Research Fronts and Areal Density of Bibliographically Coupled Publications

Edgar Schiebel

edgar.schiebel@ait.ac.at
AIT Austrian Institute of Technology GmbH, Technology Management, Donau City Straße, 1220 Vienna, Austria

Abstract

This article proposes a method for the detection and visualization of research fronts within a broader research field. A research front is defined as a grouping of publications on a research topic using the same knowledge base in a broader context. The broader context is defined by a field such as battery research, casting or tribology, while the knowledge base is defined by references.

Some publications report on the usage of keywords, co-citations or bibliographic coupling in the detection of research fronts. In this article, agglomerations of bibliographically coupled publications with a common knowledge base are identified and graphically represented by a density function of publications per area unit. The knowledge base becomes visible if publications with similar vectors of common citations are associated. All bibliographically coupled publications are positioned on a map of points generated by a spring model. The publications positioned in a two dimensional space form areas of differing density of points per area unit. The space dependent density is calculated using a digital grid. The density function of papers per area unit is visualized in a three dimensional plot. In this way, research fronts become visible and can be quantified. The proposed methodology is demonstrated based on a case study in the field of battery research.

Introduction

Several publications report on the detection and visualization of research fronts. Price (1965) introduced the concept of research fronts based on citations and Chen & Morris (2003) identified clusters of co-cited articles. In recent work Shibata et al, (2009) analysed the performance co-citation, bibliographic coupling and direct citation in detecting research fronts. Boyack & Klavans, (2010) added a bibliographic coupling-based citation-text hybrid approach to the three mentioned bibliometric approaches and compared accuracies of cluster solutions for a very large set of articles. Both author groups gave a sophisticated overview over science mapping methods to detect emerging research fronts.

The objective of the present work is to propose a visualisation technique to detect, visualize and quantify agglomerations of publications in a map. Research fronts are understood as an amount of research activity reported by publications on a research topic such as the development of a battery electrode using nanotechnology. Scientists working in such an area use previously published knowledge shared by their colleagues in a publication or in a presentation given at a conference. The trickier such a research topic, the more intensive the research work. Persson (1994) distinguishes between the research front and the intellectual basis: “In bibliometric terms, the citing articles form a research front, and the cited articles constitute an intellectual base“, cited from Chen (2006). This means that a knowledge base of a research topic can be made visible by building clusters of papers based on common references. We used this concept to draw maps of bibliographically coupled papers in order to detect research fronts. The methodology is described in this paper.

References were taken as a vector and the scalar product as a similarity measure using the Jaccard index. Large scale maps enable the inclusion of several thousand articles. We are therefore able to obtain an overview of a whole field such as battery research. Two dimensional maps show papers as points. Areas of high or low density are identified with high density areas denoting areas of high research activity on a small research topic.

Methodology:

The methodology starts with the calculation of maps of bibliographically coupled publications in the chosen research area. The map is used to calculate the density of publications per area unit and a x,y,z plot to visualize the density distribution. Further steps include the identification of areas with high research activity, assignment of thematic topics and improving clarity by introducing thresholds to include publications cited at least once and references cited no more than 99 times.

Maps of bibliographically coupled papers

Research on batteries has intensified in recent years due to the development of cars with electric traction and the associated need for storage of electrical energy. In a first step all publications with titles containing the word "battery" were downloaded from the Web of Science Database for the period 2004 to 2010. To draw the map of the bibliographically coupled articles the cited references of the publications had to be extracted and cleaned by deleting volume and pages information. The relations of the publications were stored in a table of relations as shown in Table 1.

Table 1. Table of relations

Cited references	*Publication ID*
AABAKKEN J, 2005, NRELTP62037930	ISI:000257010100014
AADLAND J, 1985, DEV PSYCHOBIOL	ISI:000266030600006
AALTO P, 2001, TELLUS B	ISI:000266556000006
AAMATUCCI GG, 1977, J POWER SOURCES	ISI:000224133300005
AARDAHL C, 2005, 2005 US DEP EN HYDR	ISI:000241012000015
AARDEMA MJ, 1998, MUTAT RES-REV MUTAT	ISI:000258809200003
AARNOUTSE C, 2005, ED RES EVALUATION	ISI:000266720900002
AARON PG, 1991, J LEARN DISABIL	ISI:000256336200003
AARONSON N, 2002, QUAL LIFE RES	ISI:000269369000002
AARONSON N, 2002, QUAL LIFE RES	ISI:000240081900021
AARONSON N, 2002, QUAL LIFE RES	ISI:000242229000013
...	...

