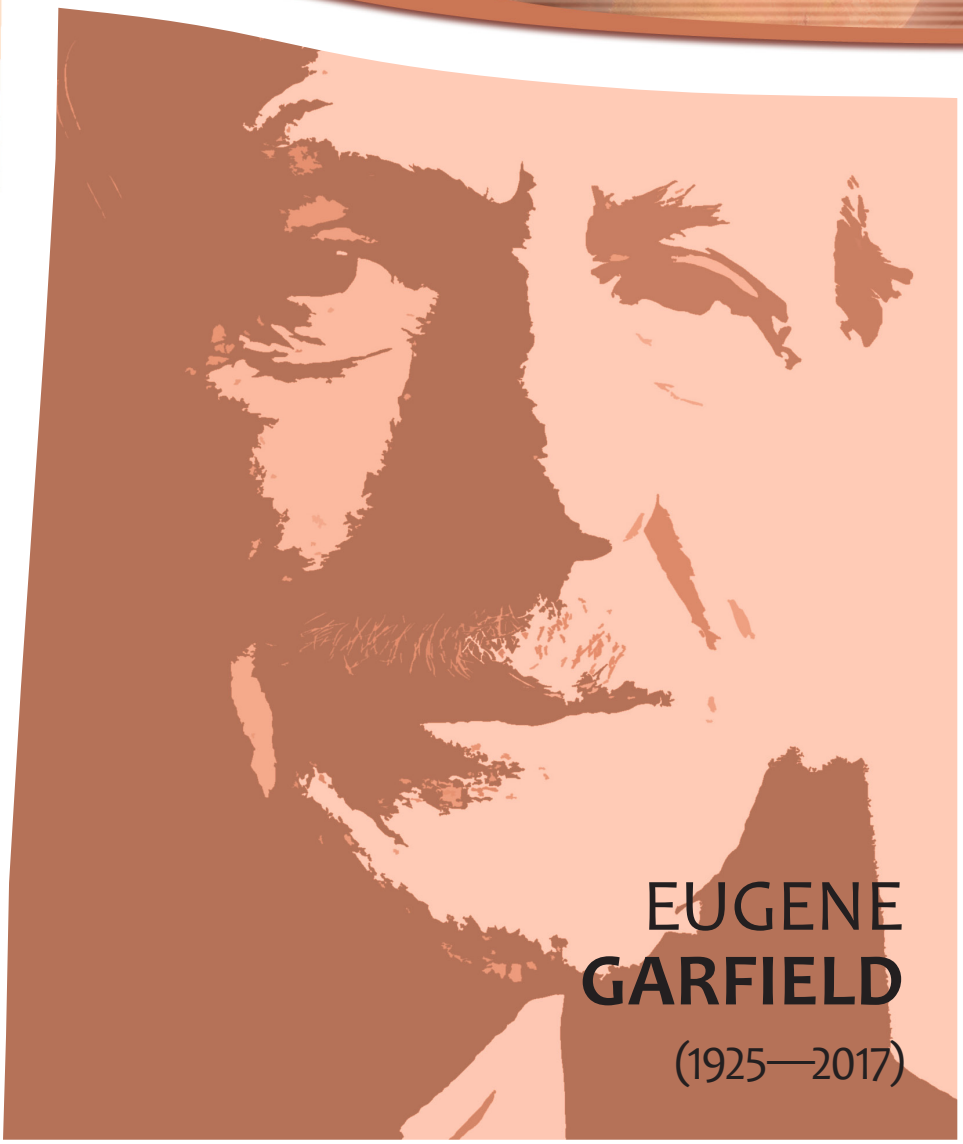
ISSI Newsletter

QUARTERLY E-NEWSLETTER OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS
ISSN 1998-5460

#49 / VOLUME 13 NUMBER 1
MARCH 2017

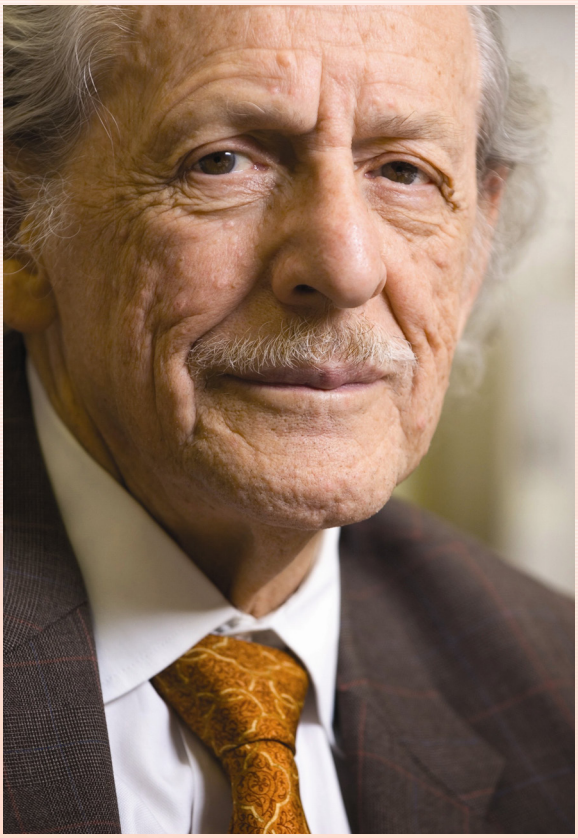
CONTENTS



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.

OBITUARY

EUGENE GARFIELD (1925—2017)



Photograph courtesy of © Meher Mistry

Dr. Eugene Garfield, founder of the Institute for Scientific Information (ISI) and one of the pioneers of scientometrics passed away unexpectedly on 26 February 2017 at the age of 91. The ISI in Philadelphia (PA, USA) was one of the first and most significant scientific information services and research centres in the world. Eugene Garfield created several bibliographic indexes and databases that form the fundament of our daily work. Without his ingenious innovation of citation indexing, the field of scientometrics, as we know it today, would not be imaginable. His industrious creative activity brought forth an oeuvre that embraces thousands of commentaries, pieces, articles, chapters and books. His famous collections of *Current Comments* and *Essays of an Information Scientists* stand just pars pro toto for his literary and scientific work. He devised the perhaps most widely used and debated scientometric product, the Impact Factor. His name is also linked with the development of many important ideas and concepts such as writing citation historiography, mapping the world of science and premature discovery or delayed recognition, just to mention three of them.

