ISSI NEWSLETTER
QUATERLY E-NEWSLETTER OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS
ISSN 1998-5460
#43 / VOLUME 11 NUMBER 3
SEPTEMBER 2015

CONTENTS

REPORT ON THE 20th STI CONFERENCE

2-4 SEPTEMBER 2015,
UNIVERSITÀ DELLA SVIZZERA ITALIANA, LUGANO, SWITZERLAND

INTRODUCTION

BENEDETTO LEPORI

The 20th Science and Technology Indicators (STI) conference was held at the Università della Svizzera italiana in Lugano from the 2nd to the 4th of September 2015 under the theme "Research Indicators under Scrutiny" (http://www.sti2015.usi.ch/). The STI annual conference series is run under the auspices of the European Network of Indicator Designers (ENID) Association and was supported by the RISIS EU-FP7 project, by the Swiss National Science Foundation and by Swissuniversities, as well as by Thomson Reuters and Elsevier.

While dealing with the development of Science and Technology Indicators, each STI conference should have a focus on a specific topic

ISSI e-Newsletter (ISSN 1998-5460) is published by ISSI (http://www.issi-society.org/).
Contributors to the newsletter should contact the editorial board by e-mail.

Lugano. *Photo copyright © Bastian Greshake. Source: Wikimedia Commons. License: Creative Commons Attribution-Share Alike 2.0 Generic*

or approach, in order to provide new directions and impulses to the community in the field. In Lugano, this focus was on the development and use of S&T indicators to characterize and understand the behaviour of research organizations, including higher education institutions, Public Research Organizations and Research Funding Organizations. This year's STI conference was therefore the expression of a longstanding process of development of indicators to characterize the profile and the position of organizations in the broader field of science and technology, the so-called positioning indicators (Lepori, Barré and Filliatreau 2008). Large initiatives in this direction include the development of databases on higher education institutions, the use of bibliometric indicators for the creation of rankings of institutional research performance, new developments concerning academic patents, the development of datasets covering different aspects of the research activities, from research projects to research impact assessment and research programs.

PROGRAM AND EVENTS

Around this federating theme, the conference program included a set of complementary elements. The backbone was represented by around 35 oral presentations of full papers, selected upon a strict peer review process and organized in thematic sessions on major topics, like academic careers, research funding, higher education institutions, bibliometric indicators, as well as by a poster session displaying a number of on-going research activities and emerging developments. The book of abstracts available on the conference website provides an overview of the richness of topics and contents presented.

The second main component was represented by four thematic panels on mixed-methods, researchers' careers, standards in funding evaluation and research responsible innovation. While also presenting novel research insights and potentially promising data sources (like in the case of careers), they also allowed for engaged discussions on central topics for the STI research, but also for its practical relevance, like in the case of evaluation standards adopted by research councils.

Two central events in the conference were the two plenary panels. The panel discussion on the *dynamics of standards* aimed at promoting a collective and interdisciplinary reflection on how standards are constructed and adopted in evaluation processes and on their implications for the outcome of evaluation and the behaviour of actors. The session moved from some recent documents on standards like the Leiden manifesto, and was introduced by a keynote speech on standard as socio-organizational constructions by David Seidl from the University of Zurich. This panel therefore expressed the blending between organizational scholarships and indicators development which characterized the whole conference. The second event was a panel on establishing a European data infrastructure, where organizational solutions to manage the future STI data infrastructure were discussed. Such issues are at the core of the EU-FP7 project on Research Infrastructure for Research and Innovation Policy Studies (RISIS), which promoted this debate.

An important contribution to the conference was given by the keynote speeches. The speeches of Filippo Wezel (Università della Svizzera italiana), David Lampel (University of Manchester) and Barry Bozeman (Arizona State University) provided novel insights on relevant topics for our community, like academic careers and mobility of employees, imitation effects generated by rankings and, finally, the diffusion of red-tape bureaucracy within universities.

The conference also marked an important event in the life of the ENID association, with the renewal of the conference committee and the nomination of Jordi Molas-Gallart as a new president, who replaced Ton van Raan. The association also decided that the next conference will take place in Valencia (2016), Paris (2017) and Leiden (2018).

CONCLUSION

The 2015 edition of the STI conference was highly appreciated by the about 140 participants from 35 different countries. It provided important theoretical insights on how STI organizations could be studied using data and indicators available in the community; it allowed presenting novel results; most of all, it was characterized by an engaged discussion and exchange climate between the participants, displaying clearly how cohesive has become the European STI community in the last decade. The choice made by the ENID association to blend in the conference series bibliometrics and informetrics with other domains of development of STI indicators like higher education, careers and funding studies has proved indeed fruitful for the whole community.

The next edition of the STI conference will be held in Valencia from the 14th to 16th of September 2016. More information is available on the conference website http://www.sti2016.org/.

REFERENCES

1. Lepori, B., Barré & Filliatreau (2008). New perspectives and challenges for the design and production of S&T indicators. *Research Evaluation*, 17(1), 33-44.

AFTER ISSI 2015

A CONFERENCE REPORT BY

ALMILA AKDAĞ SALAH

ALBERT ALI SALAH

The 15th International Society of Scientometrics and Informetrics Conference took place at Boğaziçi University in Istanbul from 29th of June to 4th of July. The Conference was jointly organised by Boğaziçi University, Hacettepe University, and the TÜBİTAK ULAKBIM (Turkish Academic Network and Information Center – The Scientific and Technological Research Council of Turkey) under the auspices of ISSI – the International Society for Scientometrics and Informetrics.

With ISSI'15, the organization committee enhanced the conference with some new approaches. The first of these was to introduce a focus theme for the conference, namely "**the future of scientometrics**". Our keynotes responded to the theme by addressing the issue from different perspectives.

The keynote given by Loet Leydesdorff demonstrated the potential of thinking of science as a complex institution. By building on the Triple Helix Model of University-Industry-Government relations, Dr. Leydesdorff showed that innovation systems can provide institutional mediation between knowledge production, wealth generation, and governance.

The second keynote, by Kevin Boyack, directly answered the challenge of the focus theme, and proposed several opportunities to expand the field of scientometrics. Dr. Boyack stressed that scientometrics at the moment only deals with only a few key issues intensively (e.g. publications, patents, citations) and called for increasing attention to funding, workforce, data and instrumentation, research objects, and innovation. The model he presented and expanded was based on Bruno Latour's model of science [1987][1].

ISSI'15 opened with a full day of workshops and tutorials. We had **five workshops** and **four tutorials**. The increasing number

1 Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard university press.

of open-source software for scientometrics presents great opportunities for researchers. Open source data analysis and visualization tools, citation exploration software, measurement of scholarly impact, and on social network analysis with the popular R software were the topics of the tutorials.

