Using noun phrases extraction for the improvement of hybrid clustering with text- and citation-based components. The example of "Information System Research"

Bart Thijs[1], Wolfgang Glänzel[2], and Martin Meyer[3]

[1] *bart.thijs@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)

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

[3] *m.s.meyer@kent.ac.uk*
Kent Business School, University of Kent, Canterbury(UK)
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
SC-Research, University of Vaasa, Lapua, (Finland)

Abstract
The hybrid clustering approach combining lexical and link-based similarities suffered for a long time from the different properties of the underlying networks. We propose a method based on noun phrase extraction using natural language processing to improve the measurement of the lexical component. Term shingles of different length are created form each of the extracted noun phrases. Hybrid networks are built based on weighted combination of the two types of similarities with seven different weights. We conclude that removing all single term shingles provides the best results at the level of computational feasibility, comparability with bibliographic coupling and also in a community detection application.

Workshop Topic
Text enhanced bibliographic coupling

Introduction
For a long time scientometrians have been using the combination of textual analyses with citation based links for many different applications. In 1991, Braam et al. (1991a; 1991b) suggested the use of co-citation in combination with word-profiles which are indexing terms and classification codes for a mapping of science. In the same year, Callon et al. (1991) demonstrated how co-word analysis can be used for studying academic and technological research. Glenisson (2005) encountered the disadvantage of the single term approach and used the Dunning likelihood ratio test (Dunning 1993; Manning & Schütze, 2000) to identify common bigrams. For this test the occurrence of each pair has to be calculated together with the frequency of each term appearing separately. The bigrams with the highest score are retained. The risk of this procedure is that pairs that are less frequent or that appear in a few variations are not selected. Also the selection of a bigram in a paper might change when additional documents are added to the dataset. It is clear that the introduction of full text analysis increased processing complexity. Janssens (2005) introduced a true integrated approach where he combines the distance based on bibliometric features with a text-based distance using a linear combination of distances (or similarities), where the parameter can be used to fine-tune the weight of the two components. Later Janssens et al. (2008) warned against the combination based on simple vector concatenation and linear combinations of similarity measures because of the completely different structures of the underlying vector spaces and they proposed a combination based on Fisher's inverse Chi-Square. They also showed that this method outperforms hitherto applied methods. This method solves the issue of different distributions

drastically but it introduces an even more complex calculation scheme. Glänzel & Thijs (2012) take a more pragmatic approach and exploit the fact that both similarities can be expressed as cosines in a vector space model and introduce a hybrid similarity as the cosine of the weighted linear combination of the underlying angles of each of the cosine similarities.
None of solutions proposed in the literature were so far able to eliminate or at least to considerably reduce the effect of different distributions in each of the two components without excessive computational requirements.
In this paper we introduce the use of noun phrase extracted by the application of *Natural Language Processing* (NLP) and we investigate different options that can be taken while using syntactical parsing and the effects of these choices on the lexical similarities and the properties of networks based on these similarities. The rationale here is that as we are using the text mining to map documents in order to identify clusters of fields or emerging topics we have to limit the textual information that we use to those elements in texts (or - more formally - those parts of speech) that actually contain the topics. Nouns or noun phrases are used as subjects, objects, predicative expressions or prepositions in sentences. Syntactic parsing as one of the applications within NLP will be used to extract the noun phrases from the abstracts; other categories, such as verb, adjective or adpositional phrases will be neglected. However, the selected noun phrase might contain an embedded phrase of these other types or even other embedded noun phrases. We will illustrate the new approach using the example of a document set on Information System Research.