There is also a direct link with our Society: *The Garfield Dissertation Doctoral Scholarship* is donated by the Eugene Garfield Foundation to foster research in informetrics, including bibliometrics, scientometrics, webmetrics and altmetrics by encouraging and assisting doctoral students in the field with their dissertation research. The award helps doctoral students attend the ISSI biennial conferences. The price has been awarded at ISSI conferences six times since 2005.

Eugene Garfield was also one of the first Editors-in-Chief of the journal *Scientometrics* and the first recipient of the Derek de Solla Price Medal awarded by our journal. The editors of Scientometrics are organising a commemoration of his academic life and work and have invited researchers, including the Derek de Solla Price Award laureates of the journal Scientometrics and the previous winners of the Garfield Dissertation Doctoral Scholarship to contribute to the memorial issue. The issue is scheduled for one of the following volumes of Scientometrics.

With Eugene Garfield's passing, our scientific community has lost one of its true pioneers and best leaders.

Wolfgang Glänzel

STI 2017—PARIS

OPEN INDICATORS: INNOVATION, PARTICIPATION AND ACTOR-BASED STI INDICATORS

CONFERENCE CALL FOR PROPOSALS

6–8 SEPTEMBER 2017
ESIEE, PARIS
2 boulevard Blaise Pascal, Cité Descartes, BP 99, 93162 Noisy-le-Grand Cedex

http://sti2017.paris

The 2017 STI conference addresses the new issues and challenges that have appeared in Science, Technology and Innovation indicators. We are witnessing sharp changes in the recent years: new areas of knowledge are appearing, new types of objects need to be taken into account, new methodologies and visualisations have been proposed, a combination of different policy interests emerging from a large variety of social actors modify the demands addressed to indicators. Most of these challenges relate to profound changes in the way science, technology and innovation relate to society; indicators — necessarily — reflect these changes, taking into account the needs and strategies of the many different actors involved.

The conference will be the opportunity to showcase results from the intense work done in recent years on the way science, technology and innovation indicators are used in relating social actors to science and technology. The range of relations between science and society has been expanded into a dizzying array of forms that include a large variety of knowledge producing activities such as: open science and open innovation, collaborative projects that include social actors, crowdsourcing in large digital platforms, the inclusion of local and "indigenous" knowledge in development programmes, closer connections between users and producers of scientific and technological devices, a growing digitalisation, a shifting balance between productive and 'access to market' activities, new developments such as the sharing economy, crowdfunding, responsible innovation, technology 'makers' and 'do-it-yourself' movements... They drive to new governance issues, but also to new requirements for indicator designers and multiple experiments that the conference should discuss.

Participatory research programmes, combined with active civil society organizations, promote a need for debates and more democratic decision-making processes. Expertise can no longer be lim-

Photographs: courtesy of © Balázs Schlemmer / ::schlemmerphoto.com::

ited to top-down application of scientific knowledge and indicators should reflect and contribute to this democratic move. How can and do indicators get involved into this democratic move? How is this enlarged participation of actors into the definition and shaping of indicators changing the modalities of their construction?

Moreover, following last year's central theme on peripheries, STI 2017 will interrogate broadly the evolving geography of ST&I, the impacts of the concentration of 'new dominant' sciences in large metropolitan areas, the expanding and diverse forms of international collaborations, all of which impose both methodological developments and new global strategies.

These objectives include new methodological developments, new methods in data processing, sharing, analysis and use, including the management of large data in a large variety of forms. Indicators, today, require not only larger databases, but also the mastering of shared technologies and collaborative technologies. Of interest are the possibilities of enlarging indicators through the use of open data.

The Conference will thus propose to engage in stimulating exchanges around these new developments concerning actor-based indicators, in a large variety of sectors from scientific and technological production, to innovations in service sectors such as tourism, leisure and culture, health, ageing, or food catering, to non-technological and organizational innovations. It will open the debates on the democratic uses of STI indicators and the specific challenges participation of a wider range of actors pose to the construction of sound, meaningful and robust indicators.

As last year in Valencia, the conference will include special tracks:

1. Data infrastructures & data quality for evolving research metrics
2. Innovation benchmarking and indicators
3. Actor-based and location-based innovation indicators: the evolving knowledge landscape
4. Measuring impact and engagement
5. Collaboration, mobility and internationalization

6. Social sciences and humanities
7. Peripheries and frontiers in science, technology and innovation

Other topics include:

1. Social media and alternative metrics.
2. How do indicators shape research agendas?
3. Evaluation of mission-oriented research.
4. Science participation and communication.
5. Inclusive innovation and grassroots innovation.
6. Indicators for sustainable development in socio-economic transitions.
7. Gender and gendered research and innovation.
8. Innovation, creativity and culture
9. Innovation in 'person-based' services (health, culture, leisure, tourism).

The conference will consider presentations on the above topics, or other specific topics, but will focus on those related to its general theme of "Open indicators".

The Conference will be used as a platform to present and discuss the results of the EU-funded "Research infrastructures for the assessment of science, technology and innovation policy" (RISIS). RISIS is developing sets of STI indicators to be openly accessible (http://risis.eu). The open access CORTEXT platform, a unique tool designed for textual analysis, managed by IFRIS will also be presented.

SUBMISSION GUIDELINES

All papers must be original and not simultaneously submitted to another journal or conference. The following paper categories are welcome:

- Short paper (max 3,000 words) with a description of a completed study
- Research in progress paper (max 1,500 words)
- Posters (max 1,000 words) with the main content of the poster/study reported. We very much encourage posters. The Venue will permit a very good series of posters sessions and they can be the opportunity of very intense and intellectually stimulating exchanges.