A co-occurrence matrix was calculated via a simple relational database operation using the table of relations. The lines and columns are defined by the paper ID and the value is the number of common references.

The concept of bibliographic coupling takes into account that similar papers largely cite the same references. The Jaccard index was used to measure similarity. Ensemble measures like the Pearson correlation coefficient or the cosines of the references vector are not appropriate as the individual similarity is required.

The similarity matrix is projected into a two dimensional space with the help of a spring model using the software BibTechMon^{TM}. In calculating the map it is important to position similar publications at a close distance to each other, irrespective of the visibility of single publications. Parameters of the software allow to manipulate the repulsive forces of not coupled papers and threshold of of the Jaccard index based on its distribution. Techniques like multi dimensional scaling (MDS) are not appropriate as they tend to be equally distributed over the space.

Spatial density function of publications

Following calculation of the map, the publications are positioned in a two dimensional space. The density function p is calculated with a two dimensional digital raster. Thus we get a density function in x and y dimension: p=f(x,y).

The size of the digital raster depends on the number of publications. It should not be much smaller than the square root of the number of publications. If we have around 10,000 publications for example, then the pixel size should be 0.01 x 0.01.
A first graphical representation shows some noise because many publications with no or only few common references with other papers show a low coupling and are indifferently positioned and not associated to agglomerations. The noise can be reduced using a two dimensional moving average. For this purpose a square of n x n pixels is moved over the map calculating the average number of publications. The new density function p' shows the spatial distribution of the bibliographically coupled publications much more clearly.
The maps were calculated using BibTechMon^{TM} and the three dimensional graphs using MS Excel. The density distribution can be further analyzed on the basis of isodensity lines.

Results for Battery Research

A search for titles containing the words "battery" or "batteries" in the databases SCI-EXPANDED, SSCI, CPCI-S and CPCI-SSH showed 8,494 publications for the period from 2004 to 2010.

Landscape of bibliographically coupled publications in battery research

Figure 1 shows a map of 7,761 publications with at least one reference. Each point in the graph is a publication. The similarity between pairs of publications was calculated using the Jaccard index applied to the number of common references. At first glance, the map shows areas of different agglomerations of publications over the two dimensional space. We have areas of high density and areas of low density.