Data source and data processing

A set of 6144 publications on 'Information Systems' is used in this study. This data set is retrieved from the *Social Sciences Citation Index* by using a custom developed search strategy focusing on 'Management Information System', 'Geographical Information System', 'Decision Support System' or 'Transaction Processing System' (Meyer et al., 2013). Publications from 1991 up to 2012 with document type Article, Letter, Note or Review are selected.
For the lexical component, the title and the abstract of the papers are processed by both Lucene[1] (version 4.0) and the Stanford Parser. Terms used in the older single term based approach were retrieved by the next pre-processing steps: title and abstracts are merged and converted to lower case. Then, this data is tokenized by punctuation and white spaces. Stop words are removed through a custom built stop word list and remaining terms were stemmed by the Snowball Stemmer available in Lucene which is an extended version of the original Porter Stemmer (Porter, 1980). All terms that occur in only one document are removed. A term-by-document matrix is constructed in a vector space model with term frequency-inverse document frequency weightings (TF-IDF). Salton's cosine measure is used as measure of document similarity (Salton & McGill, 1986).
For the extraction noun-phrases we rely on the Stanford Parser, a Java package which has been developed and distributed by the Stanford Natural Language Processing Group. In short, this parser returns the grammatical structure of sentences based on probabilistic language models. In this study we use the PCFG-parser version 2.0.5 (Klein & Manning, 2003). The format of the output of the parser are *Stanford Dependencies*, which describe the grammatical relations between words in a sentence (de Marneffe & Manning, 2008a; 2008b). In the output, nouns are tagged with NN or NNS (for plurals), noun phrases with NP. For the selection of the noun phrases from the parsing result we can choose between several options. Complete noun phrases (NP) can be selected or only restricted noun phrases in which no other noun phrase is embedded. Noun phrases can be recorded with the constituent words in the given order or the included

[1] See http://lucene.apache.org, visited in January 2015

terms can be sorted alphabetically. It is the objective of this paper to study the consequences of these options.
After selection of the type of noun phrases and the optional sorting of the terms additional processing steps are taken. Similar to the single-term based approach the Snowball stemmer is applied and stop words are removed. The stemmed terms within a single phrase are then used to create term-shingles. A term shingle is a set of subsequent terms. The length of these shingles can vary between one and the number of terms in the phrase which is the maximum. With respect to the length of the selected shingles we identified five different possibilities with different criteria on the number of terms in the phrase and on the length of the shingle. Table 1 lists the five applicable criteria on the length of the shingle.

Table 1. All possible shingles in a phrase with three terms

Tag	Criteria
(none)	None – all possible shingles are included
`Lm`	Shingle length is equal the longest possible shingle thus length = maximum. Only the full noun phrase is used in the analysis
`lm_l1`	Shingles with length one or shingles with a length equal to the maximum
`l>1`	Any shingle with length higher than one
`m1_l>1`	Any shingle with length higher than one or any single term noun phrase.

The combination of these five possible selection criteria with the two options for the type of noun phrase and the possible sorting creates twenty different scenarios for the creation of a phrase by document matrix. This matrix contains only phrases or shingles that occur in more than one document and the weighting is a slightly modified TF-IDF version where the term frequency is equal to the number of sentences in which the phrase or shingle appears. Salton's cosine is calculated to express the similarity between documents. As a result we have for each document pair up to twenty different similarities based on the different scenarios in this NLP approach.

For the citation component we calculate the cosine similarity based on bibliographic coupling (BC), that is, the number of references shared by document pairs with respect to all of their references that are indexed in the Web of Science databases.

The two components lexical and bibliographic coupling are combined by calculating a hybrid similarity as the cosine of the weighted linear combination of the underlying angles of each of the cosine similarities. This method has been introduced and described by Glänzel & Thijs (2012). A free parameter (λ) defines the *convex combination* and the weight of the two components. For document pairs, where one of the components is not defined $\pi/2$ is used as the underlying angle of this component. Document pairs with two undefined components are discarded. In this paper we will only use an NLP based lexical component with seven values of the λ parameter (0.125, 0.25, 0.33, 0.5, 0.66, 0.75 and 0.875).

Clustering of the data is done by the Pajek 'Single Refinement' implementation (Batagelj & Mrvar, 2003) of the Louvain method for community detection (Blondel et al., 2008). Prior to this clustering all singletons are removed from the network. The resolution parameter is set to 1.0, and five random restarts are requested.

Results

In this section we discuss shortly the twenty networks resulted from the different options and compare seven hybrid combinations of the bibliographic coupling component together with the selected NLP component according to the above λ parameters. The density of the networks and the outcomes of the clustering algorithm are hardly influenced by the choice of noun-phrase types nor by the ordering of the terms the phrases. We only found that restricted noun-phrases resulted in much smaller data files. However, the creation of shingles from the noun-phrases had a large influence on the results. We found out that scenario's that still allow single term phrases did not reduce the density nor did change the distribution of edge weights. As a consequence the best result was obtained when restricting the lexical component to the use of shingles with a length higher than one.