All proposals should be made through EasyChair.org, with an abstract (up to 500 words). Easychair: https://easychair.org/conferences/?conf=sti2017

Full length papers should be uploaded on EasyChair.org (usually a pdf file).

Templates for the full length papers are provided on demand at: contact@sti2017.paris. They are also available at the website of the conference: sti2017.paris.

You can submit proposals for Special sessions: proposal of 90 or 180 min. panel discussions, round tables or a coherent set of papers (2,000 words max.), by sending a request at the address: contact@sti2017.paris

All submissions (except for sessions) are to be made through EasyChair.org

SUBMISSIONS

Deadline: April 19
Easychair: https://easychair.org/conferences/?conf=sti2017

You can copy/paste title, names of authors and affiliations, and abstract within Easychair and upload a pdf document of the paper itself.

Submissions can be on the Easychair Call for proposals:

1. Short paper (max 3,000 words) with a description of a completed study
2. A research in progress paper (max 1,500 words)
3. A poster (max 1,000 words) with an abstract of the study. Poster are strongly encouraged. Opportunities will be given for both poster sessions and permanent exhibition of posters.
4. Special session: can be proposed of 90 or 180 min. panel discussions, round tables or a coherent set of papers (2,000 words max.), by sending a request at the address: contact@sti2017.paris

Website of the Conference: sti2017.paris
Enquiries at: contact@sti2017.paris

THE CHALLENGE OF COMBINING VARIOUS DATA SOURCES FOR BIBLIOMETRIC USE

MEHMET ALI ABDULHAYOGLU
ECOOM, KU Leuven, Belgium

1. INTRODUCTION

Scientometrics emerged as a discipline monitoring and measuring research literature quantitatively. The field achieved success in basic sciences and with the demand from applied and social sciences in time, its applications have been extended. The bibliographic databases (BDB), such as Clarivates Web of Science (WoS) or Elsevier's Scopus which are multidisciplinary, dynamic and closed DBs, are mostly used as data sources. This paves the way standardized, compatible, reproducible and documentable studies. However, while the field has turned into a tool for research evaluation and assessment, this has introduced some limitations and challenges. For example, citation patterns differ significantly between disciplines hindering direct comparisons or even the most extensive traditional BDBs do not cover all research areas. Additionally, one metric or one source is mostly not enough to evaluate the quality of work. Glänzel and Debackere (2003) have pointed to some limitations in this context.

Other than the above-mentioned limitations, we should point out the challenge introduced by internet which has changed the scholar communication significantly. To this end, web-based tools have been developed for specific needs. However, many times they do not meet the diverse demands. Unlike the BDBs and their strengths as given above, such internet sources may lack proper documentation, clean data and reproducibility. Therefore, this brings another challenge

to the field due to the presence of myriad number of external sources. On the other hand, although traditional BDBs provide detailed information, the data quality may not always meet the standards needed for analysis at all level of aggregation. Furthermore, today scientometrics focus on not only "communication among researchers" but also modelling and measuring the effect of science outside research communities which is called as "Scientometrics 2.0" (Priem and Hemminger, 2010). Considering the challenges the field has been and will be facing and the continuous growth of bibliographical information in the Web, combining traditional BDBs with external sources promises a lot. For example, Hoffmann et al. (2014) state that "*Traditional impact measures based on bibliographic analysis have long been criticized for overlooking the relational dynamic of scientific impact.*" The authors apply a dataset from a social media platform and show that online communication activity information may enrich established impact measures.

Besides supplementary data are mainly available from the Web or from another BDB, a non-bibliographic source can be exploited for bibliographic studies as well. On the other hand, some external data, such as publication lists (PL) or CVs of authors containing bibliographic references may be desired to be searched within a BDB. In both cases combining different sources introduce a challenging task. That is, combining data and matching individual records on the basis of specific fields and components cannot be done manually nor with bibliometric methods alone. Substantial support from computer science is needed as well to solve this task. Furthermore elaborated concepts for data integration and harmonization including their technical aspects are needed (Daraio & Glänzel, 2016; Glänzel & Willems, 2016).

In what follows, I will give an overview of the systems we have developed at ECOOM and my work done in the framework of my PhD project in order to contribute to the solution of some of the above-mentioned issues. In this context, we use WoS as the main source and split the study into two parts. In the first part, we deal with searching external sources within WoS. For this part, we devote two sections where the first one presents a text matching system to identify references from PLs or CVs within WoS while the second one focuses on text matching issue for larger data sets; that is, BDB overlapping on paper level.

In the second part, we work the other way around, that is, identifying the papers indexed in WoS in external sources. As mentioned above, such initiatives can be very valuable and necessary since WoS or other prominent BDBs cannot index all the details about the papers although they provide very detailed information. Two sections are devoted for this part. The first addresses the author name disambiguation (AND) by combining WoS and an academic social platform, namely, Researchgate (RG). The second focuses on accessing *introduction* and *conclusion* sections of the papers indexed in WoS via Crossref and processing this external text information to provide more insight into science maps.

I would like to stress that these are not the only solutions to above-mentioned issues and many other approaches are imaginable and possible. And there is, of course, no *best* solution. Finally, the practical applicability in the light of further development and challenges will give evidence of the usefulness of the methods described below.

2. IDENTIFYING EXTERNAL SOURCES IN WoS

2.1. SEARCHING FOR REFERENCES FROM PLs

The scientific world has been witnessing a rapid and continual growth in every province of its literature. Szalay & Gray (2006) had stated that "*the amount of scientific data is doubling every year*". The

massive increase in scientific data volume has inevitably caused BDBs to expand significantly in terms of number of indexed publications. This has brought the challenges in information retrieval from the BDBs. Since bibliometrics serve as a tool for benchmarking and evaluating research performance (Glänzel and Schoepflin, 1994), information retrieval from the prominent BDBs, such as WoS, plays an important role for governmental, academic or business related applications (Fisher et al., 2013). For such applications, mostly PLs or CVs of the applicants are provided and the bibliographic references in the lists are searched within a BDB. Such retrieval procedure must be as meticulous as possible and hence requires a heavy manual effort. Although it is unrealistic to replace all the manual work, suggesting matching pairs based on a text matching procedure may save the time and labor significantly which constitutes our first chapter.

