Altmetrics-based Visualizations Depicting the Evolution of a Knowledge Domain[1]

Peter Kraker[*], Philipp Weißensteiner[**] and Peter Brusilovsky[***]

[*]*pkraker@know-center.at*
Marshall Plan Scholar, Know-Center, Inffeldgasse 13/VI, Graz, 8010 (Austria)

[**]*philipp.weissensteiner@student.tugraz.at*
Graz University of Technology, Rechbauerstraße 12, Graz, 8010 (Austria)

[***]*peterb@pitt.edu*
Personalized Adaptive Web Systems Lab, University of Pittsburgh, 135 North Bellefield Avenue, Pittsburgh, 15260 (USA)

Introduction

Altmetrics have recently received a lot of attention in the scientometric community and beyond. A lot of effort is going into assessing the potential of altmetrics for evaluative purposes. Yet another strand of research has emerged, focusing on altmetrics for establishing relations between articles, journals and research fields. One problem that is considered within the latter strand of altmetrics research is knowledge domain visualization. So far, clicks (Bollen and van de Sompel, 2006) and readership (Kraker et al., 2013) have been successfully employed to map a scientific domain. These efforts, however, represent the state of a domain, but not the evolution of a domain.

The evolution of scientific domains has been addressed in the past primarily using citation-based analysis; see e.g. Garfield (1966) and Small (1999). There is, however, a significant problem with citations: it takes around two to six years after an article is published before the citation count peaks (Amin and Mabe, 2000). Therefore, citation-based visualizations, and indeed all analyses that are based on incoming citations, have to deal with this time lag. Altmetrics have thus emerged as a potential alternative to citation data. Compared to citation data, altmetrics have the advantage of being available earlier, many of them shortly after the paper has been published. In many instances, usage statistics are also easier to obtain and collect (Haustein and Siebenlist, 2011).

In this paper, we present work towards showing the evolution of scientific domain using data from a scientific conference scheduling software. Conference Navigator 3 (Parra et al., 2012) allows conference attendees to create a personal conference schedule by bookmarking talks from the program that they intend to follow. This scheduling data represents an altmetrics source which – to the best knowledge of the authors – has not been studied before.

Data and Method

All data is sourced from Conference Navigator 3. As a use case, we have chosen to analyse the 19th and 20th iteration of the Conference on User Modelling, Adaptation and Personalization (UMAP), representing the conference years of 2011 and 2012. The data sets

relevant to this analysis were article metadata of the two UMAP conferences and the scheduling information associated with these articles.

The procedure for creating knowledge domain visualizations follows the approach used in the knowledge domain visualization Head Start which employs Mendeley readership data to map a research field (Kraker et al. 2013). It is adapted from the knowledge domain visualization process described in Börner et al. (2003).

At first, individual knowledge domain visualizations were created for each year. Therefore, the article metadata and the scheduling data have been extracted from the system ranked by the number of bookmarks received. A threshold of 4 (3) bookmarks was introduced for 2011 (2012) to ensure that enough scheduling data for the subsequent processing steps was present. Then, a co-bookmarking matrix for the remaining articles was created. On top of this matrix, we performed multi-dimensional scaling (MDS) and hierarchical agglomerative clustering (HAC) to create the visualization. We used HAC with Ward's method to establish the research areas. We then employed non-metric MDS to place the articles on the map. To unclutter the articles in the map, a force-directed layout with a collision detection algorithm was used. Finally, article titles and abstracts were sent to the text mining services Zemanta and Open Calais to find appropriate labels for the areas. Both services return a number of concepts that describe the content.

As far as time series visualization goes, there are many types of visualizations, most prominently index charts and stacked graphs. In the case of knowledge domain visualizations, simple visualizations are unfortunately not able to convey all necessary dimensions of the data (in terms of ordination, size of research areas and closeness). One possibility would be to use animation to show changes in a domain over time. Psychological studies have shown, however, that people are bad at recognizing change in an object or a scene. This phenomenon is called change blindness (Simons and Rensink, 2005). Therefore, the approach of small multiples (Tufte, 1990) was chosen. In small multiples, a graph is drawn for each of the steps in a time series. Then the graphs are positioned next to each other. This approach thus allows for direct visual comparison between different representations.