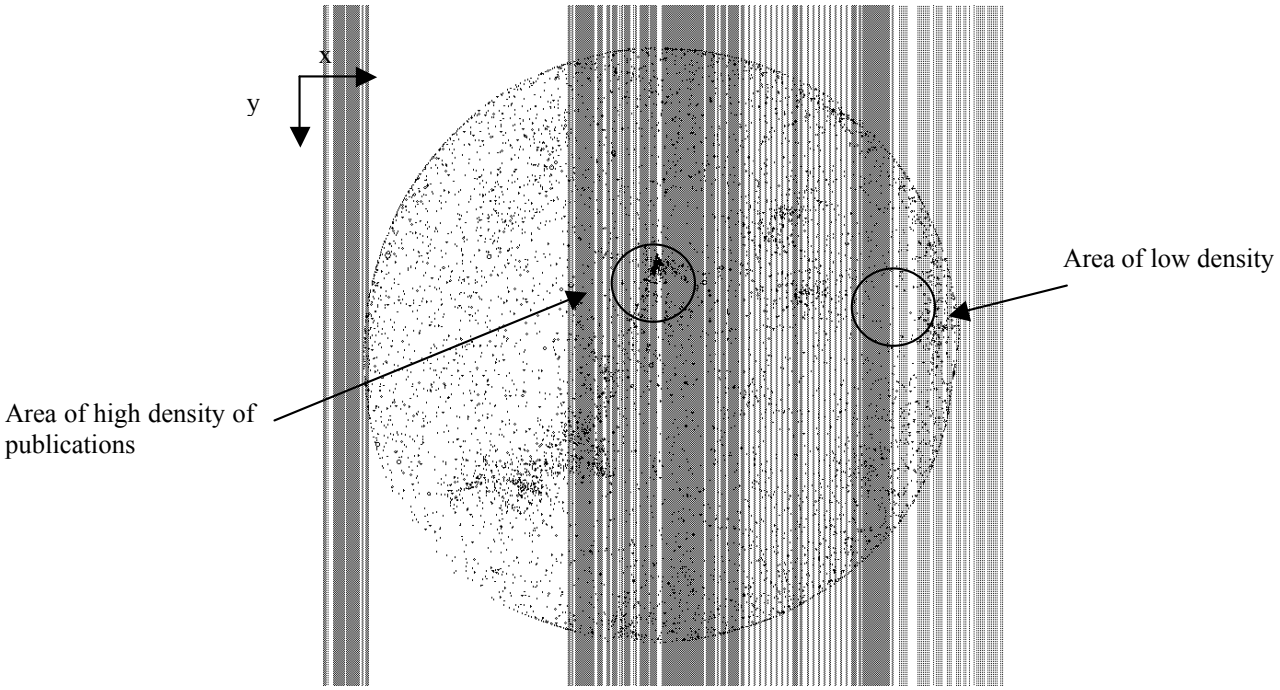

Figure 1. Map of bibliographically coupled publications. Similarity measure: number of common references normalised using the Jaccard index, projection by a spring model, with x between 0.46 and 2.06 and y between 0.45 and 2.06; 7,761 publications on battery research in the period from 2004 to 2010, BibTechMon^{TM} AIT Austrian Institute of Technology GmbH.

A digital grid of 0.01 x 0.01 was used to estimate the density distribution of the two-dimensional x,y-space. The result is shown in Figure 2. A comparison between Figure 1 and

Figure 2 shows a peak in the marked area of high density, reflecting the large number of publications associated by common references. This is an indication that a remarkable number of papers were published and refer to a common knowledge base of cited references.

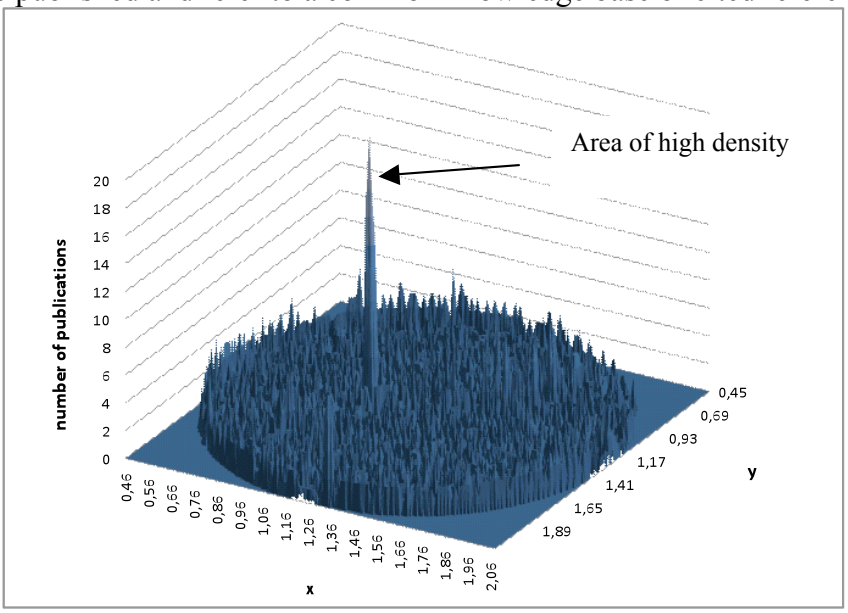

Figure 2. Bibliographically coupled publications. Density of bibliographically coupled publications per area unit 0.01 x 0.01 with x between 0.46 and 2.6 and y between 0.45 and 2.6; 7,761 publications, 2004 to 2010

The smoothened p'(x,y) is shown in Figure 3 and gives a much better view of the landscape. The distribution shows areas of high and low density. It is interesting to note that we have high narrow mountains (LiMetalPO4) as well as lower but broader mountains (anode material). We have to be aware of the fact that publications are similar if they have the same number and kind of references; this means the more similar they are the more they use the same and narrow knowledge base. We can say that areas with very high and narrow mountains represent research with relatively high activity, where all scientists refer to the same knowledge base or cite each other very often. Mountain ranges with several lower peaks represent broader research with slightly variant topics such as different anode materials.

