

# Bibliometric field delineation with heat maps of bibliographically coupled publications using core documents and a cluster approach - the case of multiscale simulation and modelling (research in progress)

Edgar Schiebel<sup>1</sup>, András Vernes<sup>2</sup>; Friedrich Franek<sup>2</sup>

<sup>1</sup>AIT Austrian Institute of Technology GmbH, Donau City Straße 1, A-1220 Vienna, Austria  
[edgar.schiebel@ait.ac.at](mailto:edgar.schiebel@ait.ac.at)

<sup>2</sup>Austrian Center of Competence for Tribology, AC2T research GmbH, Viktor-Kaplan-Straße 2-C, A-2700 Wiener Neustadt, Austria

[vernes@ac2t.at](mailto:vernes@ac2t.at); [franek@ac2t.at](mailto:franek@ac2t.at);

*Contribution at the ECOST-MEETING-TD1210-290816-077233 "Identification, location and temporal evolution of topics - data and algorithm - comparison of approaches" 2016 08 29 – 2016 08 30*

Keywords: field delineation, bibliometrics, science mapping, cluster analysis, core documents, multiscale simulation and modelling

Acknowledgement: This work was funded by the Austrian COMET-Program (Project K2 XTribology, Grant No. 849109), and has been carried out within the Excellence Centre of Tribology.

A remarkable advance has been made in the development of bibliometric approaches for the identification and delineation of research topics, see for example Shibata, N., Kajikawa, Y., Takeda, Y. and Matsushima, K. (2009), Boyack, K. & Klavans, R. (2010), Glänzel, W. (2012). In recent years traditional bibliometric concepts like co-citation analysis, bibliographic coupling (Price, D.D., (1965), Persson, O., (1994), Kessler, M.M. (1963)), have been enriched with text and semantic similarities, topic modelling (Yau, C.,K., Porter, A., Newman, N., Suominen, A. (2014) ) as well as hybrid methods, (Janssens, F., Glänzel, W., & Moor, B. (2008)). Additionally visualization tools for networks and the representation of agglomerations of similar documents in a two dimensional space (Van Eck, N.J. and Waltman, L. (2010), Chen, C (2006), Kopcsa, A. & Schiebel, E. (1998)), dendrograms of hierarchies but also word clouds are helpful instruments that have been applied or developed. Well known algorithms are cluster analysis, multidimensional scaling with spring models. Recently latent Dirichlet allocation for topic modeling has been added.

However the identification of thematic issues or fields from thousands of publications is still a challenge. A reason is that some issues and after all emerging research topics very often do not have sharp borders to others. The reason is that publications that are thought to build delineated communities in a network of similar documents tend to be more or less strong connected to other communities that are located nearby. For example, in the case of bibliographic coupling often cited pioneering, influential and fundamental publications are responsible for cross linkages.

In this contribution a semi-automated graphically assisted procedure for the delineation of subfields is suggested. It is an attempt to combine several bibliometric approaches. The idea is to work with a reduced number of publications that are core documents having been selected by a minimum of the sum of the similarity to all other objects (Glänzel, W., & Thijs, B. (2011))., to map these documents in a two dimensional space with a spring model (Kopcsa, A., Schiebel, E., (1998)), to select documents with

a high centrality in a hot zone of an agglomeration of similar documents and to enlarge this core set with documents that have a minimum similarity to each element of the core. That way, documents have been agglomerated to a hot zone by the so called second order similarity but the selection of subsets is done with the help individual similarities (Colliander & Ahlgreen, 2012; Thijs, B., Schiebel, E. & Glänzel, W., 2013). In a next step the coupling elements of a these selected subsets (cited references) are visualized on the map and demonstrate the spreading of the underlying knowledge base of a selected and delineated research front (Schiebel, E. (2015)).

The names of the research fronts are inspired by the most relevant keywords, the content of the cited documents and the content of the documents of the research front.

The resulting research fronts are compared with the results of a cluster analysis in a concordance matrix. The comparison of the clusters with the research fronts gives an additionally view on the subfields to be delineated. Subsets of documents that form a strong and delineated community can be identified identically as a research front from the map but also as a well separated cluster. Other research fronts can be subdivided in compact sub fields.