Considering the above-mentioned challenge, we developed a short text matching system based on character n-grams (Abdulhayoglu et al., 2016). Character n-grams are the decomposed successive components of a text. For example, the character 3-grams *__g _gl glä län änz nze zel* are the decomposed chunks of the text "*Glänzel*". Unlike word n-grams, they have lots of advantages when erroneous or misspelling texts are present. For example, when matching "*Glänzel*" vs. "*Glanzel*" using the word system, a similarity score of 0 will be returned due to the character "ä". On the other hand, when the same task is repeated with character 3-grams, a Salton similarity score of 0.57 or an edit distance based similarity score of 0.85 (Kondrak, 2005) is returned due to the common chunks, such as *__g _gl nze zel*.

In our paper, we present an optimum setting by trying different character n-gram sizes and threshold values. In addition, we create a baseline using word unigrams to make a comparison between two systems. We show that character 3-grams and a cosine similarity score of 0.60, above which is accepted as a correct match, present the best results for our dataset. As a result, we obtain an accuracy of 96.0% and 94.7% for our character and word based systems, respectively. Our approach proves decrease in manual work and speed up the information retrieval task.

2.2. BDB OVERLAPPING

Although we demonstrate the success of our matching procedure for the desired information retrieval study in the previous chapter, it is incapable of dealing with larger datasets due to its high complexity, that is, $O(kn^2)$ which indicates how running time grows depending on input size where k and n refer to number of features and sample size, respectively. In other words, we were previously able to match about 10.000 references with WoS records in a manageable time. However, when this number was increased to 50.000, it took about 10 hours to return the matching results. If millions of records were in question, the system would be useless.

One important task in the field is measuring the overlap between BDBs (Gavel & Iselid, 2008; Bosman et al., 2006) due to some indications of BDB overlapping which include the cost issue as the most important one amongst others. Hood & Wilson (2003) give a summary of the issue in detail. So this means that a system matching millions of records with other millions needs to be developed if a paper based comparison is to be conducted. This is indeed a harsh challenge which led the researchers to use only journal information from the BDBs rather than employing papers. However, as Pao (1993) states, two BDBs may index different publications from an identical journal issue and hence paper based comparison may yield more detailed and reliable results or at least confirm the previous literature works. In this context, we devote the second chapter to this high volume matching challenge as a complementary to the previous approach employing WoS and Scopus. Today, thanks

to the advancements in big data world and thus in distributed computing, such previously tedious tasks are now possible.

For our application, we use LSH algorithm which fulfills the text matching process with a significantly lower complexity ($O(kn)$) unlike our previous approach (Ravichandran et al., 2005). In our case, for a WoS record, it first finds some of its neighbors from Scopus DB instead of searching the entire Scopus. This approach offers approximate results and thus may sometimes miss the correct matches. However, this is the trade-off between the accuracy and the running time of the algorithm. Therefore, once we obtain matching suggestions, we follow some rules to retain the identical matches because a match with a very high similarity score does not always guarantee the identical match. For example, a suggested pair is retained as identical if both papers have identical *journal names* or *ISSN numbers*, *issue numbers*, *volumes*, *begin pages*. The more details can be seen in our paper which will be published in *Lecture Notes in Computer Science* soon.

As for the data and the results of the application, 1.6 & 2.2 million papers, published in 2011, are selected from WoS and Scopus, respectively. From our experiments, we show that matching suggestions can be obtained in less than one hour. Furthermore, our best matching results show that at least 70% of WoS records is also indexed by Scopus which seems to be in line with previous similar literature works. Furthermore, since we match two BDBs on paper base, we can also say something about the cited papers. The number of papers, published in 2011 and cited at least once in WoS, is 1,127,239 and 1,002,478 (88.93%) of them are also indexed by Scopus. Moreover, when highly cited publications are in question, similar results were obtained. For example, there are 304 and 9,652 publications cited more than 500 and 100 times, respectively where 264 (86.84%) and 8,741 (90.56%) of them are also found to be indexed by Scopus.

3. FINDING WoS PUBLICATION IN EXTERNAL SOURCES

So far, we have explained two cases where external data is desired to be found in WoS. From this section, we explain two opposite cases in which WoS records are to be found in some external sources to enrich or confirm the results which are based only on WoS records.

3.1. AUTHOR NAME DISAMBIGUATION (AND) WITH THE HELP OF EXTERNAL SOURCES

In this section, we study AND which is a very important task for information science fields and bibliometrics as well. Especially for individual level analysis, detecting true authors for each publication is crucial. This task is quite challenging especially for common names like Chinese names (for example, *Wang* surname appears more than 800,000 times in our WoS DB.). Although WoS provides common metadata, such as *author name, affiliation, address*, *co-authors* and more detailed and distinguishing info, such as *e-mail address* and *researcher ID*, it may be still inadequate. For example, only 13.79% and 8.57% of the authors in the entire WoS have either an e-mail or RID assigned, respectively. To tackle this lack of more distinguishing information, we try to leverage some external sources, that is mainly RG and the authors' web pages as a complement to RG.

