Using Hybrid Methods and 'Core Documents' for the Representation of Clusters and Topics: The Astronomy Dataset

Wolfgang Glänzel[1,2] and Bart Thijs[1]

[1] *wolfgang.glanzel@kuleuven.be, bart.thijs@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
[2]Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
Based on a dataset on Astronomy & Astrophysics a hybrid cluster analysis has been conducted. Hybrid clustering was based on a combination of bibliographic coupling and textual similarities using Louvain method at two resolution levels. The procedure resulted in seven and thirteen clusters, respectively. The statistics reflect a high quality of classification. For labelling and interpreting clusters, *core documents* are used. The results of these two scenarios are presented, discussed and compared with each other. The two scenarios clearly result in hierarchical structures that are analysed with the help of a concordance table. Furthermore, the core documents help depict the internal structure of the complete network and the clusters.
This work has been done as part of the international project 'Measuring the Diversity of Research' and in the framework a special workshop on the comparative analysis of algorithms for the identification of topics in science organised in Berlin in August 2014.

Conference Topic
Methods and techniques (special session on algorithms for topic detection)

Introduction
Within the framework of the event series on 'Measuring the Diversity of Research' a special workshop on the comparative analysis of algorithms for the identification of topics in science was organised in Berlin in August 2014. A dataset downloaded from Thomson Reuters Web of Science covering the annual volumes 2003–2010 was shared with all contributors in order to test the various algorithms and techniques and to compare the results of the different approaches. On the basis of the shared Astronomy & Astrophysics dataset the following analysis has been conducted at our institute. In particular, the topic structure of the subject defined by the set was analysed using two different but related techniques. A cluster analysis was based on bibliographic coupling and textual similarity. And *core documents* (Glänzel & Czerwon, 1996) defined on the same links were used to represent topics within the subject and to depict the internal structures of both subject and clusters (cf. Glänzel & Thijs, 2011). Main results are presented in the following, but changing parameters of the algorithm and of the combination of the components leads to further results.
Currently a new and more robust method for the measurement of textual similarities and thus for the revision of the lexical component is in development. A comparison of the results of the present study with those of the new algorithm is part of the ongoing project and will be presented on a later occasion, when available.

Methodological aspects
The advantage of using hybrid lexical–citation based methods, notably of combinations of term-frequency and bibliographic coupling, has already been discussed in previous studies (e.g., Glenisson et al., 2005; Boyack & Klavans, 2010). However, at this level of aggregation (topics within the same field or discipline) we have encountered several specific problems that have already been reported in earlier studies in the context of the detection of emerging topics (e.g., Glänzel & Thijs, 2012). Terms and phrases might become less specific since they express common knowledge base and vocabulary while others might gain more 'information

value'. The most important TF-IDF keywords and terms alone are often not specific enough for topic description and labelling. Thus a larger set of terms is needed to describe topics at this level. A possible solution has already be discussed already in earlier studies (e.g., Glänzel & This, 2011): On one hand, depending on the level of aggregation *and* the discipline under study, the weight of the two components can be adjusted and, on the other hand, instead of the best TF-IDF terms *core documents* can be used to describe and label clusters. In order to apply the hybrid clustering we have only vertices with *positive* degree (i.e., documents with at least one link) taken into account. Furthermore, we have removed all papers with publication years outside the period 2003–2010. Table 1 shows the description of the dataset.

Table 1. The input dataset.
[Data sourced from Thomson Reuters Web of Science Core Collection]

Data	Documents	Percentage
Original dataset	111514	100.00
Not present in ECOOM Database	103	0.09
Publications in 2003-2010	110412	99.01
Excluded from all analysis	1205	1.08