The following procedure is proposed and presented for discussion in the workshop:

- Collection of a set of publications of research on multiscale simulation and modeling (MSSM) downloaded from Web of Science.
- Calculation of sum similarities (Jaccard index) of bibliographically coupled publications
- Selection of a subset of publications with a threshold of the sum similarity
- Calculation and visualization of a two dimensional map of the subset of bibliographically coupled publications with a spring model.
- Filter and visualize the local density of the number of publications weighted with the similarity to draw agglomerations of publications with a heat map.
- Graphically assisted selection of documents in the center of a red heat zone
- Selection of additional elements of the community by thresholds of similarity (Jaccard index, number of common references)
- Visualization of the coupling elements (cited references) to examine the compactness of the selected community
- Providing lists of TFIDF ranked keywords, authors, organizations, cited references and abstracts of the publications of the community to name the research fronts
- Calculation of a cluster analysis (Pearson, Ward or other similarities and linkage methods) of the subset of bibliographically coupled publications. Visualization with a circular dendrogram.
- Visualization of selected clusters in the heat map
- Comparison of the identified communities of both approaches (Concordance matrix)
- Conclusions (exclusion of cited reviews, highly cited fundamental publications, preprocessing of the set of publications by excluding publications with weak second order sum similarities, workload,...)

Exemplary results:

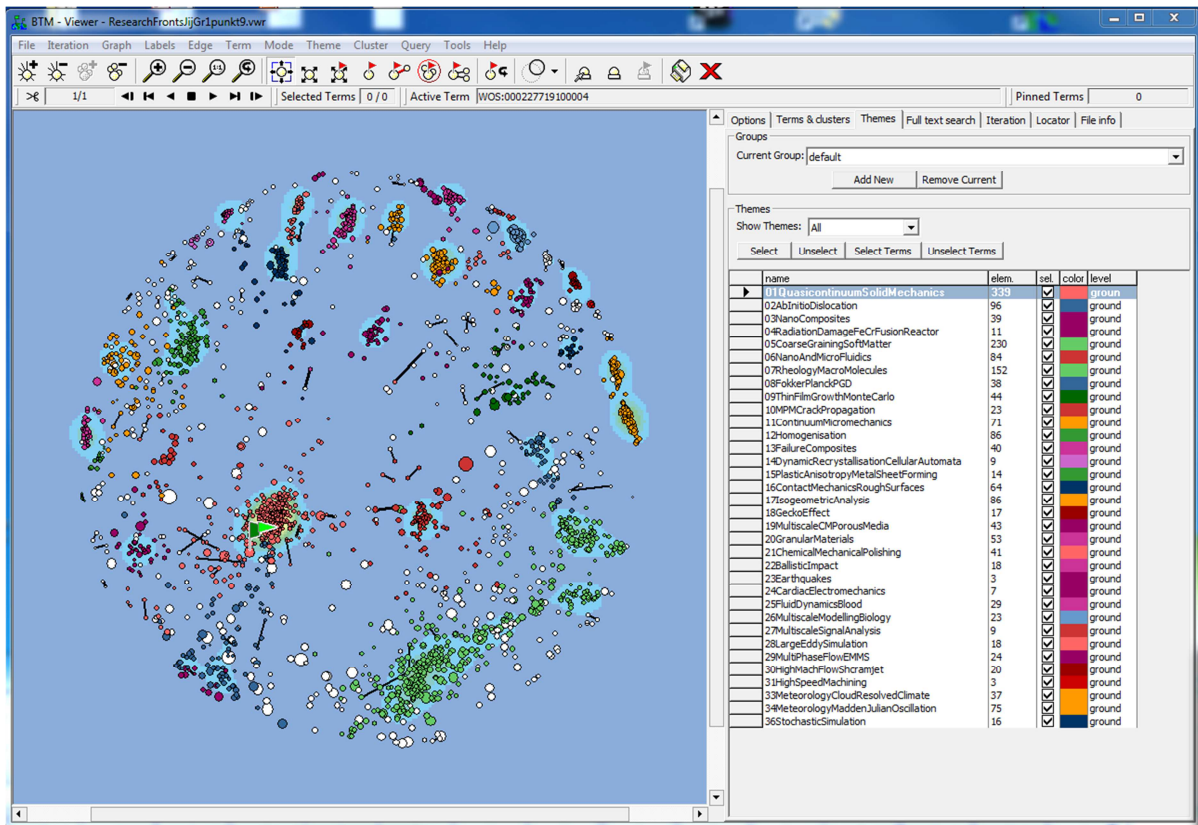


Figure 1: Heat map of the agglomeration of bibliographically coupled MSSM publications (bubbles), spring model, colored red heat zones (local density of documents weighted by sum similarity), reduced number of publications (2325 out of 8145, threshold for the sum of Jaccard indices for each node: 1.9), publications of a research fronts are marked with the color of the research front, data retrieved October 20, 2014