The satellite workshops of the conference reflected the diversity of the field. In "**Mining Scientific Papers: Computational Linguistics and Bibliometrics",** researchers in bibliometrics and computational linguistics are brought together to study the ways bibliometrics can benefit from large-scale text analytics and sense mining of scientific papers, thus exploring the interdisciplinarity of Bibliometrics and Natural Language Processing. The workshop on "**Grand challenges in data integration for research and innovation policy**" dealt with problems of big, open and linked data. The "**Forecasting science: Models of science and technology dynamics for innovation policy"** workshop discussed methodology for predicting the circumstances leading to scientific or technological innovation. "**Workshop on Bibliometrics Education**" brought together educational institutions, employers, professional societies, and Bibliometrics researchers and professionals to tackle this problem. Finally, "**Google Scholar and related products"** was a highly interactive workshop on the benefits and limitations of some of the most important citation tools.

We hosted **four special sessions** on a range of topics, including performance indicators, algorithms for topic detection, empirical evaluation of education, research and innovation, and how scientometrics can be used to improve and inform university rankings. These special sessions included poster presentations, panel discussions, invited speakers, and public debates. The session on university rankings showed how a few well-cited papers could have significant impact on the ranking of universities, countries, and even regions.

The **Doctoral Forum**, organized by Andrea Scharnhorst and Judit Bar-Ilan, was a meeting of senior researchers and selected doctoral students for presenting and discussing research projects and an excellent way for students of getting valuable feedback, along with strong networking opportunities. This is the sixth ISSI Doctoral Forum

Photos' copyright: © Kenan Özcan

Photos' copyright: © Kenan Özcan

and we are extremely happy about the interest it continues to receive from the community. This year, the forum had almost the same number of mentors as the students. With Kevin W. Boyack, Rodrigo Costas, Santo Fortunato, Wolfgang Glaenzel, Sybille Hinze, Andrea Scharnhorst, Cassidy Sugimoto, Gunnar Sivertsen, and Theresa Velden as mentors, the forum was able to provide one-to-one mentoring.

During the pre-conference day Katy Boerner gave a talk on *Places & Spaces: Mapping Science*. Places & Spaces is an exhibit curated by the Cyberinfrastructure for Network Science Center at Indiana University, and it was hosted at the conference venue for the duration of ISSI'15. The exhibit aims to introduce science mapping techniques to the general public and to experts across diverse disciplines for educational, scientific, and practical purposes. It is meant to inspire cross-disciplinary discussion on how to best track and communicate scholarly activity and scientific progress on a global scale.

The first phase of Places & Spaces is ended after the 10th iteration. During these first ten years, each year the exhibit curators announced a focus theme, and made a selection of 10 maps for each year. After ten years, the exhibit with 100 science maps has traced the evolution of science maps, featuring the best examples of knowledge domain mapping, novel location-based cartographies, data visualizations, and science-inspired art works. During the ISSI'15 all the 10 iterations of Places Spaces exhibition was open to ISSI participants, as well as to the general public.

At the last day of the conference, the Derek de Solla Price Award of the international journal *Scientometrics* was given to Mike Thelwall, Professor of Information Science at the University of Wolverhampton (UK), in a special session organized for this purpose. This award recognizes excellence through outstanding, sustained career achievements in the field of quantitative studies of science and their applications. Unfortunately Mike Thelwall was not able to join the ceremony, and in his name, Kayvan Koshua received the medal from Wolfgang Glaenzel, and read the laudation.

At the closing ceremony, the prestigious Eugene Garfield Doctoral Dissertation Scholarship was given by the Eugene Gar-

field Foundation. Ronald Rousseau named the prize winner, Cathelijn Waaijer from Centre for Science and Technology Studies (CWTS, Leiden University, the Netherlands), with a dissertation entitled "Academic career systems: The most important factor in academic research is not infrastructure, but people". Zohreh Zahedi from the same institution received an honorable mention.

The ceremony continued with handing in awards on the following categories: outstanding paper award, outstanding student paper award, outstanding poster award, and outstanding reviewer award. We hope that in the future these awards will be continued in recognition of outstanding work, as well as outstanding feedback given by the scientific committee. Hereby, we would like to express our thanks to our reviewers, whose comments guided us in choosing the award winning papers, which are:

- **Interdisciplinarity and Impact: Distinct Effects of Variety, Balance, and Disparity** by *Jian Wang, Bart Thijs, Wolfgang Glänzel (outstanding paper)*
- **Research Data Explored: Citations versus Altmetrics** by *Isabella Peters, Peter Kraker, Elisabeth Lex, Christian Gumpenberger and Juan Gorraiz (outstanding paper)*
- **A Preliminary Study of Technological Evolution: From the Perspective of the USPC Reclassification** by *Hui-Yun Sung, Chun-Chieh Wang and Mu-Hsuan Huang (outstanding paper)*
- **Understanding Relationship Between Scholars' Breadth of Research and Scientific Impact** by *Shiyan Yan, Carl Lagoze (outstanding student paper)*

The outstanding poster award is chosen by generating a poll among the conference participants and asking their opinion. The program chairs shortlisted a number of posters based on the reviewer comments, and the participants voted for posters in this list. The most voted poster among this selection was **Tweet or Publish: A Comparison of 395 Professors on Twitter** by *Timothy D. Bowman*.

We also wanted to acknowledge our gratitude to our most prolific reviewers. Andrea Scharnhorst, Jacqueline Leta, Thed van Leeuwen and Alesia Zuccala, who all gave valuable feedback to our participants via their comments, received the Outstanding Reviewer awards.

A last word must be devoted to the conference statistics. ISSI'15 received 354 submissions: 164 full papers, 64 research-in-progress papers, 106 posters and 20 ignite talks. The overall acceptance rate was 58% with 207 papers. Among these we had 82 full papers (%50 acceptance), 41 research-in-progress papers (%64 acceptance), 71 posters (%67 acceptance) and 13 ignite talks (%65 acceptance). The ignite talks were to increase discussion of underrepresented topics and novel ideas. Because of the large number of papers, and to allow proper discussion for each paper, four parallel sessions were implemented. Several poster sessions were organized, each containing a relatively manageable number of posters. The conference brought together 302 researchers from 46 countries and the works of 458 researchers were presented.

ORGANISATION AND COMMITTEES

CONFERENCE CHAIRS
Albert Ali Salah
Yaşar Tonta
M. Mirat Satoğlu

PROGRAMME CHAIRS
Almıla Akdağ Salah
Cassidy Sugimoto
Umut Al

WORKSHOPS & TUTORIALS CHAIR
Caroline Wagner

DOCTORAL CONSORTIUM CHAIRS
Andrea Scharnhorst
Judit Bar-Ilan

LOCAL ORGANIZATION CHAIR
Heysem Kaya

PETER INGWERSEN— AWARDEE OF THE 2015 TONY KENT STRIX AWARD

ABOUT THE AWARD

It occurred just a couple of months ago that the ISSI Newsletter had the pleasure to inform its readers about the third honorary doctorate of our valued colleague, Peter Ingwersen (Professor Emeritus at The Royal School of Library and Information Science, Copenhagen, Denmark; and amongst many other awards, recipient of the 2005 Derek de Solla Price Medal), but it seems that after a long and adventurous career he just cannot sit back and relax in his comfortable armchair in his home in Malmö, as in recognition of his academic achievements newer and newer travelling destinations are waiting for him. This time he had to fly to London to receive the prestigious Tony Kent Strix Award.