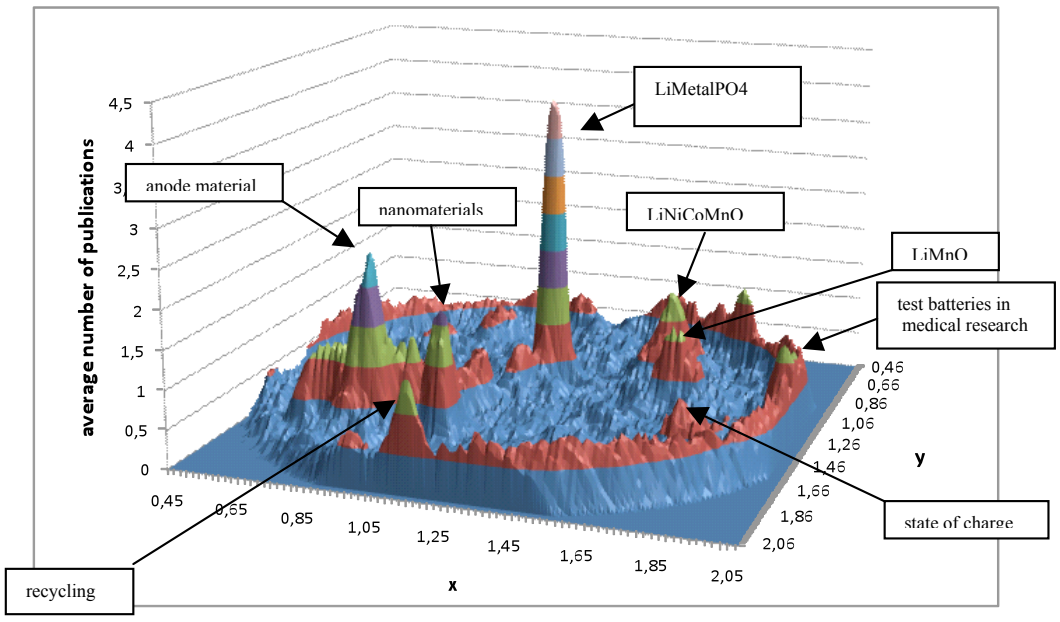

Figure 3. Two dimensional areal density distribution of bibliographically coupled publications, moving average of 7x7 pixels; 7,761 publications, 2004 to 2010

Many publications in a research field are not cited and therefore do not play a visible role in the knowledge base of a research front. Therefore all non-cited articles were excluded. Additionally, some references are cited extremely often and occupy a central position in a co-citation network. These may be very important in a broad sense but do not help to identify clear research fronts. We therefore only included references that were not cited more than 100 times.