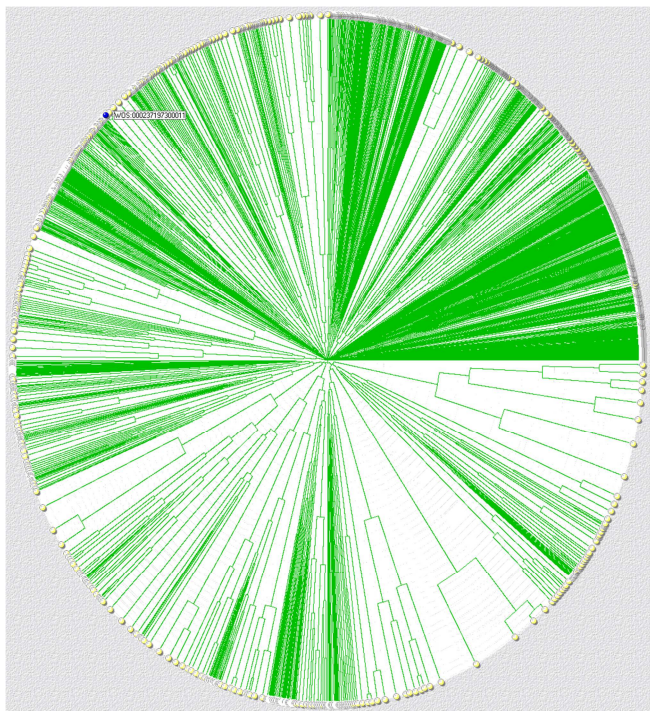


Figure 2: Cluster tree of bibliographically coupled publications: Pearson Correlation of second order Jaccard similarity, Ward Linkage



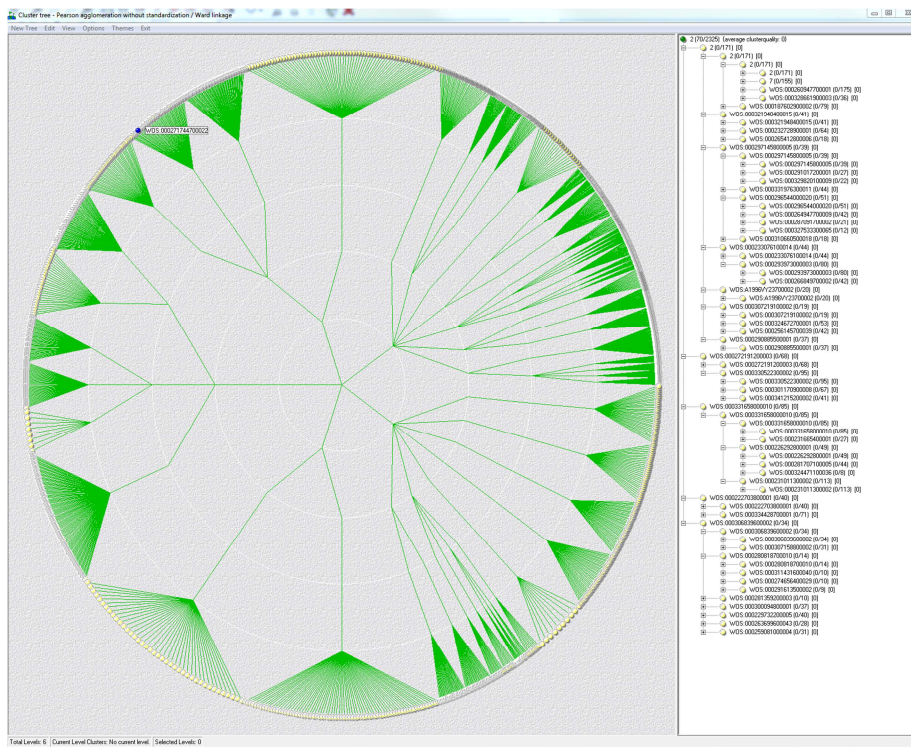


Figure 3: Cut circular cluster tree of research fronts with 5 equidistant clusters levels with 5/19/36/50 clusters