The Strix Award is annually presented in memory of Dr Tony Kent (†1997), a past Fellow of the Institute of Information Scientists (UK), and is given in recognition of an outstanding contribution to the field of information retrieval in its widest sense. Dr Tony Kent made a major contribution to the development of information science and information services both in the UK and internationally. The name 'Strix' was chosen to reflect Tony's interest in ornithology (a particular kind of owl), and as the name of the last and most successful information retrieval packages that he created.

The candidates for the Award are assessed and the Award is issued by the joint body of the UK eInformation Group (UKeiG); the Chemical Information and Computer Applications Group of the Royal Society of Chemistry; the International Society for Knowledge Organisation UK and the British Computer Society Information Retrieval Specialist Group. This year's awarding ceremony took place on 21 October at the Enterprise Search Europe Conference in London.

In the first place Peter Ingwersen had been recommended for the Award by two close colleagues and friends, Christina Lioma (Department of Computer Science, University of Copenhagen) and Birger Larsen (Department of Communication, Aalborg University Copenhagen) but later on several colleagues from American, British, Chinese, Danish, Italian, Finnish, Scottish, Spanish and Swedish academic circles joined to support the recommendation.

For more information on the Award, visit the website of CILIP (Chartered Institute of Librarians and Information Professionals): http://www.cilip.org.uk/uk-einformation-group/awards-and-bursaries/tony-kent-strix-award.

PERSONAL REFLECTIONS – BY PETER INGWERSEN

It was a great surprise to me to have the honor to be presented with the prestigious Tony Kent Strix Award 2015, given "in recognition of outstanding contribution to the field of information retrieval (IR) in its widest sense". The Award is managed by an Executive Committee and UKeiG in partnership

with the Royal Society of Chemistry, the International Society for Knowledge Organisation UK Chapter (ISKO UK) and the British Computer Society Information Retrieval Specialist Group (BCS IRSG).

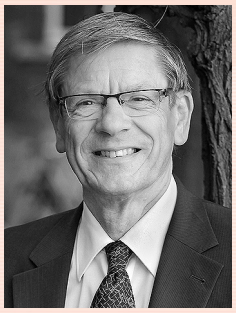
I have now had the opportunity to look into the nomination text and the 20 supporting letters from outstanding IR researchers worldwide. I am very grateful for their acknowledgements of my contributions to the field over three decades of research. Like with citations received to your work you obtain a spectrum of perspectives of your research and activities, which you often were not really aware of. It seems that I have had (and perhaps still have) a much stronger role as mentor than I perceived. I really appreciate that achievement.

The ceremony took place as part of the Enterprise Search Europe Conference, 2015, held in London at the Olympia Conference Center, Kensington, October 21. The last time I participated in that similar conference was in 1993. At that time the IT explosion was just at its beginning. The personal computer was indeed beginning to become a major player, but the Internet was not a hot topic and no smart phones existed. On the other hand the conference discussed, as it did this year, how to decrease the gap between IR research and practical applications of academic results in business and libraries. In 1993 the information profession was, unknowingly, at the end of the 'intermediary' era, because the end-user was increasing its direct access to information due to the PC, and five years later due to the global Internet and the Web. It is interesting to observe that while value-adding of information was synonymous with online bibliographic (academic) databases and OPACs prior to the Web, the value-adding now takes place at search time. Albeit, bibliographic databases still exist as part of the hidden Web, value-added by means of human interference and costly, but the 'free' algorithmic-based search engines on the web form the basic access point for almost all kinds of end-users, laymen as well as academics. If one re-

Photo copyright: © Internet Librarian International

members how difficult it was to convince and transfer the 'ranking principle' in experimental IR to major online vendors and applied search environments 1976-1996, it is interesting to observe how fast that principle became the only available and accepted search mode on the Web. That principle together with other algorithmic solutions provided by IR research (like automatic pseudo-relevance feedback, misspelling correction, auto-provision of alternative query strings or recommender-based social 'personalization' of searching) constitutes now-a-days the value-adding provided for end-users. The alternative query strings provided by search engines are like older times' look-ups in thesauri or descriptor lists: recognition is easier than creation. Given these novel kind of search facilities, with no set combinations available in a still fundamentally archaic Boolean-like file structure on the Web, it is amazing how long time it took to recognize the importance of user-orientation by the IR research community. I somehow see my award as a sign of this recognition.

EINSTEIN'S GREAT IDEA AND ITS IMPACT: ON THE 100th ANNIVERSARY OF GENERAL RELATIVITY



FRED Y. YE
1) School of Information Management, Nanjing University, China
2) Jiangsu Key Laboratory of Data Engineering & Knowledge Service, Nanjing, China
yye@nju.edu.cn

RONALD ROUSSEAU
1) Department of Mathematics, KU Leuven, Belgium
2) Library and Information Science (IBW), University of Antwerp, Belgium
ronald.rousseau@kuleuven.be

ABSTRACT: A great idea can transform the minds of scientists and of science's knowledge structure itself. Einstein's article on general relativity is not just one of his publications, but also, and more importantly, a great and universal idea. Coming to its 100 anniversary, scientists are still inspired and enlightened by its intellectual content.

INTRODUCTION

After Einstein published his famous article on special relativity (Einstein, 1905), he realized that the principle of relativity could also be extended to gravitational fields, linking gravity to space-time curvature. In collaboration with the mathematician Marcel Grossmann he basically found a gravitational theory by the end of 1913 (Einstein & Grossmann, 1913; Norton, 1984). Yet it took him another two years to clean up the last technical details (Einstein, 1915a, b). Finally, in 1916 he presented a first complete exposé of the so-called general relativity theory (Einstein, 1916). This theory updated Newton's theory of gravity which had been around for more than two hundred years. Nowadays the consequences of general relativity speaking most to the imagination of many lay persons is the existence of black holes and the origin of the universe in the Big Bang, this none in the least by popular works such as Stephen Hawking's *A Brief History of Time* (1988) and Stephen Hawking and Roger Penrose's *The Nature of Space and Time* (1996).