We applied Louvain method (Blondel et al., 2008) using Pajek (Batagelj & Mrvar, 2003) to this dataset. The reason for this choice was that hierarchical clustering with Ward used in previous projects (e.g., Thijs et al., 2013) often results in a heterogeneous "hotchpotch" cluster of objects that can otherwise not be assigned. Therefore we decided to apply Louvain method. We conducted a hybrid clustering with two components: *bibliographic coupling* (BC) and *textual similarity* (TS), where we used a weight of 0.75 for BC and 0.25 for TS according to the algorithm described in Glänzel & Thijs (2011). In particular, the underlying similarity measure r is defined as the cosine of the linear combination of the underlying angles between the vectors representing the corresponding documents in the vector space model, i.e.,

$$r = \cos\left(\lambda \cdot \arccos(\eta) + (1-\lambda) \cdot \arccos(\xi)\right), \quad \lambda \in [0,1],$$

where η is the similarity defined on bibliographic coupling and ξ the textual similarity. The λ parameter defines the convex combination, $\arccos(\eta)$ and $\arccos(\xi)$, respectively, denote the two underlying angles. Furthermore, we have conducted the clustering at two resolution levels, namely 0.7 and 1.4. The results of these two scenarios will be presented and briefly discussed in the following section.

Results

The results using both resolution levels are briefly summarised in Table 2. The number of documents, that could not been clustered, is marginal. The number of clusters has almost doubled (from 7 to 13) with growing resolution. The solutions for the two resolution levels are presented in Tables 3 and 4. Except for the tiny cluster (#13) on atmospheric turbulence in the second solution, all clusters are of reasonable size. This is expressed by the frequency, i.e., the number of documents per cluster (columns 2–4). The description of the clusters, shown in the last column of the tables, have been derived from the most important TF-IDF terms and the titles of the *core documents*, where the core documents have been determined according to see Glänzel (2012) on the basis of the *degree h-index* of the hybrid document network. In particular, core documents are represented by core nodes, which, in turn, are defined as nodes with at least h degrees each, where h is the h-index of the underlying graph. Or, to express this simpler, degrees of documents are ranked in descending order and the h-core is formed by the documents the degrees of which do not undercut their rank value. This method has proved

efficient in *local* clustering, that is, in clustering of fields or disciplines, where the network h-core usually represents the order of magnitude of 1% of the total document set (see Glänzel, 2012).

Table 2. Description of parameters and results. [Data sourced from Thomson Reuters Web of Science Core Collection].

Number of vertices		108937
Number of edges		87602281
Density		1.5%
All Degree Centralization		0.13
Method		Louvain (Pajek)
Hybridity parameter		$\lambda = 0.75$
Resolution	0.7	1.4
Number of Clusters	7	15
Documents not Clustered	360	360
Modularity	0.61	0.49

Table 3. Scenario 1 (description of structures in the seven-cluster structure). [Data sourced from Thomson Reuters Web of Science Core Collection].

Cluster	Freq	Freq%	CumFreq	CumFreq%	Label
1	20634	18.7%	20634	18.7%	Star Clusters
2	12149	11.0%	32783	29.7%	Terrestrial planets/Extra Solar Planets
3	14365	13.0%	47148	42.7%	Solar Flares
4	17036	15.4%	64184	58.2%	Star Formation
5	20173	18.3%	84357	76.5%	Dark Energy
6	15023	13.6%	99380	90.1%	Gamma Ray Burst
7	10820	9.8%	110200	99.9%	Neutrino

Table 4. Scenario 2 (description of structures in the 13-cluster structure). [Data sourced from Thomson Reuters Web of Science Core Collection].

Cluster	Freq	Freq%	CumFreq	CumFreq%	Label
1	11569	10.5%	11569	10.5%	Star Clusters / Globular Clusters
2	9470	8.6%	21039	19.1%	Disk around a brown dwarf or young star
3	12163	11.0%	33202	30.1%	Extrasolar planetary systems
4	15060	13.7%	48262	43.8%	Solar Flares
5	6481	5.9%	54743	49.6%	Dark Matter Halo: Formation of galaxies
6	10075	9.1%	64818	58.8%	Star formation
7	7523	6.8%	72341	65.6%	Dark Energy
8	9005	8.2%	81346	73.8%	Astrophysical jets and accretion discs
9	10298	9.3%	91644	83.1%	Brane-world black hole
10	5503	5.0%	97147	88.1%	Radio Pulsars
11	2336	2.1%	99483	90.2%	Gamma Ray Burst
12	10224	9.3%	109707	99.5%	Neutrino
13	477	0.4%	110184	99.9%	Atmospheric turbulence

Table 5. Core-document representation of Cluster #5 based on h-core. [Data sourced from Thomson Reuters Web of Science Core Collection].