In the literature, there are some works exploiting external sources for AND (Kanani et al., 2007; Yang et al., 2008). However, they all apply Google's Custom Search Engine service (CSE), which allows the user to issue queries programmatically. This approach is limited today in two ways. First, the number of free query searches is bounded to 100 per day. Second, if the aim is to access and retrieve data from an author's web page, it will be too costly to implement the approach for each different web page having different designs. There-

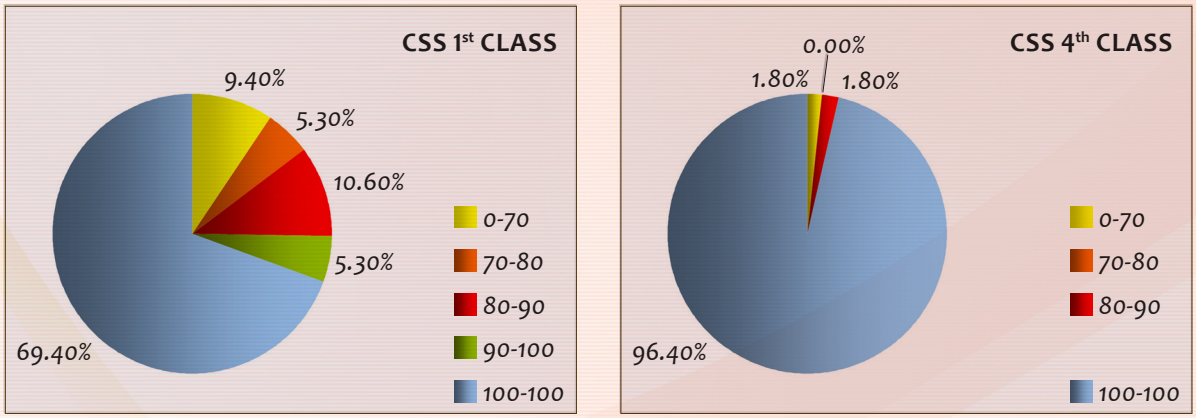


Figure 1. Joint results from CC clustering – RG&CSE when they are compared to manually validated data in the first and fourth CSS-class.

fore, we use this approach as a complementary to our RG procedure dealing with relatively smaller datasets.

RG is an academic social platform where publication lists of authors can be found in their corresponding RG profile pages. In our paper (Abdulhayoglu & Thijs, 2017), we leverage this structured source and present a system comprising three stages. In the first stage, using only the available data in WoS, we apply a mix of supervised and a network based clustering techniques (Connected Components (CC)) to create set of papers each of which presents one author. In the second stage, we search the papers from the formed sets in RG to see if they come from the same RG author profile page to confirm cluster results. In the final stage, we try to confirm the papers which cannot be found in RG.

We apply a data set of 10,940 publications with their corresponding 31,983 co-authors indexed in WoS. The first stage forms 22,120 author clusters and 16,900 (76.4%) of them can be confirmed by RG in the second stage. However, we notice that most of the clusters are the singletons, that is, they contain only one member. This does not make sense for bibliographic analysis. Therefore, we retain only the clusters having at least 10 members which result in 183 clusters. For this data, the clusters, formed by stage 1, have an F score of 0.955. When the clusters are confirmed through RG and CSE stages respectively, we observe an F score of 0.947. It is promising that almost the same accuracy is obtained with the confirmed data.

Despite promising results, there are some issues need to be mentioned. We see that authors are prone to show their selected or high impact papers either on RG or in their web pages. Indeed, this causes the correct papers in the clusters to be removed since they do not appear in RG. Figure 1 summarizes this fact with respect to citation behaviors of confirmed publications. To produce the figure, we detect poorly and highly cited publication classes for each cluster by means of Characteristic Scores and Scales (CSS) methodology developed in our institute. CSS is a method classifying observations into self-adjusting categories and can be applied to assess the eminence of citation impact (Glänzel & Schubert, 1988; Glänzel et al., 2014). So, in the figure, CSS 1st and CSS 4th classes mean the poorly and highly cited categories, respectively.

In the figure above on the right hand side, it tells that for 96.40% of the authors, we confirm all their highly cited publications. This ratio drops (69.40%) significantly when the poorly cited publications are in question. This tells us that only relying on RG or author web pages may mislead the bibliometric results. For example, relative indicators are calculated by dividing the number of citations by the number of publications where relying only RG confirmation will weigh highly cited papers more and thus overestimate the met-

Figure 2. Pruning incorrect cluster members from a formed cluster through RG&CSE

ric. As a future work, our procedure can be applied only to those authors having very common surnames. Figure 2 shows an example to this case. As seen, 9 authors are gathered in one cluster in which 7 are incorrect. Through our confirmation procedure those incorrect ones are pruned.

3.2. ENRICHING SCIENCE MAPS THROUGH EXTERNAL SOURCES

In our last chapter, we address another important issue which is drawing science maps exploiting full text and especially *introduction* and *conclusion* sections which are not indexed in WoS. In the literature, there are abound valuable studies on science maps (Glänzel & Thijs, 2011; Boyack & Klavans, 2010). Such studies have already proven the success of hybrid methods that is the use of bibliographic coupling along with textual information mostly from titles and abstracts. Additionally, in their recent study, Thijs et al. (2015) improve the hybrid clustering results by means of Natural Language Processing (NLP) applications.

Being aware of such successful applications, we study a dataset in which papers have almost no strong reference links thus very low bibliographic coupling. Therefore, the science maps will rely on textual information. In this connection, the literature says that introduction and conclusion sections are the other essential parts of the scientific papers (Shah et al., 2003; Galeas et al., 2009). Motivated by those findings and very few studies having used section information, we devote this last section to produce science maps for a medical field by using introduction and conclusion sections along with titles and abstracts.

To access the section information, we apply a free API service (Text and data mining (TDM)) from Crossref which is an association bringing together more than 4,000 publishers, such as Elsevier, Wiley, Springer, Taylor & Francis etc. Using Digital Object Identifier (DOI) numbers from WoS in the service, we access 1.082 publications from the field of *Integrative & Complementary Medicine*. All the accessed papers are published by Hindawi Publishing Corporation (HPC) which presents them in the XML format which is very structured and easy to extract the texts from the desired sections.