GENERAL RELATIVITY AND CITATION RESULTS

One might expect that Einstein's tour de force made him a citation star. However, this turns out not to be true. The search strategy AU=(Einstein, A.) and PY=1915 in the Web of

	selected field	total publications
1	astronomy astrophysics	10,524
2	physics particles fields	8,286
3	physics multidisciplinary	8,236
4	physics mathematical	1,956
5	physics nuclear	651
6	multidisciplinary sciences	547
7	mathematics applied	407
8	mathematics	396
9	physics applied	291
10	history philosophy of science	262
	social sciences interdisciplinary	14
	humanities multidisciplinary	10

Table 1. *Selected fields obtained by searching for the topic 'general relativity'*

Science (WoS) in May 2015 led to a scanty 79 citations over a century (1915-2015) for "Zur allgemeinen Relativitätstheorie" (in English: On the general theory of relativity) and 84 citations for "Feldgleichungen der Gravitation" (in English: The field equations of gravitation). His complete exposé of general relativity ("Die Grundlage der allgemeinen Relativitätstheorie" (in English: The basics of general relativity theory) received 830 citations from 1916 till 2015 (again in the Web of Science).

However, when searching for TS=(General relativity) over the period 1915-2015, we obtained 19.974 publications as output. Among these the highest cited paper is Israel's article published in Nuovo Cimento (Israel, 1966), cited 1341 times, followed by 'Hehl et al., General relativity with spin and torsion – foundations and prospects' (1976) cited 1340 times.

These simple searches show that the idea of general relativity is much better known, or at least more cited, than Einstein's original publications. They illustrate what Merton (1968) and Garfield (1975) described as obliteration by incorporation, the undercitation of well-established concepts and the articles in which they were introduced because they are considered to be part of the canonical knowledge of their field.

DIFFUSION OF THE "GENERAL RELATIVITY" IDEA OVER DIFFERENT FIELDS

The influence of Einstein's general relativity theory has gone beyond physics. As a simple illustration we searched the Web of Science (WoS) using the following query: (TS=general relativity, time span: 1915-2015, data collected on May 11, 2015). Partial results are shown in Table 1, where we show some physics related fields and two multidisciplinary ones.

PENETRATION OF SCIENTIFIC IDEAS

These data tell us that just counting citations to Einstein's publications, would indicate a limited impact. However, if we consider the impact of the idea of 'general relativity' we see that it influenced many fields, including the social sciences and humanities. The case of Einstein's articles on general relativity illustrates that an idea as published in a paper is only a part of a publication, but this idea has real impact if it is useful in its field and, most importantly, has lasting influence if it leads to multidisciplinary penetration, as sketched in Figure 1.

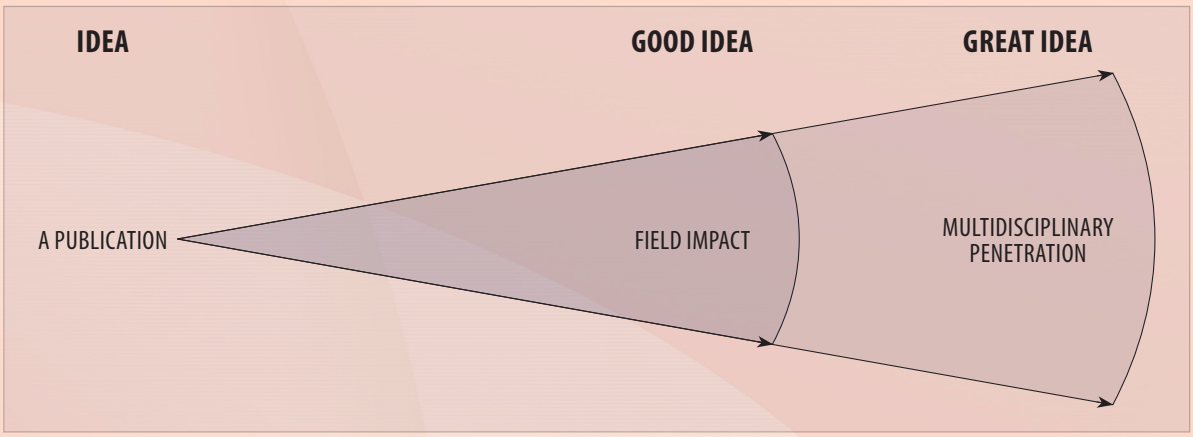


Fig. 1 *Symbolic representation of the penetration of a scientific idea*

These observations point to the fact that the penetration of an idea proposed in an article is much more important than the article's impact itself. When we encounter ideas having a multidisciplinary penetration, then these are great ideas, able to promote real knowledge progress in the world.

CONCLUSION

If one wants to apply informetric methods to measure an idea, the idea must have become a 'subject' or 'topic' in a field, so that it can be retrieved by a topic search in a database. We say that such an idea is a good idea if it is useful in a field, and that it is a great idea when it has penetrated many other fields. This can be observed by studying its diffusion (through received citations) in these fields (Wagner et al., 2011).

Most importantly, a great idea can transform the minds of scientists and of science's knowledge structure itself. General relativity it is not just one of Einstein's publications, but also, and more importantly, a great and universal idea. Coming to its 100th anniversary, scientists are still inspired and enlightened by the brilliance of its intellectual light.

Acknowledgements: Fred Y. Ye is supported by the National Natural Science Foundation of China Grant No 71173187. Ronald Rousseau is supported by National Natural Science Foundation of China Grant No 71173185.

REFERENCES

1. Einstein, A. (1905). Zur Elektrodynamik bewegter Körper (English translation: On the electrodynamics of moving bodies). *Annalen der Physik*. 17:891-921.

2. Einstein, A. (1915a). Zur allgemeinen Relativitätstheorie. (English translation: On the general theory of relativity). *Sitzungsberichte der Königlich-Preussischen Akademie der Wissenschaften*, (part 2), 778-786.

3. Einstein, A. (1915b). Feldgleichungen der Gravitation (English translation: The field equations of gravitation). *Sitzungsberichte der Königlich-Preussischen Akademie der Wissenschaften*, (part 2), 844–847.

4. Einstein, A. (1916). Die Grundlagen der allgemeinen Relativitätstheorie. (English translation: The Foundation of the general theory of relativity). *Annalen der Physik*, 49: 769–822 (1916).

5. Einstein, A. & Grossmann, M. (1913). Entwurf einer verallgemeinerten Relativitätstheorie und einer Theorie der Gravitation. *Zeitschrift für Mathematik und Physik*, 62, 225-261.

6. Garfield, E. (1975). The "Obliteration Phenomenon" in science – and the advantage of being obliterated! *Current Contents*, 51-52, December 22. Reprinted in Garfield E. (1977). *Essays of an Information Scientist*, *2*, 396-398.