Results
The result of the visualization procedure for a single year detailed above can be seen in Figure 1. The blue bubbles represent research areas. The centre point of each circle was calculated as the means of the coordinates of the articles based on the MDS result. The size of the areas is determined by number of bookmarks that the papers have received. Spatial closeness implies topical similarities. As can be seen, "User modeling" is the area with most papers and most bookmarks. It is closely connected to several other larger areas, including "Recommender system". A second cluster of areas can be found on the right hand side of the visualization, involving "Intelligent tutoring system", "Adaptive system", and "Problem solving".

Figure 2 shows the evolution of the topical overview using small multiples. To aid the user in detecting changes between the representations, we introduced two visual helpers. First, a grid is drawn to help with comparing size and position of the research areas. Second, whenever users hover over an area, the corresponding area is highlighted in the other representation, and a line is drawn between the two entities. There are three areas that are present in both years: "User modelling", "Recommender system" and "Intelligent tutoring system". While the relative position of the areas to each other has not changed much, the area with the most papers and bookmarks is now "Recommender system".

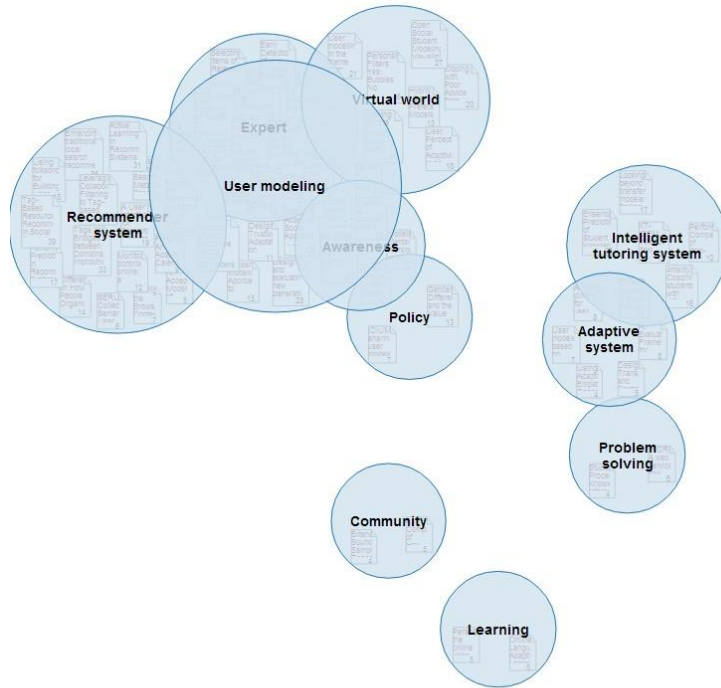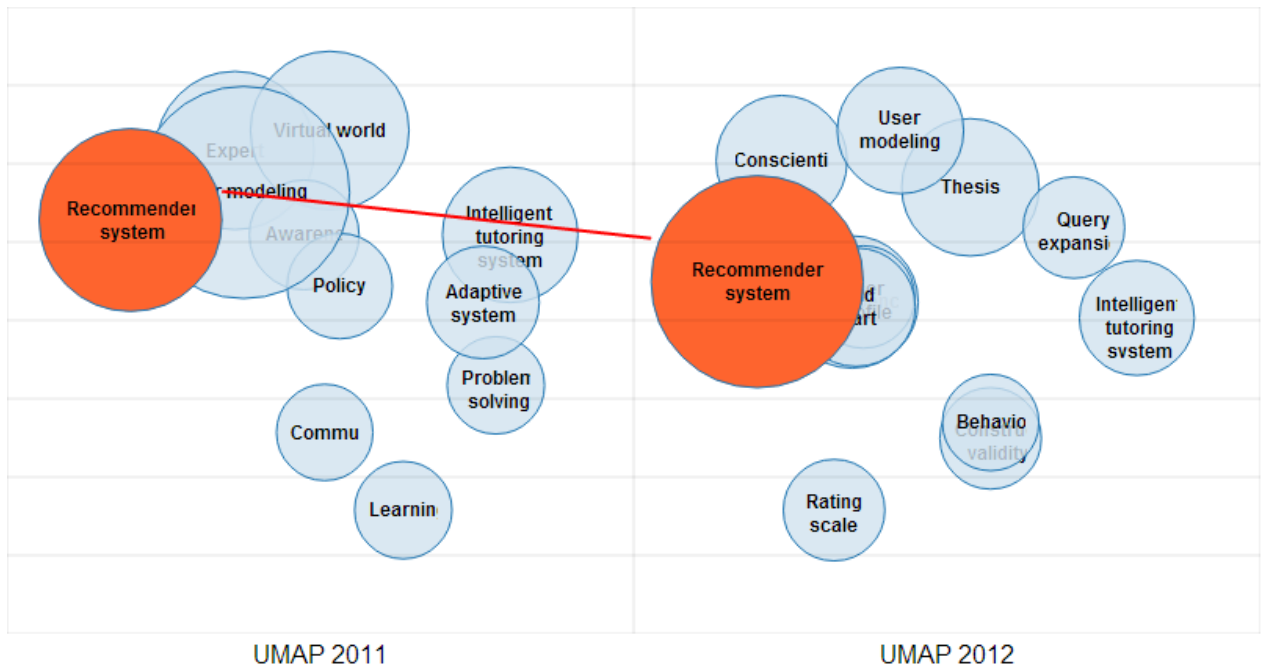Figure 1: Topical visualization of UMAP 2011.