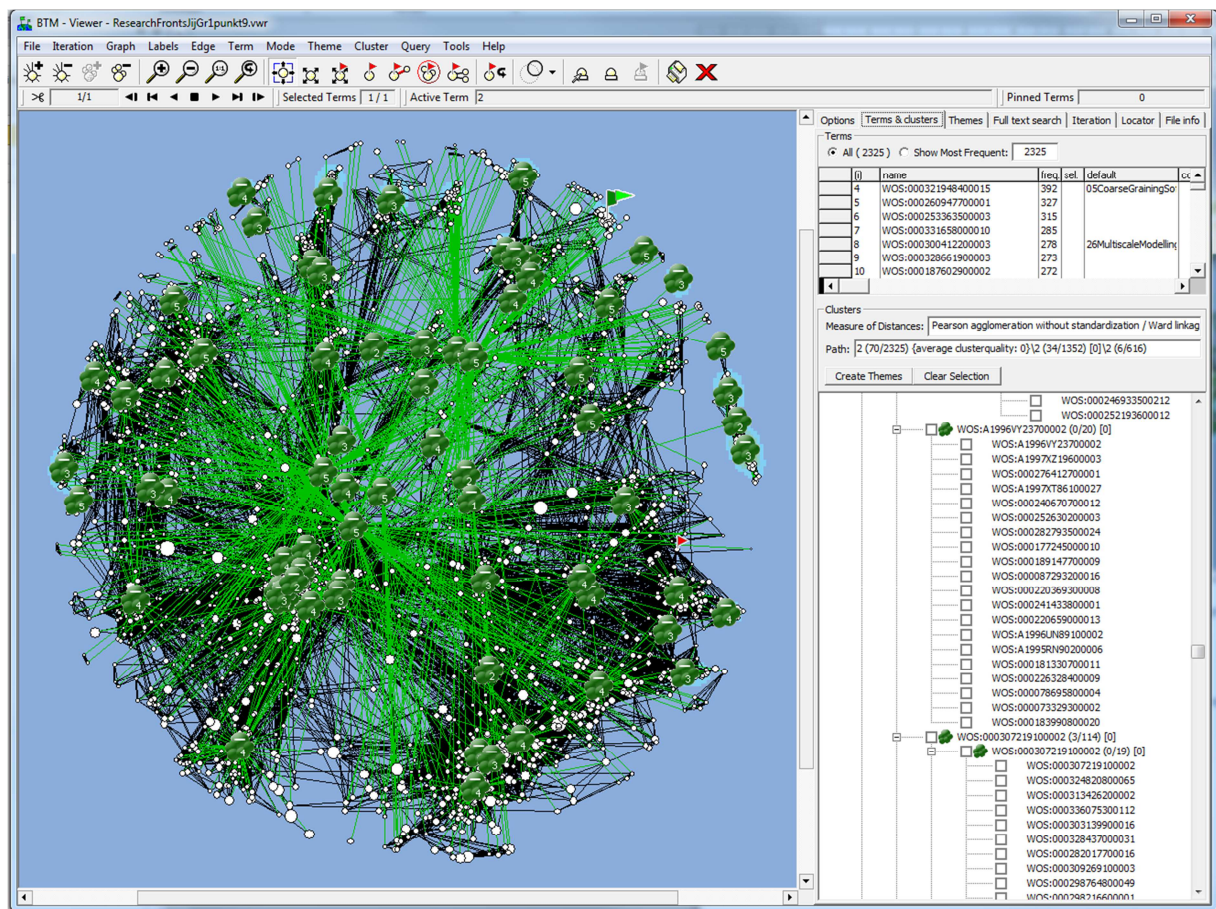


Figure 4: Visualization of the Cluster structure in the heat map of Figure 1, black lines are edges of the similarities and green lines indicate the cluster category