7. Hawking, S. (1988). A *Brief History of Time: From the Big Bang to Black Holes*. New York: Bantam Books.

8. Hawking, S. & Penrose, R. (1996). *The Nature of Space and Time.* Princeton University Press. ISBN 0-691-03791-4.

9. Hehl, F.W., von der Heyde, P., Kerlick, G.D. & Nester, J.M. (1976). General relativity with spin and torsion: Foundations and prospects. *Reviews of Modern Physics*, 48(3), 393-416.

10. Israel, W. (1966). Singular hypersurfaces and thin shells in general relativity. *Il Nuovo Cimento B*, 44(1), 1-14.

11. Merton, R.K. (1968). The Matthew effect in science. *Science*, 159(3810), 56-63.

12. Norton, J. (1984). How Einstein found his field equations: 1912-1915. *Historical studies in the Physical Sciences*, 14(2), 253-316.

13. Penrose, R. (2004). *The Road to Reality: a complete guide to the laws of the universe.* London: Jonathan Cape.

14. Wagner, C.S., Roessner, J.D., Bobb, K., Klein, J.T., Boyack, K.W., Keyton, J., Rafols, I. & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1), 14-26.

STUDYING THE STRUCTURE OF "INFORMATION SYSTEMS" RESEARCH USING SOME IMPROVEMENT OF LEXICAL SIMILARITIES THROUGH NOUN PHRASES EXTRACTION IN HYBRID DOCUMENT CLUSTERING

BART THIJS
KU Leuven,
ECOOM and Dept. MSI,
Leuven, Belgium
bart.thijs@kuleuven.be

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

MARTIN MEYER
1) Kent Business School,
University of Kent,
Canterbury, UK
2) KU Leuven, ECOOM &
Dept. MSI, Leuven, Belgium
3) SC-Research, University
of Vaasa, Lapua, Finland
m.s.meyer@kent.ac.uk

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. We discuss twenty different extraction-shingling scenarios and compare their results. Some scenarios show no improvement compared with the single term lexical approach used so far. But when all single term shingles are removed from the dataset the network has properties which are comparable with those from a bibliographic coupling based network. Next, 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.

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 & b) 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. Later, Noyons and Van Raan (1994) constructed geometrically organized maps based on co-occurrence of keywords in patents and publications to illustrate possible links between science and technology. In the same year Zitt and Bassecoulard (1994) used lexical and co-citation analysis for trend detection and analysis. In 2005, Glenisson et al. (2005a and 2005b) started using full text instead of the rather limited set of keywords originating from subject heading, titles or abstracts and they could compare the performance of the full text approach with the combined approach with title/abstract and reference-based text analysis. The application of bibliometric indicators allowed to finetune the clusters found after full text mining. Processing these full texts introduced new problems that were less likely to occur with keyword based approaches. Stemming was required to reduce the English words to their stem (the Porter stemmer was used for this purpose, see Porter, 1980). The dimensionality of the representation of documents in a vector space grew and Singular Value Decomposition was introduced to reduce this dimensionality. Glenisson 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 the following scheme:

$$D^{INTEGR} = \lambda D^{TEXT} + (1 - \lambda) D^{BIBL}$$

While introducing this simple weighted linear combination, he immediately lists some issues that need to be solved. One of these is the different distributional characteristics of the combined similarities and another is the choice of the weighting parameter which is set manually and is quite arbitrary. In this 2005 paper the authors use a Silhouette Value per Cluster as introduced by Jain and Dubes (1988) as an estimator for λ. 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. Each similarity is converted into their corresponding *p*-value based on a cumulative distribution function of the similarities in a completely randomized dataset. This method solves the issue of different distributions drastically but it introduces an even more complex calculation scheme. Glanzel & 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. In what follows we first describe a small data set that is used for the development of the methodology. Then, a short introduction to NLP is given, however, we refer to existing literature on this topic for an in-depth discussion

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 similarity measure between documents (Salton & McGill, 1986).

NOUN PHRASE EXTRACTION

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 is a syntactic tree which describes 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 detailed information on the parsing procedures we refer to their website: http://nlp.stanford.edu.

Prior to the parsing, the DocumentPreprocessor, a Java Class provided in the Stanford Parser package, is used to extract separate sentences from the abstracts. Each sentence is numbered for retrieval purposes afterwards.

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

LABEL	RESULT OF THE STANFORD PARSER
	(ROOT
	(S
A	(NP
A1	(NP (NNS Results))
	(PP (IN of)
A2	(NP (DT the) (NN study))))
	(VP (VBP show)
	(SBAR (IN that)
	(S
B	(NP (NN information) (NNS systems) (NN downsizing))
	(VP (MD may)
	(VP (VB produce)
C	(NP
C1	(NP (NNS benefits))
	(PP (JJ such) (IN as)
C2	(NP
C2a	(NP (VBN improved) (NN information) (NNS systems))
	(, ,)
C2b	(NP (VBN improved) (JJ organizational) (NN structure))
	(, ,)
C2c	(NP (JJR higher) (NN productivity))
	(, ,)
	(CC and)
C2d	(NP (JJR lower) (NN cost)))))))))))

Table 1. *Result of parsing the example sentence*
[Data sourced from Thomson Reuters Web of Science Core Collection]

Next, each sentence is parsed and each resulting noun phrase is numbered sequentially. Table 1 presents the output of the parsing of one sentence for one of the selected papers. We have added the labels. This sentence will be used as an example throughout the paper

> `Results of the study show that information systems downsizing may produce benefits such as improved information systems, improved organizational structure, higher productivity, and lower cost.`

For the selection of the noun phrases from the parsing result we have several options. It is the objective of this paper to study the consequences of these options. Each option or scenario is tagged differently.

- For the selection of noun phrases we have two options: Complete noun phrases (NP) can be selected or only leaf noun phrases in which no other noun phrase is embedded. Noun phrase *A* (`Results of the study`) in the example is such a complete noun phrase. But also *A1* (`Results`) is a complete noun phrase. However, *A1* has no embedded NP while *A1* is embedded in *A*. Analogously *C*, *C1*, *C2*, *C2a*, *C2b*, *C2c* and *C2d* are complete noun phrases while only *C1*, *C2a*, *C2b C2c* and *C2d* have no embedded NP. In this paper we will use the tag `NP` to denote complete noun phrases and `NPwoNP` for noun phrases without embedded noun phrases.
- Noun phrases can be recorded with the constituent words in the given order or the included terms can be sorted alphabetically. In our data set we found both following phrases '`information`

management system' and 'management information system'. Papers using these different versions of the same concept would not be linked to each other. As mentioned in the introduction a bi-gram detection method like the Dunning likelihood ratio test would not be able to detect these variation. Tags for the sorted scenarios get the suffix '_o'.

After selection of the type of noun phrase 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. Table 2 presents the shingles that can be identified for noun phrase *B* in the example (information systems downsizing). The number of terms in the phrase is 3 and that is also the length of the longest possible shingle in this phrase.