We process the extracted texts following the suggested NLP procedure by Thijs et al. (2015). In addition, we also extract the *Medical Subject Headings* (MeSH) terms from the texts automatically by applying an API provided by NIH (https://ii.nlm.nih.gov/MTI/index.shtml) and redraw the maps. VOSviewer tool (van Eck & Waltman, 2009) is used to produce the maps. The science maps produced based on NLP and MeSH terms are in line and present almost the same topics. The observed topics are twofold. On the one hand, they present what kind of alternative medicines are applied to treat the diseases. The most salient one is *Traditional Chinese Medicine* and its derivatives. Additionally, Korean medicine (Sasang constitution) is observed especially for heart related diseases. Finally, India originated treatments such as yoga or meditation etc. are observed. On the other hand, what kind of diseases are in question is presented. They are heart related diseases, cancer, obesity and diabetes, inflammatory and bacterial diseases, wounds, brain related diseases, gastric diseases, liver related diseases, rheumatoid arthritis, osteoporosis.

We observe that the textual information from titles and abstracts are quite adequate to detect the topics. That is, when using in-

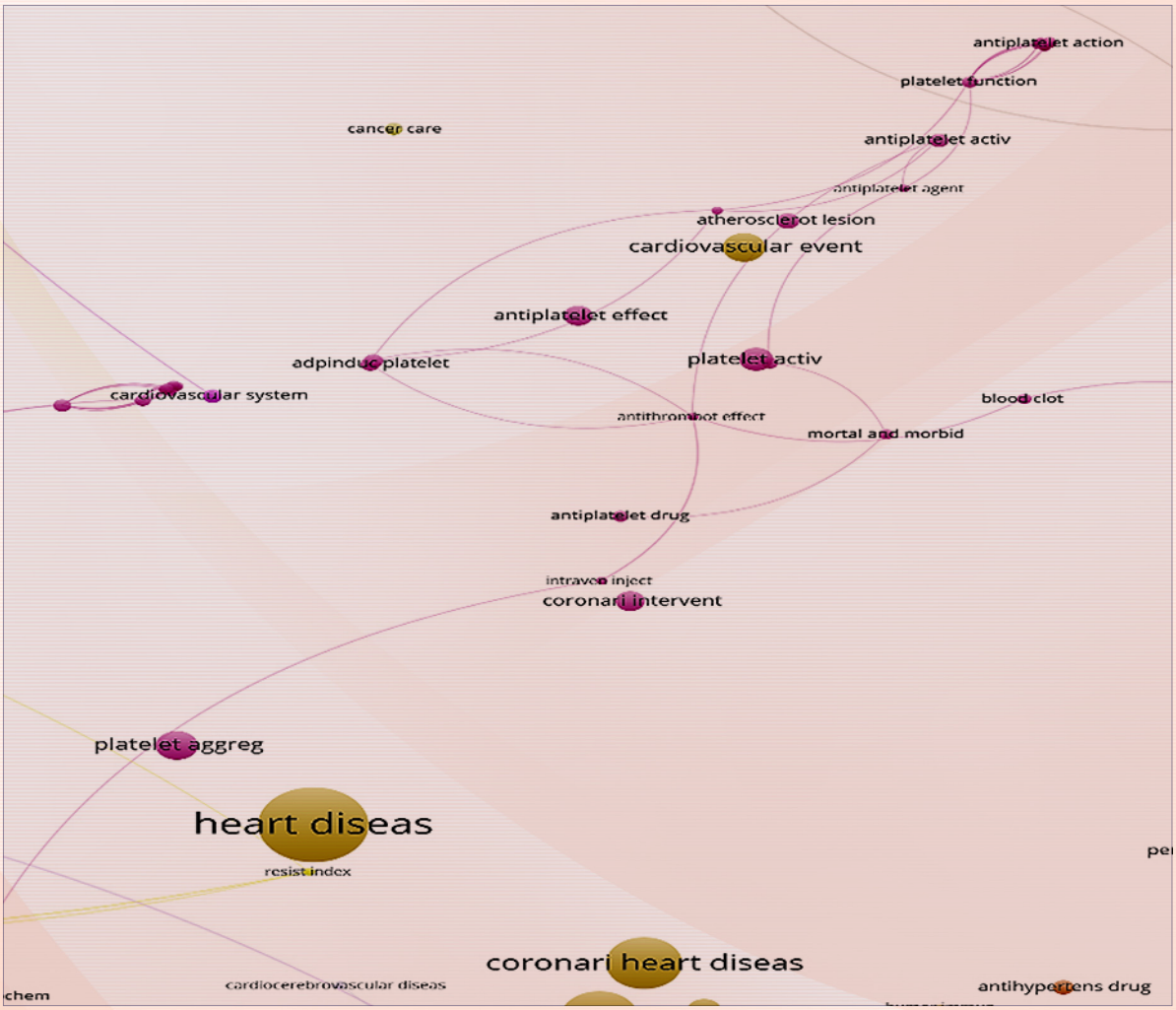


Figure 3. A part of a cluster (cardiovascular and heart diseases) based on the co-occurrence of NPs extracted from titles, abstracts, introductions and conclusions

troduction and conclusion additionally, we do not observe a new topic. In a way, this is supportive to the hybrid methods using title and abstract information on their text side. However, when we check the topics in detail, we observe some interesting results from the maps drawn based on all internal and external data. For example, in the cardiovascular diseases cluster, we see many terms related to *platelet aggregation*, *anti-platelet therapy*, *anti-platelet activity* etc. These terms appear together with other terms from the cluster, such as *cardiovascular diseases*, *cardiovascular system*, *cardiovascular event* etc. more frequently due to included terms in the introduction section. Figure 3 shows that specific cluster part.

Such detail is not present when only abstracts and titles are applied. The main reason for this is that the use of an alternative drug may be mentioned for a specific disease in the abstract as the main topic of the study, whereas the drug's or related components' effects on other diseases may be mentioned in introduction as background information. We observe similar cases in about ten publications in our dataset. For example, Chen et al. (2012) study liver microsomes when using a Chinese drug. The authors give some more general details about the medicine in the introduction. As a result, more term pairs, which do not exist in the abstract and the title, can be retrieved including for example:
cardiovascular disease - platelet aggregation
cardiovascular disease - antiplatelet effect
cardiovascular outcome - antiplatelet effect
cardiovascular outcome - platelet aggregation

In the maps drawn based only on abstract and title data, we can see the platelet aggregation term very close to wound healing related terms which makes sense due to its important role in coagulation. On the other hand, there is no clear link for the relation between platelet aggregation and heart failure or cardiovascular diseases. Please note that this study will be submitted to a journal in the medical field.