UT	Degree	Rank	Title
000261696000006	111	1	Non-linear isocurvature perturbations and non-Gaussianities
000278201600003	99	2	Non-Gaussianity of quantum fields during inflation
000260529800008	96	3	Conditions for large non-Gaussianity in two-field slow-roll inflation
000261260200020	88	4	A curvaton with a polynomial potential
000278201600004	86	5	Local non-Gaussianity from inflation
000238060100019	84	6	Non-Gaussianities in two-field inflation
000186983100013	83	7	Generalized chaplygin gas with alpha=0 and the Lambda CDM cosmological model
000246571300004	82	8	Cleaned 3 year Wilkinson Microwave Anistropy Probe cosmic microwave background map: Magnitude of the quadrupole and alignment of large-scale modes
000253980700030	82	9	Non-Gaussianity analysis on local morphological measures of WMAP data
000276102300001	81	10	Scale dependence of local f(NL)
000270036800016	79	11	Non-Gaussianity beyond slow roll in multi-field inflation
000235669800017	78	12	Testing primordial non-Gaussianity in CMB anisotropies
000259692800055	77	13	Anomalous CMB North-South asymmetry
000185760100005	76	14	WMAP and the generalized Chaplygin gas
000250363000004	75	15	Alignment and signed-intensity anomalies in Wilkinson Microwave Anisotropy Probe data
000221258900057	74	16	Numerical analysis of quasinormal modes in nearly extremal Schwarzschild-de Sitter spacetimes
000264762500065	74	17	Modeling gravitational recoil from precessing highly spinning unequal-mass black-hole binaries
000220092300012	73	18	Non-Gaussianity in the curvaton scenario
000242409800004	72	19	Non-Gaussianity of the primordial perturbation in the curvaton model
000242449600008	72	20	A numerical study of non-Gaussianity in the curvaton scenario
000245928000021	70	21	Exploring the properties of dark energy using type-Ia supernovae and other datasets
000248953800006	70	22	Primordial non-Gaussianity in multi-scalar slow-roll inflation
000253764800075	70	23	Further insight into gravitational recoil
000243171800001	68	24	Inflationary trispectrum for models with large non-Gaussianities
000252864000020	68	25	Non-Gaussianity in the modulated reheating scenario
000243171800040	67	26	Primordial trispectrum from inflation
000275514800001	67	27	Disks in the sky: A reassessment of the WMAP ""cold spot""
000278201600005	67	28	Use of delta N formalism-difficulties in generating large local-type non-Gaussianity during inflation
000221258900023	66	29	Curvature and isocurvature perturbations in a three-fluid model of curvaton decay
000221277400044	66	30	Dirac quasinormal modes of the Reissner-Nordstrom de Sitter black hole
000266501900050	66	31	Trispectrum versus bispectrum in single-field inflation
000272271900003	66	32	The subdominant curvaton
000243725400002	65	33	The non-Gaussian cold spot in the 3 year Wilkinson Microwave Anisotropy Probe data
000244679500013	65	34	Mapping the large-scale anisotropy in the WMAP data
000255424300029	65	35	Generation and characterization of large non-Gaussianities in single field inflation
000235939700023	64	36	On the large-angle anomalies of the microwave sky
000250954900032	64	37	A note on the large-angle anisotropies in the WMAP cut-sky maps
000257290600085	64	38	Anti-de Sitter universe dynamics in loop quantum cosmology
000245405900001	63	39	Constraints on the generalized Chaplygin gas model from recent supernova data and baryonic acoustic oscillations
000183377200050	62	40	Generation of dark radiation in the bulk inflaton model
000188864800011	62	41	Large scale structure and the generalized Chaplygin gas as dark energy
...	...	...	...
000256378700020	60	59	A low cosmic microwave background variance in the Wilkinson Microwave Anisotropy Probe data
000259700200011	60	60	Consistency relations for non-Gaussianity