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. Each possibility is labelled by an appropriate tag and this is combined with the above mentioned tags. Table 3 lists the five tags and the applicable criteria.

LENGTH	SHINGLE
1	Information
1	Systems
1	Downsizing
2	information systems
2	systems downsizing
3	information systems downsizing

Table 2. *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.

Table 3. *All possible shingles in a phrase with three terms*

The combination of these five possible selection criteria with the 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.

We have to remark here that this parsing procedure is very demanding on computation resources. Parsing these 6144 documents on one machine with a single Intel Xeon-E5-1650 processor with six cores and 2 threads per core took about 36 hours with a fully multithreaded implementation of the parsing procedure. As of version 3.4 of the Stanford Parser released in June 2014 it contains a much faster parser which can reduce processing time. Another advantage of this NLP approach is that the processing is sentence-based which allows distributed processing over many machines. The computational complexity increases linearly with the size of the dataset. Opposed to the bi-gram detection method discussed above, the result of this procedure is independent from other papers in the dataset.

BIBLIOGRAPHIC COUPLING

For the citation component we calculate a bibliographic coupling cosine similarity based on the unique reference codes that Thomson-Reuters provides in its custom dataset. These codes are assigned to references in indexed papers and allow identifica-

tion of common references between indexed documents without the requirement that also the cited document is indexed. This choice improves the application of bibliographic coupling in those fields where many cited documents are not indexed.

HYBRID APPROACH

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.

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

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 (Glänzel & Thijs, 2012). Please note that Janssens (2005) used the λ parameter to weight the lexical component as opposed to our notation. For document pairs where one of the similarities is missing $\arccos(0)$ is uses as the underlying angle of this component. Document pairs where both similarities are missing are discarded. In this paper we will only use one version of the NLP based lexical component instead of all twenty but 7 different values of the λ parameter will be used (0.125, 0.25, 0.33, 0.5, 0.66, 0.75 and 0.875).

CLUSTERING

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 and 5 random restarts are requested. We report the number of clusters and the modularity of the clustering of each of the twenty nine networks that are created in data preparation (1 Single Term, 1 bibliographic coupling, 20 NLP versions and 7 hybrid networks).

RESULTS

As a kind of benchmark, the results of the single term and the bibliographic coupling network are presented in Table 4. This table illustrates the problem of different distributional characteristics already mentioned by Janssens (2005). The bibliographic coupling network is very sparse with a density of only 6.6%. The Single term network is nearly complete with a density of 97.5% and average degree close to the number of nodes in the network (5991.8 vs. 6144). In the bibliographic coupling network 392 documents are singletons without any link with other documents, in the single term network there are only 3. These singletons are removed for the calculation of the weighted degree and the clustering.

The comparison of the weighted degree distribution (see Figure 1) proves once more the different nature of both networks.

Consequently, also substantial differences in the results from the Louvain clustering are observed. In the `BibC` network, 16 clusters are found with a modularity of 0.35. However, eight small clusters with less than 10 documents are present. The modularity of the clustering of the `SingleT` network is extremely low, only 0.04 which means that despite the four large clusters, it is not possible to detect clear distinct topics in this network.

	DENSITY	AVERAGE DEGREE	SINGLE	WEIGHTED DEGREE			COMMUNITY DETECTION		
				AVERAGE	MED.	MAX.	NC	MOD.	<10
`BibC`	6.6%	403.6	392	14.6	10.7	94.6	16	0.35	8
`SingleT`	97.5%	5991.8	3	217.2	219.0	392.5	5	0.04	1

Table 4. *Network properties and clustering results for single term and bibliographic coupling networks [Data sourced from Thomson Reuters Web of Science Core Collection]*

Another problem arises from the near completeness of the Singe Term network. The number of possible undirected links between documents grows quadratic with the number of nodes in a network and is given by the formula

$$\frac{n(n-1)}{2},$$

where n denotes the number of documents. With this dataset of 6144 documents we already have 18.4 million lexical links. Using this approach comes with computational challenges for large datasets.

COMPARING NLP SCENARIOS

Next we compare the twenty NLP-based scenarios and their networks in Table 5. Due to the fact that in the complete NP approach the length of the noun phrases is higher than in the approach without embedded NPs, the files after parsing for NP contained many more lines and were about 50% larger in size. This is especially the case when comparing the two options without any restriction on length of the retrieved shingles. Ordering the phrases has no effect on number of returned lines or file size. For each scenario we calculated the total number of unique phrases found in the document set, the average number of unique phrases per document and the highest number of phrases in any document. After the construction of networks, also density and average degree are calculated. As we have the number of unique phrases and average number of phrases per document we could compute a density in a complete random network. In such a random network a number of phrases equal to the observed average number of phrases is assigned to each document. Possible links between documents can then be calculated and subsequently also the density in this completely random network. We used this formula to approximate the random density:

$$1-\left(1-\frac{C(u,a)-C(u-1,a)}{C(u,a)}\right)^{a}, \qquad (2)$$

where u denotes the total number of unique phrases, a the average number of phrases and $C(u,a)$ denotes the number of combinations of set u with a elements. Table 5 also reports this random density and the ration between the observed and the random density.

Several observations can already be made from the results presented in Table 5. The selection of type of noun phrase has an influence on the set of unique phrases and on the average and maximum number of

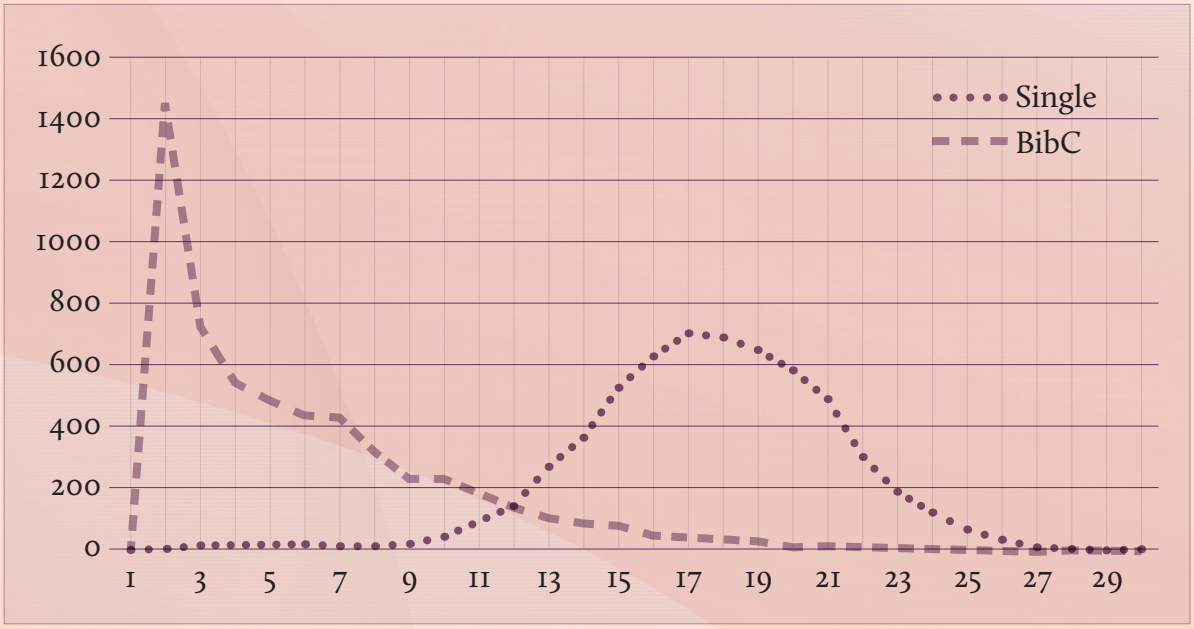


Figure 1. *Distribution of weighted degree for the bibliographic coupling network [Data sourced from Thomson Reuters Web of Science Core Collection]*