CONCLUSION

In my PhD, I study the interaction between WoS and external sources to either tackle some challenges in information retrieval from WoS or enriching WoS sources. In the era of information, more similar applications can be carried out thanks to abound bibliographic sources. I show that combining different bibliographical sources are possible and they may produce quite valuable results for not only bibliometricians but also the researchers from other fields such as medical fields.

REFERENCES

Abdulhayoglu, M. A., & Thijs, B. (2017). Use of Researchgate and Google CSE for author name disambiguation. *Scientometrics*. DOI: 10.1007/s11192-017-2341-y, in press.

Abdulhayoglu, M. A., Thijs, B., & Jeuris, W. (2016). Using character n-grams to match a list of publications to references in bibliographic databases. *Scientometrics*, 109(3), 1525-1546.

Bosman. J., Mourik, I. V., Rasch. M., Sieverts. E., & Verhoeff. H. (2006). Scopus reviewed and compared: The coverage and functionality of the citation database Scopus. including comparisons with Web of Science and Google Scholar. *Utrecht University Library*.

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

Chen, H., Zhang, Y., Wu, X., Li, C., & Wang, H. (2012). In vitro assessment of cytochrome P450 2C19 potential of Naoxintong. *Evidence-Based Complementary and Alternative Medicine*, 2012 Article ID 430262.

Daraio, C., Glänzel, W., (2016). Grand challenges in data integration –state of the art and future perspectives: an introduction. *Scientometrics*. 108(1), 391-400.

Fisher, J., Wang, Q., Wong, P., & Christen, P. (2013). Data cleaning and matching of institutions in bibliographic databases. *Organization*, 238, 99–103.

Galeas, P., Kretschmer, R., & Freisleben, B. (2009). *Document relevance assessment via term distribution analysis using Fourier series expansion*. In: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, 277–284. ACM.

Gavel. Y. & Iselid. L. (2008). Web of Science and Scopus: a journal title overlap study. *Online information review*. 32(1). 8–21.

Glänzel, W., Debackere, K., (2003), *On the opportunities and limitations in using bibliometric indicators in a policy relevant context*. In: R. Ball (Ed.), Bibliometric Analysis in Science and Research: Applications, Benefits and Limitations, Jülich.

Glänzel, W., Thijs, Debackere, K., (2014). The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment. *Scientometrics*. 101(2), 939-952.

Glänzel, W., Schubert, A., (1988). Characteristic Scores and Scales in assessing citation impact. *Journal of Information Science.* 14, 123–127.

Glänzel, W., & Schoepflin, U. (1994). Little scientometrics, big scientometrics... and beyond? *Scientometrics*, 30(2), 375–384.

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

Glänzel, W., Willems, H., (2016). Towards Standardisation, Harmonisation and Integration of Data from Heterogeneous Sources for Funding and Evaluation Purposes. *Scientometrics*. 106(2), 821-823.

Hoffmann, C.P., Lutz, C. & Meckel, M. (2014), *Impact Factor 2.0: Applying Social Network Analysis to Scientific Impact Assessment*. 47th Hawaii International Conference on System Sciences. 1576-1585. DOI: 10.1109/HICSS.2014.202

Hood. W. W. & Wilson. C. S. (2003). Overlap in bibliographic databases. *Journal of the American Society for Information Science and Technology*. 54(12). 1091-1103.

Kanani, P. H., McCallum, A., & Pal, C. (2007). *Improving Author Coreference by Resource-Bounded Information Gathering from the Web*. In: R. Sangal, H. Mehta, R.K. Bagga (eds), IJCAI, 429-434.

Kondrak, G. (2005). *n-Gram similarity and distance*. In: M.P. Consens, G. Navarro (eds), String Processing and Information Retrieval, 12th International Conference, SPIRE 2005, Buenos Aires, Argentina. Lecture Notes in Computer Science 3772, Springer, 115–126.

Pao. M. L. (1993). Term and citation retrieval: A field study. *Information Processing & Management*. 29(1), 95-112.

Priem, J., Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*. DOI:10.5210/fm.v15i7.2874.

Ravichandran. D., Pantel. P. & Hovy. E. (2005). *Randomized algorithms and NLP: using locality sensitive hash function for high speed noun clustering*. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics.622-629.

Shah, P. K., Perez-Iratxeta, C., Bork, P., & Andrade, M. A. (2003). Information extraction from full text scientific articles: Where are the keywords? *BMC bioinformatics*, 4(1), 20.

Szalay, A., & Gray, J. (2006). 2020 Computing: Science in an exponential world. *Nature*, 440(7083), 413-414.

Thijs, B., Glänzel, W., & Meyer, M. (2015). *Using noun phrases extraction for the improvement of hybrid clustering with text-and citation-based components. The example of "Information System Research*. In Proc. of the Workshop Mining Scientific Papers: Computational Linguistics and Bibliometrics, 15th International Society of Scientometrics and Informetrics Conference (ISSI), Istanbul, Turkey. 28-33. Available via: http://ceur-ws.org.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer: A computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.

Yang, K. H., Peng, H. T., Jiang, J. Y., Lee, H. M., & Ho, J. M. (2008). *Author name disambiguation for citations using topic and web correlation*. In: B. Christensen-Dalsgaard, D. Castelli, B. Ammitzbøll Jurik, J. Lippincott (eds), Research and advanced technology for digital libraries. Springer Berlin Heidelberg, 185–196.