Table 5 lists the core documents of Cluster #5 of the first scenario with seven clusters as an example. The degrees given in the table also illustrates the role of core documents in the cluster: Core documents are by definition strongly interlinked with many other documents and therefore play a representative and central part in a network. And they are suited to depict the internal structure of the complete network, of a cluster or of parts of it. In this context Cluster #5 has not been chosen by chance. The core documents of this cluster form the centre of the structure. Links connecting core documents reveal the internal structure of both the field under study and the clusters as the links with other core documents of the same cluster as well as with those of other clusters are distinctly apparent. Beside this cluster, also cores documents of cluster 7 play a central part. This is shown in Figure 1. Core documents of cluster 5 are marked in pink, those of Cluster 7 in auburn.

By contrast, Figure 2 presents the concordance between the two scenarios. Indeed the two resolutions results in a different number of clusters as already have been shown in Tables 3 and 4. Now the question arises of whether the two approaches yield completely different structures or almost concordant hierarchic structures, where the choice of the resolution would go with merging and splitting clusters, respectively. The first case would, of course, be problematic and point to the possible inappropriateness of methodology, while latter case testifies consistency of the chosen method. Cluster concordance of the results of the two scenarios are visualised in Figure 2.

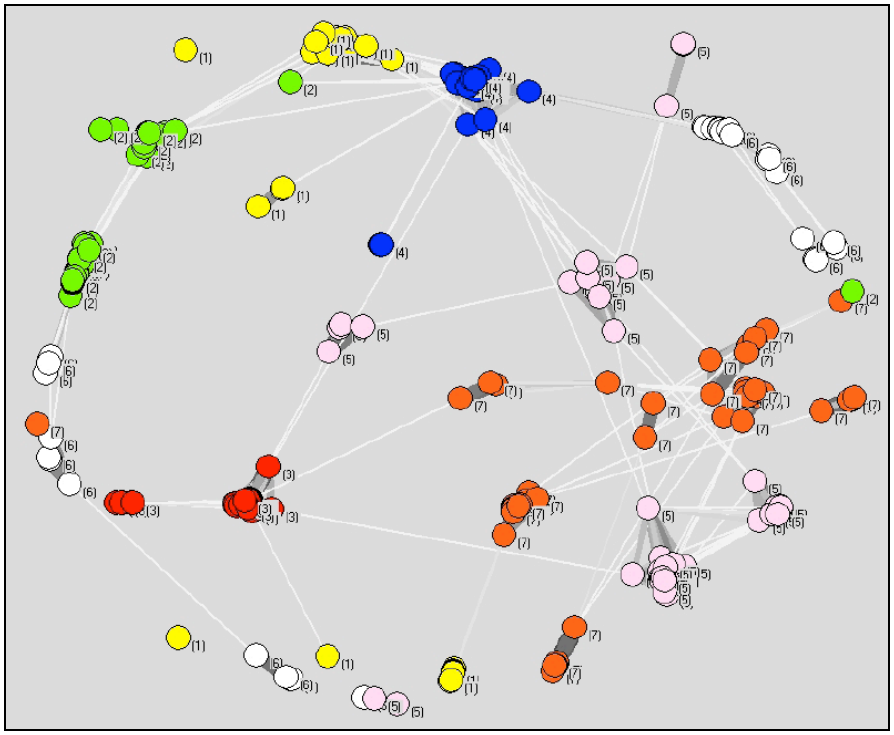

Figure 1. Structure of core documents in 7 clusters according to scenario 1 (Pajek with Fruchterman-Rheingold layout) [Data sourced from Thomson Reuters Web of Science Core Collection].