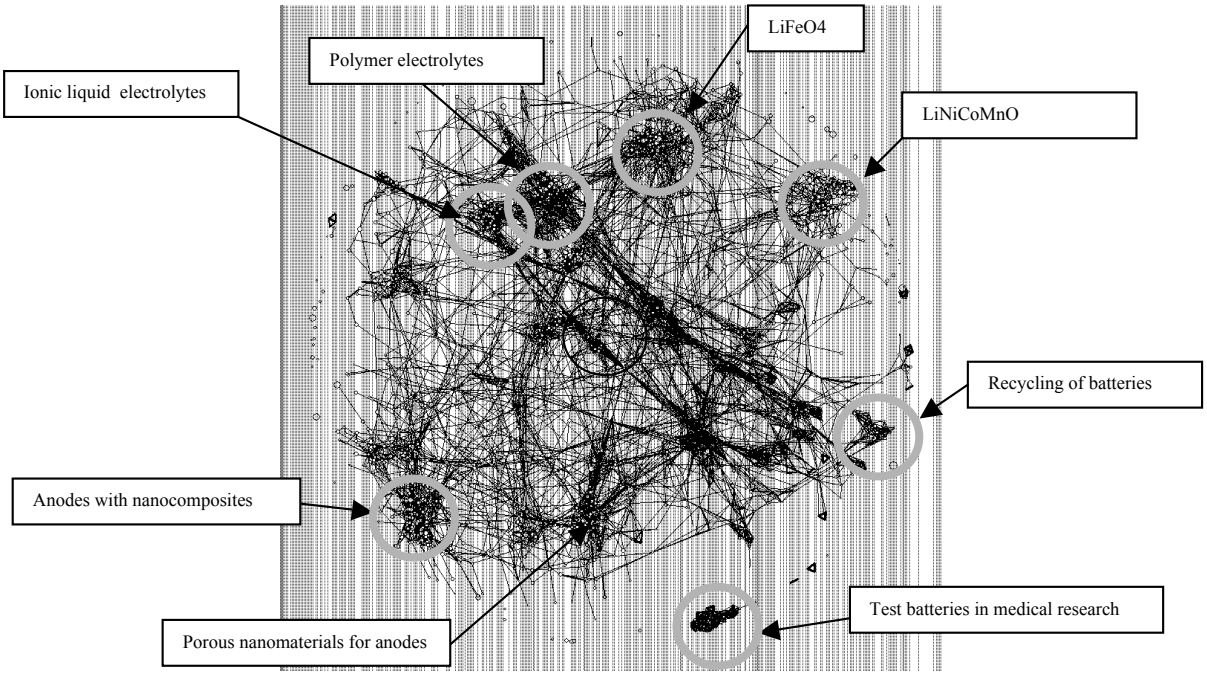

Figure 4. Map of bibliographically coupled publications; similar to Figure 1 but including only publications with at least one citation and references with less than 100 citations; similarity measure: number of common references normalised using the Jaccard index, projection by a spring model, threshold for Jaccard: 0.5; 1,826 publications, 2004 to 2010.

The results of this procedure are presented in Figure 4. The remaining 1,826 publications show a much more clearly structured coupling. We see more than 15 agglomerations of publications with different knowledge bases. This allows a more in-depth view into research activities: materials for electrodes like LiFePO4, LiNiMnO4, LiNiCoMnO, recycling of batteries, porous nanomaterials for anodes, nanocomposites for anodes, additives for stability, flame retardant additives and ionic liquid electrolytes.

The corresponding two dimensional density distribution of the average number of publications per area unit is shown in Figure 5. This illustrates the research fronts by the altitude (maximum of density) of the mountains.