Figure 2: Evolution of the domain of UMAP from 2011 to 2012, with the area "Recommender system" highlighted

Discussion and Future Work
The first results from this type of analysis are encouraging. Using small multiples allows for a comparison of knowledge domain visualizations over various years. The work, however, also revealed several weaknesses of the current approach. First, the topology needs to be

improved. As many of the areas are overlapping, it becomes harder and harder to disambiguate between the different clusters. One way to overcome this problem might be to employ force-directed placement.

Second, scheduling data in CN3 is sparser than readership data in Mendeley due to the fact that the audience of CN3 is restricted to conference participants. This means that the results of the clustering vary more when choosing different threshold values. Therefore, we want to explore supplementing bookmarks with content-based measures.

Third, the continuity between the two years is very low. There are only three areas that are present in both iterations. Therefore, it might be worthwhile to explore moving time windows of two years to show how the different papers move when associated with earlier or later years.

Finally, it will be important to evaluate the method and the interface. One way to go about this would be to ask experts to critically review the visualizations, and to give them the ability to manipulate size and location of papers and areas. This would allow for comparing the automatically created visualizations to the experts' perception of the field. Another possibility would be to contrast the visualization based on conference papers with the evolution of the field based on other types of literature (e.g. journal articles).

References
Bollen, J., & van de Sompel, H. (2006). Mapping the structure of science through usage. *Scientometrics*, *69*(2), 227–258. doi:10.1007/s11192-006-0151-8

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179–255. doi:10.1002/aris.1440370106

Garfield, E., Sher, I., & Torpie, R. (1964). *The use of citation data in writing the history of science* (p. 75).

Haustein, S., & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics*, *5*(3), 446–457.

Kraker, P., Jack, K., Schlögl, C., Trattner, C., & Lindstaedt, S. (2013). Head Start: Improving Academic Literature Search with Overview Visualizations based on Readership Statistics. In *Web Science 2013*.

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

Parra, D., Jeng, W., Brusilovsky, P., López, C., & Sahebi, S. (2012). Conference Navigator 3: An online social conference support system. In *UMAP Workshops* (pp. 1–4).

Simons, D. J., & Rensink, R. A. (2005). Change blindness: past, present, and future. *Trends in Cognitive Sciences*, *9*(1), 16–20. doi:10.1016/j.tics.2004.11.006

Tufte, E. R. (1990). *Envisioning Information* (p. 127). Graphics Press.