Hybrid Approach 13 Clusters \ Hybrid Approach 7 Clusters	1	2	3	4	5	6	7
1	94	0	0	5	0	1	0
2	96	1	0	2	0	1	0
3	1	95	3	1	0	0	0
4	1	0	91	0	0	7	0
5	1	0	0	86	2	1	10
6	0	0	0	100	0	0	0
7	0	0	0	2	98	0	0
8	2	0	0	4	0	94	0
9	0	0	0	0	97	0	2
10	1	3	1	1	42	51	0
11	1	2	0	1	0	95	0
12	0	0	0	1	2	1	95
13	5	61	24	0	2	0	7

Figure 2. Cluster concordance: scenario 1 – scenario 2 (overlap in %). [Data sourced from Thomson Reuters Web of Science Core Collection]

The document overlap in the corresponding clusters is expressed in per cents and, in order to facilitate interpretation, marked in different colours. Percentages sum up to 100% by rows. If one neglects the light-weight Cluster #13 in the second scenario, which actually represents just 0.4% of the total, one observes an almost perfect concordance of three clusters in scenarios 1 and 2 (#2 = #3, #3 = #4 and #7 = #12), one cluster splits up into two others (#4 = #5+#6) and finally two clusters split up into three clusters each, namely #5 = #7+#9+#10 and #6 = #8+#10+#11. Thus Cluster #10 in scenario 2 is the only one that breaches the strict hierarchy in the structures of the two scenarios. Its documents are almost equally distributed over Clusters #5 and #6 in scenario 1. The tiny one (#13) in the second

scenario can be considered a small sub-cluster of #2 in the first one, where it represents just slightly more than 2% of the documents of the total cluster.

Conclusions

Our main conclusions refer to two issues, firstly to the *clustering results* and secondly to the role of *core documents*. As to the clustering, both scenarios resulted in an almost perfect hierarchic structure. Cluster concordance and hierarchy was strong except for the cluster on 'Radio Pulsars' in the 13-cluster solution. This cluster was almost evenly spread over the clusters on 'Dark Energy' and 'Gamma Ray Burst' in the seven-cluster solution. Nevertheless, hierarchical assignment of 'Atmospheric Turbulence' in scenario 2 was also somewhat "fuzzy", but had a main concordance of more than 60% of documents with 'Coronal Loop' in the first scenario. In all other cases concordances were around or even above 90% document overlap.

The second group of remarkable observations refer to core documents. These documents represent the links across clusters as well as the internal topic structure of the clusters. In this context we have to repeat that core-document identification is in principle *independent* of clustering and thus does not require any cluster analysis or community detection, but it can be seamlessly integrated into clustering exercises, provided the same type of links, i.e., bibliographic coupling, co-citation, text similarity or hybrid, are used. Core documents reinforce the observation concerning centric results of the hybrid clustering. Core documents of the clusters on 'Dark Energy' and 'Neutrino' actually form the centre of the structure. The choice of the two resolution levels resulted in a hierarchic structure confirming the appropriateness of the applied method.

Acknowledgements

This work has been done as part of the international project 'Measuring the Diversity of Research' and in the framework a special workshop on the comparative analysis of algorithms for the identification of topics in science organised in Berlin in August 2014. The project and workshop series was jointly organised by the Humboldt Universität and Technische Universität Berlin. We would like to acknowledge their support of our study.

References

Batagelj, V. & Mrvar, A. (2003). *Pajek–Analysis and visualization of large networks*. In: M. Jünger & P. Mutzel (Eds.), Graph drawing software (pp. 77–103). Berlin: Springer.

Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008.

Boyack, K. W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404.

Glänzel, W. & Czerwon, H.J. (1996), A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, *37*(2), 195–221.

Glänzel, W. & Thijs, B. (2011), Using `core documents' for the representation of clusters and topics. *Scientometrics*, *88*(1), 297–309.

Glänzel, W. & Thijs, B. (2012), Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics*, *91*(2), 399–416.

Glänzel, W. (2012), The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics*, *93*(1), 113–123.

Thijs, B., Schiebel, E. & Glänzel, W. (2013), Do second-order similarities provide added-value in a hybrid approach? *Scientometrics*, *96*(3), 667–677.