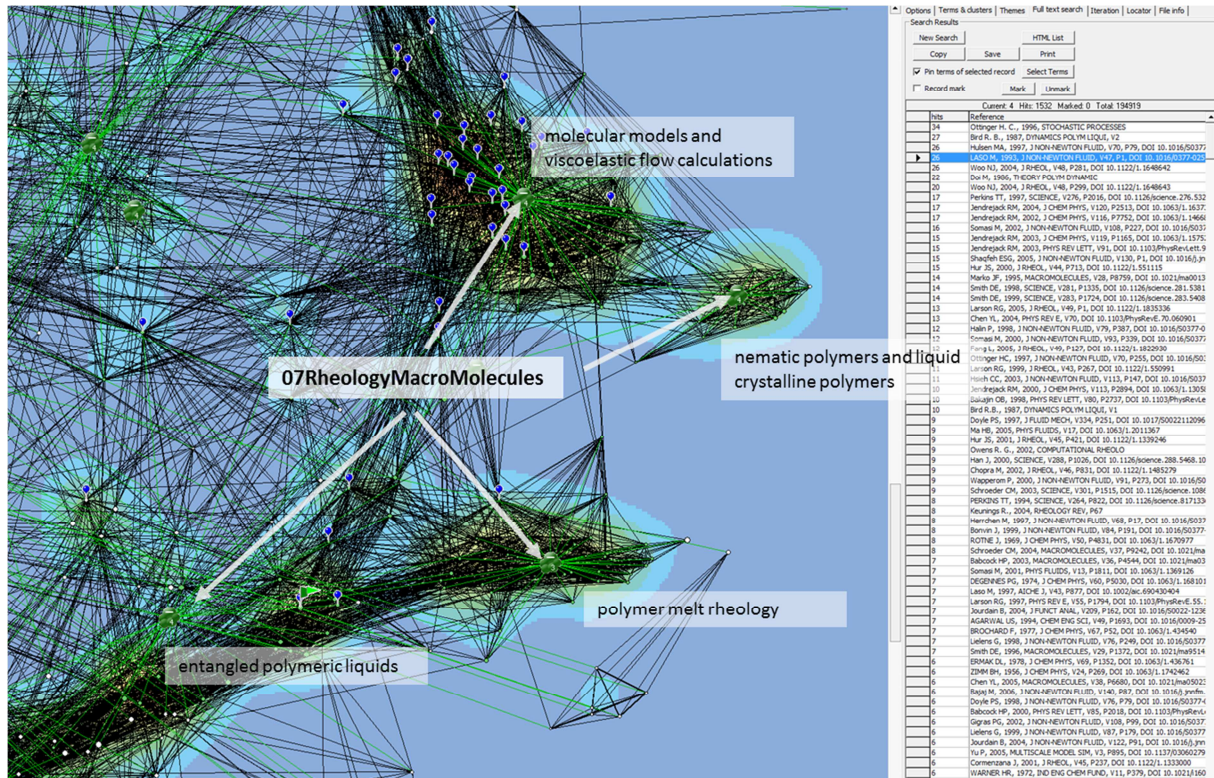


Figure 5: Cluster structure in the heat map of Figure 1 for the research front 07Rheology MacroMolecules



## References

- Boyack, K. & Klavans, R., (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., (2006) CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. *JASIST*, 57(3): 359-377
- Colliander, C., & Ahlgren, P., (2012) Experimental comparison of first and second-order similarities in a scientometric context. *Scientometrics*, 90(2), 675-685. doi:10.1007/s11192-011-0491-x
- Glänzel, W., & Thijs, B., (2011) Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88(1), 297-309. doi:10.1007/s11192-011-0347-4
- Glänzel, W., (2012) Bibliometric Methods for Detecting and Analysing Emerging Research Topics. *PROFESIONAL DE LA INFORMACION*, 21(2), 194–201. doi:10.3145/epi.2012.mar.11
- Janssens, F., Glänzel, W. & Moor, B., (2008) A hybrid mapping of information science. *Scientometrics*, 75(3), 607–631. doi:10.1007/s11192-007-2002-7
- Kopcsa, A. & Schiebel, E., (1998) Science and technology mapping: A new iteration model for representing multidimensional relationships. *Journal of the American Society for Information Science* 49 (1), 7–17, doi:10.1002/(SICI)1097-4571(1998)49:1<7:AID-ASI3>3.0.CO;2-W.
- 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, N., Kajikawa, Y., Takeda, Y. and Matsushima, K. (2009) Comparative study on methods of detecting research fronts using different types of citation. *JASSIS*, 60: 571–580. doi:10.1002/asi.20994
- Schiebel, E. (2015) Mapping the Spreading of Cited References over Research Fronts of Bibliographically Coupled Publications; in: *Proceedings of the 14th International Symposium on Information Science (ISI)*", F. Pehar, C. Schögl, C. Wolff (ed.); Verlag Werner Hülsbusch, Glückstadt (2015), ISBN: 978-3-940317-81-0; 404 - 409.
- Thijs, B., Schiebel, E., & Glänzel, W. (2013) Do second-order similarities provide added-value in a hybrid approach? *Scientometrics*, 96(3), 667-677. doi:10.1007/s11192-012-0896-1
- Van Eck, N.J. and Waltman, L. (2010) Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538
- Yau, C.,K., Porter, A., Newman, N., Suominen, A. (2014) Clustering scientific documents with topic modeling, *Scientometrics* 100: 767. doi:10.1007/s11192-014-1321-8