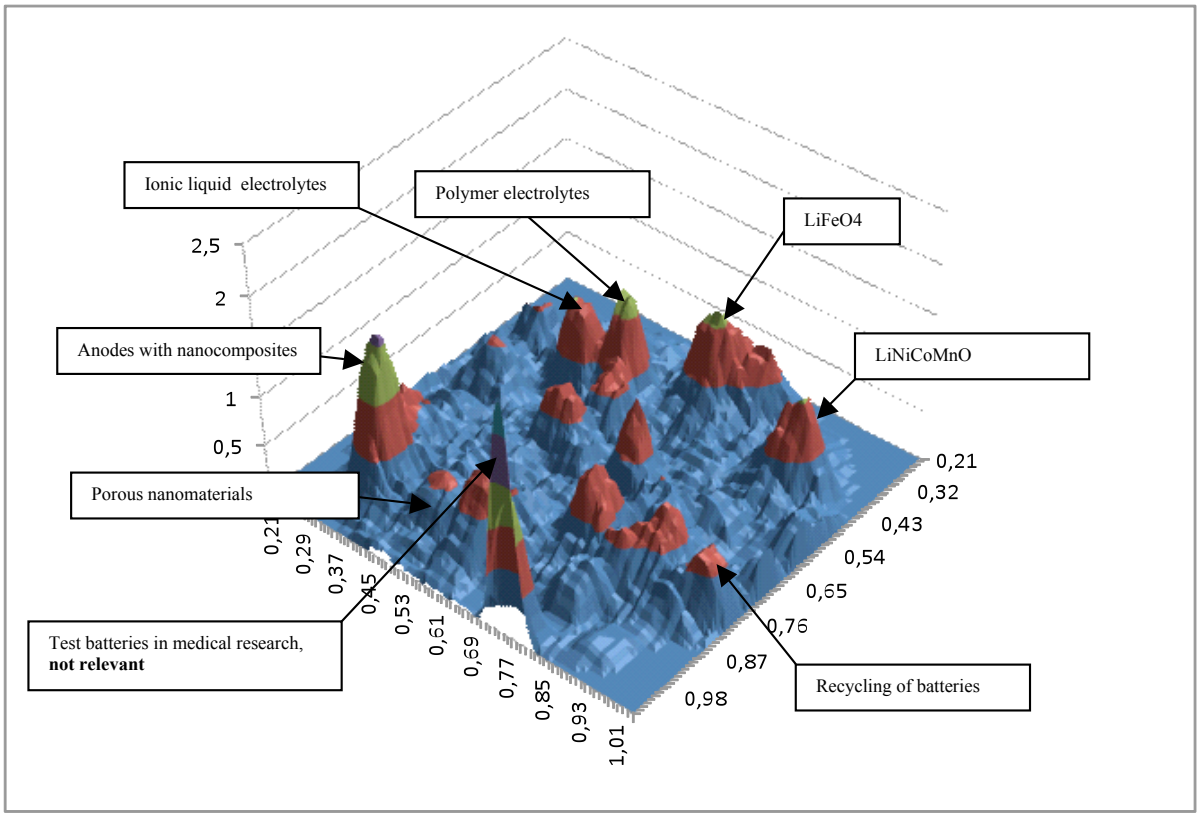

Figure 5. Two dimensional areal density distribution of bibliographically coupled publications, moving average of 7x7 pixels; similar to Figure 3 but including only publications with at least one citation and references with less than 100 citations; 1,826 publications, 2004 to 2010

Conclusions

The usage of co-word, co-citation or bibliographic coupling techniques has been reported in the literature to detect and visualize research fronts. The agglomeration of publications based on a common knowledge base that can be made visible by coupling publications on the basis of common references constitutes a practicable approach. While clustering methods or maps are already in use, studies have so far failed to provide the context of a whole research field and to quantify density distributions of research fronts. The proposed approach produces a map of a whole research field with thousands of publications in a relational context of bibliographic coupling.

The first step involves drawing a structured map of the whole field, thus revealing first agglomerations of publications with common knowledge bases. The quantification and visualization of the spatial density representing the number of publications per area unit opens a new perspective. Research fronts become visible as elevations in a three dimensional space. The altitude and thickness of mountains give an indication of the activity in a research front measured by the number of publications. Thickness and different peaks in a mountain range show thematic subtopics in a broader context such as use of different electrode materials.

This kind of analysis helps to monitor main topics of research interest in the scientific community. Long term planning activities in a research organisation can be reflected by the current state of research.

Further research will be performed by calculating lines of constant density. The number of publications within the area bounded by the lines indicates the dimension of the activity, while the extension of the area defines the broadness of the knowledge base. In a further step, the research fronts will be classified by age of cited references and year of publication.

References

Boyack & Klavans, (2010). Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach represents the Research Front Most Accurately?, JASIST 61(12): 2389-2404.

Chen, C., & Morris, S. (2003). Visualizing evolving networks: Minimum spanning trees versus pathfinder networks. Proceedings of IEEE Symposium on Information Visualization (pp 67-74), Seattle, WA: IEEE Computer Society Press

Chen, C (2006). CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. JASIST, 57(3): 359-377

Persson, O (1994). The Intellectual base and research fronts of JASSIS 1986-1990. JASSIS 45(1): 31-38

Price, D.D. (1965) Networks of scientific papers. Science, 149, 510-515

Shibata et al, (2009). Comparative Study on Methods of Detecting Research Fronts Using different Types of Citation. JASIST 60(3):571-580