	PHRASES			AVERAGE DEGREE	DENSITY	RANDOM DENSITY	OBSERVED/ RANDOM
	UNIQUE	AVERAGE	MAXIMUM				
NP	33433	73.7	239	5908.41	96.2%	15.1%	6.4
NPo	34335	74.7	242	5908.41	96.2%	15.1%	6.4
NPwoNP	26978	63.6	201	5870.08	95.6%	14.1%	6.8
NPwoNP_o	27323	64.2	203	5870.08	95.6%	13.9%	6.8
NP_lm_l1	19078	62.0	200	5908.41	96.2%	18.3%	5.3
NP_lm_l1_o	19351	62.2	202	5908.41	96.2%	18.0%	5.3
NPwoNP_lm_l1	17762	55.1	174	5870.08	95.6%	15.7%	6.1
NPwoNP_lm_l1_o	17932	55.3	175	5870.08	95.6%	15.5%	6.1
NP_lm	16080	25.5	94	3546.63	57.7%	3.8%	15.1
NP_lm_o	16353	25.8	96	3548.91	57.8%	4.1%	14.2
NPwoNP_lm	15043	24.6	91	3520.17	57.3%	4.1%	14.1
NPwoNP_lm_o	15213	24.8	93	3522.20	57.3%	4.0%	14.2
NP_m1_l>1	30435	37.2	111	3983.86	64.9%	4.4%	14.7
NP_m1_l>1_o	31338	38.2	115	3996.44	65.1%	4.5%	14.4
NPwoNP_m1_l>1	24259	33.2	103	3919.09	63.8%	4.4%	14.5
NPwoNP_m1_l>1_o	24604	33.7	106	3927.19	63.9%	4.6%	13.9
NP_l>1	27237	21.6	85	1475.91	24.0%	1.8%	13.6
NP_l>1_o	28151	22.6	82	1506.60	24.5%	1.9%	13.2
NPwoNP_l>1	21147	17.9	64	1350.85	22.0%	1.5%	14.5
NPwoNP_l>1_o	21493	18.4	64	1370.96	22.3%	1.5%	14.9
SingleT	6891	68.2	236	5991.76	97.5%	49.1%	2.0

Table 5. *Results of noun phrase extraction and network properties for twenty scenarios and the single term approach. [Data sourced from Thomson Reuters Web of Science Core Collection]*

phrases in the documents. As also observed with the file sizes, processing the complete noun phrases generates more data but when looking at the degree and density we don't see a large difference in network properties. When we apply sorting of the terms inside the phrases we detect a slight increase in the number of phrases that are included. Some of these sorted phrases are not excluded anymore if they appear on more than one paper. We find only very small differences in average degree and density between sorted and unsorted scenarios.

The largest effect comes from the selection of length of shingles. When no additional criteria are applied we observe an average degree and density close to the properties of the single term approach. Also when we select in each phrase only the individual terms complemented with the complete phrase (tagged with `_lm_l1`) we obtain similar results. The construction of these type of networks will suffer from the same computational complexity as the single term network. Next, the four scenarios where only the complete phrase is retained (`_lm`) have the lowest number of unique phrases but still have a density slightly above 57%. The fourth set of scenarios where the single term phrases and all shingles with length more than one are selected, have large sets of unique phrases and a density which is a bit higher than the third set of scenarios. Both the latter sets have density ratios between 13.9 and 15.1. It is clear that removing shingles with size 1 from longer phrases has a large effect on the density of the network. This effect is even more pronounced in the last set of scenarios where all shingles with size one are removed even when the noun phrase consists only of one term. This last criteria implies that many phrases are neglected. This is also reflected by the average number of phrases in the documents. Now the density also drops below 25%.

	SINGLETONS	WEIGHTED DEGREE			COMMUNITY DETECTION		
		AVERAGE	MEDIAN	MAX	NC	MOD.	<10
NP	5	86.59	86.85	168.07	7	0.102	1
NPo	5	85.35	85.60	167.04	8	0.102	0
NPwoNP	5	88.08	88.31	174.58	8	0.107	2
NPwoNP_o	5	87.38	87.63	166.82	8	0.108	2
NP_lm_l1	5	113.73	113.95	233.21	7	0.092	1
NP_lm_l1_o	5	112.83	113.04	233.26	6	0.920	0
NPwoNP_lm_l1	5	110.35	110.76	227.28	8	0.098	2
NPwoNP_lm_l1_o	5	109.64	110.13	227.08	7	0.097	2
NP_lm	24	47.44	46.29	123.49	8	0.107	1
NP_lm_o	24	46.81	45.56	122.78	7	0.107	1
NPwoNP_lm	24	48.69	47.27	140.19	8	0.108	1
NPwoNP_lm_o	24	48.18	46.76	139.71	8	0.108	1
NP_m1_l>1	13	34.23	33.32	101.48	10	0.147	2
NP_m1_l>1_o	13	33.79	32.86	100.13	9	0.147	2
NPwoNP_m1_l>1	13	37.46	36.42	110.33	10	0.147	2
NPwoNP_m1_l>1_o	13	37.24	36.19	109.51	10	0.143	2
NP_l>1	35	15.12	13.60	105.81	12	0.341	2
NP_l>1_o	34	15.32	13.81	103.87	12	0.330	2
NPwoNP_l>1	38	16.39	14.62	119.63	14	0.348	2
NPwoNP_l>1_o	38	16.64	14.86	118.50	12	0.338	2
SingleT	3	217.16	219.03	392.51	5	0.043	1

Table 6. *Weighted Degree of reduced network and Louvain Community Detection result [Data sourced from Thomson Reuters Web of Science Core Collection]*

Next singletons are removed from each of these networks, weighted degree is calculated and the Louvain community detection is used to cluster the documents. The results can be found in table 6. First we observe that the selection of type of noun phrases or the sorting of terms has no effect on any of the results in this table. From the difference between average and median of the weighted degree and the ratio between average and the maximum we learn that the first two sets of scenarios (no restriction on length of shingles or all one term singles complemented with the complete phrase) have a distribution that is close to normal just as the single term approach. Also the modularity coefficient is low with values around 0.1

In the next two sets of scenarios one can observe that the distribution of the weighted degree deviates from a normal distribution. The scenario, where only the complete phrase is, has also a low modularity. When we start to include all shingles with a length higher than one the clustering performs better. In the last set of scenarios where also the single term phrases are excluded we find the highest modularity and weighted degree distribution that is close to the distribution found in bibliographic coupling.