For the second analysis we use hybrid combinations. In Table 2 the results for the two trivial combinations, i.e., $\lambda = 0$ and 1 is included for reference. After the hybrid combination, 25 documents remained singletons in the network and were removed. We would like to recall that the appropriate choice of the weight parameter λ used to be crucial for the quality of the clustering result with a possible distortion of the results by too much weight on the single term lexical approach (Janssens et al. 2008). However, Table 2 clearly shows that the distribution of weighted degree is not distorted by any particular choice of the λ parameter. Also, for each of the chosen values a modularity above 0.3 is obtained.

Table 2. Results of hybrid clustering with different weight parameters

	Weighted Degree			Community Detection		
Weight λ	Average	Median	Max	NC	Mod.	<10
NLP (λ= 0)	16.64	14.86	118.50	12	0.338	2
0.125	16.26	14.88	104.66	12	0.322	2
0.25	15.90	14.82	90.66	11	0.312	2
0.33	15.68	14.58	81.62	11	0.308	2
0.5	15.19	13.64	62.22	10	0.310	3
0.67	14.73	12.40	69.24	10	0.317	3
0.75	14.47	11.62	75.95	10	0.323	4
0.875	14.11	10.48	85.27	10	0.333	3
BC (λ= 1)	14.62	10.71	94.59	16	0.350	8

Table 3. Cramer's V measurement of association

	NLP	0.125	0.25	0.33	0.5	0.66	0.75	0.875
0.125	0.85							
0.25	0.79	0.86						
0.33	0.76	0.80	0.89					
0.5	0.66	0.71	0.74	0.74				
0.66	0.63	0.65	0.68	0.71	0.90			
0.75	0.62	0.64	0.66	0.69	0.87	0.93		
0.875	0.59	0.61	0.63	0.66	0.84	0.93	0.91	
BC	0.30	0.33	0.40	0.44	0.65	0.70	0.77	0.75

When looking at the number of clusters, it evolves from 12 in the lexical component to 10 in the $\lambda = 0.5$ weighting scheme to 16 in the link component. When we look at the correspondence of cluster assignment between two schemes we observe higher stability between schemes with λ values closer to each other. Cramer's V measures are calculated between all schemes and plotted in Table 3.

Application

This section outlines briefly the results of our partitioning of the hybrid network with $\lambda = 0.5$ weight on both components at three levels with increasing resolution (0.7, 1.0 and 1.5). As mentioned above, we used a data set on in Information System Research for our analyses. Level I resulted in three large clusters and two pairs or triplets of papers with no link to any other documents. These pairs/triplets (five papers at level I) are removed from further analysis. At level II we found seven clusters and three pairs/triplets (8 papers) and level III has 19 clusters and the same eight papers were grouped in three pairs/triplets. Although the three levels consist of independent runs of the Louvain cluster algorithm we can observe a near-perfect hierarchical structure. This is confirmed by Cramér's-V values of 0.94 between level I and II, 0.93 between I and III and 0.84 between levels II and III. The labels of each cluster at the three levels are taken from the titles of *core documents* within each cluster. These core documents have been determined according to Glänzel & Thijs (2011) and 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 and only edges with a minimum weight of 0.15 are retained. At the lowest level, the three clusters contain publications that fit in broad categories, such as '*planning/development/ implementation'* (cluster I.2 with 3855 papers), '*user and technology acceptance'* (cluster I.3 with 1302 papers), and '*decision support systems'* (cluster I.1 with 957 papers).

Level I Resolution = 0.7			Level II Resolution = 1.0			Level III Resolution = 1.5		
Cluster	# Pubs	Label	Cluster	# Pubs	Label	Cluster	# Pubs	Label
1	957	Decision Support Systems	a	955	Decision Support Systems	1	807	Decision Support System
						7	111	Communication, Virtual Teams
2	3855	Development, Implementation, Planning	b	1119	Development, (Open source) Software, Planning	2	398	Development Projects
						12	537	Design Science in IS Research
						15	43	Conceptual Modeling
			c	1414	Strategic IS Planning, Management	5	454	Strategic Planning
						8	429	Performance Measurement
						13	51	Executive Perspective
						11	359	HRM & Accounting
						18	210	Enterprise Resource Planning
			e	1108	Supply Chain	6	24	Innovation, Assimilation & Diffusion
						9	107	Outsourcing
						14	338	Firm Performance
						10	442	Supply Chain Management
						19	334	Open Source
			f	376	Intangible Asset Management	16	367	Knowledge Management
						22	26	Customer Relationship Management
			h	48	Security	20	78	Security
3	1302	User Oriented	d	1092	Satisfaction, Technology Acceptance Model	3	288	Satisfaction, Service Quality
						4	709	Technology Use and Acceptance

