discussion

6.1. Recency

In the co-readership analysis, the mean age of publications is 7.3 years with 80% of articles published within 10 years of data collection. 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

\footnote{Calculation based on the exact date of data collection on 10/08/2012}
might be represented in the visualization. However, in a comparable co-citation mapping effort in educational technology by Cho et al. (2012), the mean age of papers was 14.1 years (Median = 14.1 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.

This result shows that in terms of recency, co-readership analysis is much more up-to-date than co-citation analysis. Indeed, co-readership analysis may be closer to bibliographic coupling in terms of recency. In comparison to bibliographic coupling, however, co-readership visualization has a couple of advantages; first of all it is a dynamic method, meaning that the results can change over time. Second, the data employed (readership statistics) allow to select the publications to be included in the analysis by the information given in the user profile.

\[17\] All calculations based on the publication year of the most recent article in the analysis (2011)
6.2. Biases in the Visualization

An analysis of the results shows that the visualization not free from biases. First, all of the papers are in English, even though educational technology is often researched by local communities that communicate in their native language [Ely 2008]. Second, the knowledge domain visualization represents an education-dominated view that lacks areas related to computer science.

Biases in usage statistics analyses were first mentioned by Bollen and Sompel [2008] in a study of downloads in an institutional repository. The authors found great differences in the correlation of usage impact factor and journal impact factor depending on the user base. The authors therefore concluded that these biases occur due to sample characteristics.

Biases affect all scientometric analyses. A problem that arises in citation studies is the selection of the corpus. Criteria for the inclusion of authors and papers in the analysis have an impact on the result. The difference between traditional citation-based analyses and the co-readership analysis is that in the latter case it is easier to explain the biases using information encoded in Mendeley user profiles. Consequently, we looked into the sample characteristics of users in educational technology (n=2,153 user profiles).

At first, we analysed the geographical distribution of users. One of the reasons for the fact that all of the papers are in English is surely that English is the lingua franca in science and research [Tardy 2004]. But most likely, this dominance of English also stems from the fact that there is a strong bias towards English-speaking countries on Mendeley.

This assumption is backed up by the results of the geographical analysis (see Figure 10). Out of 2,153 users, 927 (43.1%) have chosen to list a country in their user profile. In total, 70 countries have been named, but the distribution is highly skewed. There is an emphasis on the US and the UK with a total number of 423 users (45.6%). In fact, when adding Canada and Australia, English-speaking countries have a share of over 54.3%. 56 countries with a low share of users have been summed up under “Other” (21.7%). This shows that Mendeley users come from a wide variety of countries, but that there is a strong
focus on English speaking countries. Two facts play an important role with regards to this focus: first, Mendeley originated in the UK and has an office in the USA. Second, the Mendeley software is only available in English for now.

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 the first step that his or her sub-discipline is not represented in Mendeley and choose another sub-discipline.
7. Conclusions and Future Work

In this paper, we presented a methodology and prototype for creating knowledge domain visualizations from readership statistics. We propose co-readership as a measure of subject similarity. The procedure put forward is fully automated with the exception of choosing the number of publications to include (as it has a profound impact on the clustering result) and correcting some of the labels from the naming algorithm. The subsequent visualization created from co-readership patterns for the field of educational technology comprised of 13 areas, generated from 91 publications with a combined readership of 13,630. The areas 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 might be 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 Cho et al. (2012). This demonstrates that co-readership analysis is able to represent more recent aspects than co-citation. In order to better understand the differences between similarity measures, however, comparison studies between co-citation analysis, bibliographic coupling, and co-readership analysis 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. Educational technology is a diverse field with many influences; but it would not be appropriate to generalize the results to all research fields. The question is whether the same analysis would work as well on a larger set of publications and for other fields and disciplines. Each discipline has its own theories, methods, accepted practices, in short: its own culture. Just like publication and citation practices are fundamentally different for the natural sciences and the humanities, cultural differences might also affect the usage of social reference management systems. In the future, this study must therefore be repeated in other fields of research.

Reproduction of the study in other fields should be easily possible, as the developed procedure is highly automated. 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. It remains to be proven, however, that the presented procedure scales up to larger collections of documents. Both hierarchical clustering and multidimensional scaling have a high computational complexity. Therefore, it might be worthwhile to include other algorithms such as force-directed layout for ordination, and k-means clustering for the establishment of areas. In order to be able to place a given document in several clusters, it would also be interesting to explore factor analysis.

Finally, it seems promising to harness information encoded in the user profiles, such as location, discipline, and career stage, for visualization. Accordingly, it would be possible to create visualizations based on a certain share of the worldwide user base. This opens up the possibility of comparing 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 with a granular level of detail. Data like this could be used to fuel pathfinder networks and other means of showing the development of a domain.
Acknowledgments

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


URL http://arxiv.org/abs/1205.5611

URL http://dl.acm.org/citation.cfm?id=1255273


URL http://www.springerlink.com/index/10.1007/s11192-006-0151-8


URL http://doi.wiley.com/10.1002/asi.21419


URL http://doi.wiley.com/10.1111/j.1365-2729.2010.00359.x

URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC116406/


URL http://doi.wiley.com/10.1111/j.1467-8535.2008.00810.x


URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:From+Bibliographic+Coupling+to+Co-Citation+Analysis+via+Algorithmic+Historio-Bibliography#0

URL http://www.jstatsoft.org/v22/i07/paper/


URL http://www.springerlink.com/index/10.1007/s11192-010-0172-1