Based on these findings and together with the reduced data size for leaf noun phrase we chose the scenario with this tag: `NPwoNP_l>1_o` as the best possible approach. The sorted version is selected as it includes slightly more phrases than the unsorted version and solves the problem of multiple variation of the same concept.

HYBRID COMBINATION

In this section we compare seven hybrid combinations of the bibliographic cou-

WEIGHT Λ	WEIGHTED DEGREE			COMMUNITY DETECTION		
	AVERAGE	MEDIAN	MAX	NC	MOD.	<10
NPwoNP_l>1_o (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
0.66	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
BibC (λ= 1)	14.62	10.71	94.59	16	0.350	8

Table 7: *Results of hybrid clustering with different weight parameters [Data sourced from Thomson Reuters Web of Science Core Collection]*

	LEXICAL	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	
BibCpl	0.30	0.33	0.40	0.44	0.65	0.70	0.77	0.75

Table 8.*Value of Chi-Squared test between two weighting schemes [Data sourced from Thomson Reuters Web of Science Core Collection]*

pling component together with the selected NLP component. The different weights for the components are (0.125, 0.25, 0.33, 0.5, 0.66, 0.75 and 0.875). In Table 7 the results for the two uncombined components 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 7 clearly shows that the distribution of weighted degree is not distorted by any choice of the λ parameter. Also, for each of the chosen values a modularity above 0.3 is obtained.

When looking at the number of clusters, it evolves from 12 in the lexical component to 10 in the λ=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. Chi-square values are calculated between all schemes and plotted in Table 8.

APPLICATION

This section outlines briefly the results of our partitioning of the hybrid network (with 50% weight on both components) at three levels with increasing resolution (level I with resolution of 0.7; II with 1.0 and III with resolution 1.5). As mentioned above, we used a data set of 6144 publications 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

LEVEL I RESOLUTION = 0.7			LEVEL II RESOLUTION = 1.0			LEVEL III RESOLUTION = 1.5		
CLUS-TER	# PUBS	LABEL	CLUS-TER	# PUBS	LABEL	CLUS-TER	# PUBS	LABEL
1	957	Decision Support Systems	a	955	Decision Support Systems	1	807	Decision Support Systems
						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 2. *Cluster solution at three levels with increasing resolution [Data sourced from Thomson Reuters Web of Knowledge]*

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 8 papers were grouped in three pairs/triplets. These findings are summarized in Figure 2. 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).

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 chain*' (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. At a resolution of 1.5 the following picture emerges:

- *Strategic IS planning* with 1414 papers: can be subdivided in two large clusters on *strategic planning* (*III.5*) and *performance measurement* (*III.8*); 2 midsized clusters on *HR & Accounting* (*III.11*) and *ERP* (*III.18*); and a small cluster containing *executive perspectives* (*III.13*).
- *Development/OSS/Planning* with 1119 papers: The cluster '*Design science in IS research*' accounts for more than half of the papers and focuses on IT implementation and methods (III.12); another large cluster is centred on *Systems development projects* (*III.2*); a smaller cluster on *conceptual models* (*III.15*), with close to 50 papers, contains also bibliometric studies on, e.g., 'citation classics'.
- *Supply chain*: As one would expect the largest grouping is associated with *Supply Chain Management* (*III.10*), followed by *firm performance* (*III.14*) and *open source* (*III.19*); a small but still substantial cluster focuses on *outsourcing* (*III.9*), another small cluster can be linked to *innovation, assimilation and diffusion* (*III.6*).

The midsized Level II cluster on *intangible assets* can be further split up into *knowledge management* (*III.16*) accounting for most of the papers (367) and *customer relationship management* (CRM, III.22, 24 papers).

Clustering for the other Level I areas (*decision support systems, user and technology acceptance*) remains broadly the same. Only at level III with a resolution set at 1.5 one can identify a finer grained structure for these clusters:

- *Decision support systems* (*cluster III.1*) can be further differentiated from a cluster (III.7) around *communication and virtual teams*;
- *User satisfaction and service quality* (cluster III.3) are differentiated from *usage and acceptance of technology* (cluster III.4).

CONCLUSIONS

Based on the data presented in this paper we can conclude that the extraction of noun phrases from abstracts and titles can

improve the lexical component in the hybrid clustering. However, using the noun phrase itself is not sufficient for the improvement. The scenario, where the complete retrieved noun phrase is used, only reduces the density of the network but does not have an effect on the clustering afterwards. Only when the data is restricted to shingles with at least two terms constructed out of the noun phrases an improvement in the clustering is observed. This solution has several advantages over the other scenarios. It has a small set of unique terms, the density of the network is reduced to a quarter of the network constructed on single terms and the distribution of weighted degree is similar to the distribution in bibliographic coupling. As a consequence 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.

We even found out 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 only drawback for the application of a NLP-based procedure is that the intense preprocessing step requires a lot of computational power.

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, 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). Stanford Dependencies manual.

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. (2012). Using `core documents` for detecting and labelling new emerging topics. Scientometrics, 91(2, 399-416.

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

Glenisson, P., Glänzel, W., & Persson, O. (2005). Combining full-text analysis and bibliometric indicators. A pilot study. Scientometrics, 63(1), 163–180.

Jain A., Dubes R. (1988). Algorithms for clustering data, Prentice Hall.

Janssens, F., Glenisson, P., Glänzel, W. & De Moor, B. (2005). Co-clustering approaches to integrate lexical and bibliographical information. In Proc. of the 10th International Conference of the International Society for Scientometrics and Informetrics (ISSI) p. 284-289, Stockholm: Karolinska Univerity Press

Janssens, F., Leta, J., Glänzel, W. & De Moor, B. (2006). Towards mappin library and information science. Information Processing and Management, 42, 1614-1642.

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, pp. 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, August 15-17, 2013 Tartu, Estonia.

Noyons, E. C. M., & Van Raan, A. F. J. (1994). Bibliometric cartography of scientific and technological developments of an R&D field. The case of Optomechatronics. Scientometrics, 30, 157–173.

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

Salton, G. & McGill, M.J. (1986). Introduction to modern information retrieval. New York: McGraw-Hill, Inc.