Figure 1. Cluster solution at three levels
[Data sourced from Thomson Reuters Web of Science Core Collection]

Given the size of the *planning/development/implementation* cluster and the hierarchical structure of the different levels, there is value in exploring the clustering at a higher resolution which allows us to develop a more differentiated understanding of the IS literature that falls in this category. At Level II with a resolution of 1.0 we identify 5 clusters. There are three large clusters: '*II.c strategic IS planning*' (1414 papers), '*II .b development /OSS /planning'* (1119 papers), '*II.e supply chai*n' (1108). Smaller clusters were also found with one midsized cluster: '*II.f intangible assets*' (376) and one small but emergent topic: '*II.h security*' (48). This last cluster is not further partitioned at level III. The three large Level II clusters can be divided further. The obtained hierarchical structure for the three levels is shown in Figure 1.

Conclusions

Based on the data presented in this paper we can conclude that the extraction of noun phrases from abstracts and titles can considerably improve the lexical component in the hybrid clustering. However, using the noun phrase itself is not sufficient for the improvement. Only if data is restricted to shingles with at least two terms constructed out of the noun phrases an improvement in the clustering is observed. We found that many of the shingles only appear once in each document which allows us to bring the calculation of similarities in the lexical approach more in line with the bibliographic coupling by abandoning the TF-IDF weighting and adopting a binary approach. The new approach was tested in a hybrid combination and resulted in a valid clustering of the field of '*Information System Research'* with different resolution levels and changing weights for both components. This methodology has several advantages over the other scenarios. The risk of distorting the network by choosing not the optimum parameter or even an inappropriate parameter in the hybrid approach is distinctly reduced. It seems that the parameter will not be used anymore in a function to set the right focus on the document set but to change the viewpoint while the clustering stays in focus.

References

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

Blondel, V.D., Guillaume, JL., Lambiotte, R. & Lefebvre, E. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10, P10008

Braam, R.R., Moed, H.F., van Raan, A.F.J. (1991a). Mapping of science by combined cocitation and word analysis, Part 1: Structural aspects. *JASIS*, 42 (4), 233–251.

Braam, R.R., Moed, H.F., van Raan, A.F.J. (1991b). Mapping of science by combined cocitation and word analysis, Part II: Dynamical aspects. *JASIS*, 42 (4), 252–266.

Callon, M., Courtial, J. P., Turner, W., & Brain, S. (1983). From translations to problematic networks. An introduction to co-word analysis. *Social Science Information*, 22(2), 191–235

de Marneffe, MC & Manning, C.D. (2008). *The Stanford typed dependencies representation*. In: COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation.

Dunning, T. (1993) Accurate methods for the statics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

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.

Glenisson, P., Glänzel, W., Janssens, F & De Moor, B. (2005) Combining full text and bibliometric information in mapping scientific disciplines. *Inf. Proc. & Management*, 41, 1548–1572.

Janssens, F., Glenisson, P., Glänzel, W. & De Moor, B. (2005), *Co-clustering approaches to integrate lexical and bibliographical information*. In: P. Ingwersen, B. Larsen: Proc. of the 10th ISSI Conference, Karolinska University Press, Stockholm, 285–289.

Janssens, F., Glänzel, W. & De Moor, B (2008) A hybrid mapping of information science. *Scientometrics*, 75 (3), 607–631.

Klein, D. & Manning, C.D. (2003), *Accurate Unlexicalized Parsing*. Proceedings of the 41st Meeting of the Association for Computational Linguistics, 423–430.

Maning, C.D. & Schütze, H. (2000), *Foundations of Statistical Natural Language Processing*. Combridge: MIT press.

Meyer, M., Grant, K., Thijs, B., Zhang, L., Glänzel, W. (2013), *The Evolution of Information Systems as a Research Field*. Paper presented at the 9th International Conference on Webometrics, Informetrics and Scientometrics and 14th COLLNET Meeting, Tartu, Estonia.

Porter, M.F., (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.