DRAKKAR: A GRAPH BASED ALL-NEAREST NEIGHBOUR SEARCH ALGORITHM FOR BIBLIOGRAPHIC COUPLING

BART THIJS
KU Leuven, FEB, ECOOM; Leuven; Belgium
bart.thijs@kuleuven.be

Abstract. Drakkar is a novel algorithm for the creation of bibliographic coupling graphs in huge document spaces. The algorithm approaches this as an All-Nearest Neighbour search problem and starts from a bipartite graph constituted by the citing publications and the cited references and the directed citations connecting them. The approach is inspired by dimensionality reduction techniques like Random Projection and Locality Sensitive Hashing which use global random functions for dimension or feature selection. The proposed algorithm enables the use of local selection functions at the level of the individual nodes. For the particular case of bibliographic coupling the selection functions are based on the boat-shaped information distribution associated with the indegree of the cited references. This distribution resembles the typical symmetrical shape of a Viking ship (called 'Drakkar' in Dutch, hence the name). An experiment with several different random functions reveals that focussing on the end of the distribution related to the references with low indegree results in a graph with accurate strong links but many false negatives while the other end of the distribution can detect most links but underestimates the strength of the link. The algorithm is implemented in GraphX, the library for distributed graph processing within Spark. It is using Pregel's messaging framework.

Keywords: Nearest Neighbour Search, Bibliographic Coupling, GraphX, Pregel, Bulk Synchronous Parallel

INTRODUCTION

Since its introduction more than fifty years ago by Kessler (1963) bibliographic coupling has been used in numerous applications and studies. It has a proven track record for document clustering and for retrieval and subject delineation purposes. However, in the advent of large bibliographic databases covering several millions of documents and the growing focus on the creation of global document networks (Klavans & Boyack, 2011),

the application of bibliographic coupling is lacking behind. An important challenge for large scale application of bibliographic coupling (BC) in global clustering exercises or large domain studies is the computational and storage resources required for the creation of such BC-networks. Depending on the chosen representation of the underlying data, different strategies for the calculation of the cosine similarities can be applied.

In a relational database one could store citing publication-citing references pairs and use a query that joins such a table with itself on the cited references. An aggregate function can then count the number of joint references for each publication-publication pair.

Alternatively, data can be stored in a (sparse) matrix representation of a document-feature space where the cosine similarity is based on the dot product of this matrix and its transposed. But without any optimisation this calculation would take up to $O(n^2m)$-time for n documents and m features (cited references). Using dimensionality reduction techniques like PCA or SVD could reduce the computational complexity by lowering the m-factor but not without an additional cost as these techniques imply a matrix decomposition which would require substantial computation time even when using an iterative implementation. Based on the Johnson–Lindenstrauss lemma (Johnson & & Lindenstrauss 1984), Random Project reduces the high dimensional space to a subspace with much lower features while preserving the distance between documents. However, such a dimensionality reduction does not eliminate the n-by-n document comparison and implies new projections whenever new documents that extend the feature space are added.

The first problem of the n-by-n comparison can successfully be solved by the application of Locality Sensitive Hashing which is a common technique applied in record linkage problems (eg. Karapiperis & Verykios, 2016). LSH uses several random hashing functions for mapping with high probability documents with a great similarity into the same buckets (see Ravichandran et al 2005 or Rajaraman & Ullman 2010). Documents that often co-occur in these buckets have a high likelihood to be similar and the number of pairwise cosine calculation can thus be drastically reduced by limiting it to those document pair with high co-occurrence. The cosine similarity can be approximated based on the number of co-occurrences in buckets.

The application of LSH for bibliographic coupling comes with two main drawbacks that have some substantial consequences on its applicability. LSH does not solve the issues that are confronted when extending the document-feature space. New hashing functions have to be created and all documents have to be assigned to new buckets in order to be able to calculate the similarity with documents in the prior set and the newly added ones. The second drawback is related to the existence of false positives. Given the extremely sparse nature of bibliographic coupling it is quite likely that the set of hashing functions only selects those features that are absent in a large set of papers which do not share any reference. Consequently, these papers are all assigned to the same buckets despite the distance between them. This can only be solved by increasing the number of hashing functions, by increasing the dimensionality of the functions or by avoiding the approximation of the cosine similarity by actual calculating this value. Each of these solutions come with a substantial computational cost.

This paper takes an alternative approach by exploiting the properties of a graph representation of the underlying citation data. The creation of a bibliographic coupled network can be considered as an all-nearest neighbour search (ANNS) problem in a huge feature space represented as a bipartite graph.

MESSAGE PASSING

The proposed algorithm is based on the Pregel messaging framework developed at

Google (Malewicz et al. 2005). This framework builds on the Bulk Synchronous Parallel model (Valiant, 1990) by implementing a sequence of supersteps. These supersteps start with the parallel calculation of vertex properties either based on existing properties of the vertex or based on the incoming messages from the previous superstep. In a second step, messages are sent to neighbouring vertices containing calculated properties. The last step in the superstep is the aggregation of the incoming messages at the receiving vertices. Typical Pregel based programs run an iteration of a superstep until some prior defined stopping criterion is reached.

Given the bipartite nature of the graph underlying our bibliographic coupling ANNS problem it is impossible to apply an iteration of a single superstep multiple times. Therefore, this algorithm consists of three distinct supersteps.

- *Superstep I.*
  - Step 1. Publication and references calculate their degree, thus the number of outgoing or incoming edges or links.
  - Step 2. Each publication sends a message containing its identifier and out-degree to cited reference across all the outgoing edges. (Dashed line in figure 1)
  - Step 3. Each reference collects the received messages into an ordered list.
- *Superstep II.*
  - Step 1. Each reference decides if it will send out the list and to which of the citing publications. If a reference decides not to send it becomes inactive
  - Step 2. Each active reference sends messages across incoming links. Each message contains a list with the identifiers and properties of those publications that appear after the identifier of the recipient in the ordered list. (Dotted line in figure 1)
  - Step 3. Each publication collects the incoming messages
- *Superstep III.*
  - Step 1. Each publication calculates the occurrence of each identifier in the joined set of messages. A Salton cosine similarity is now

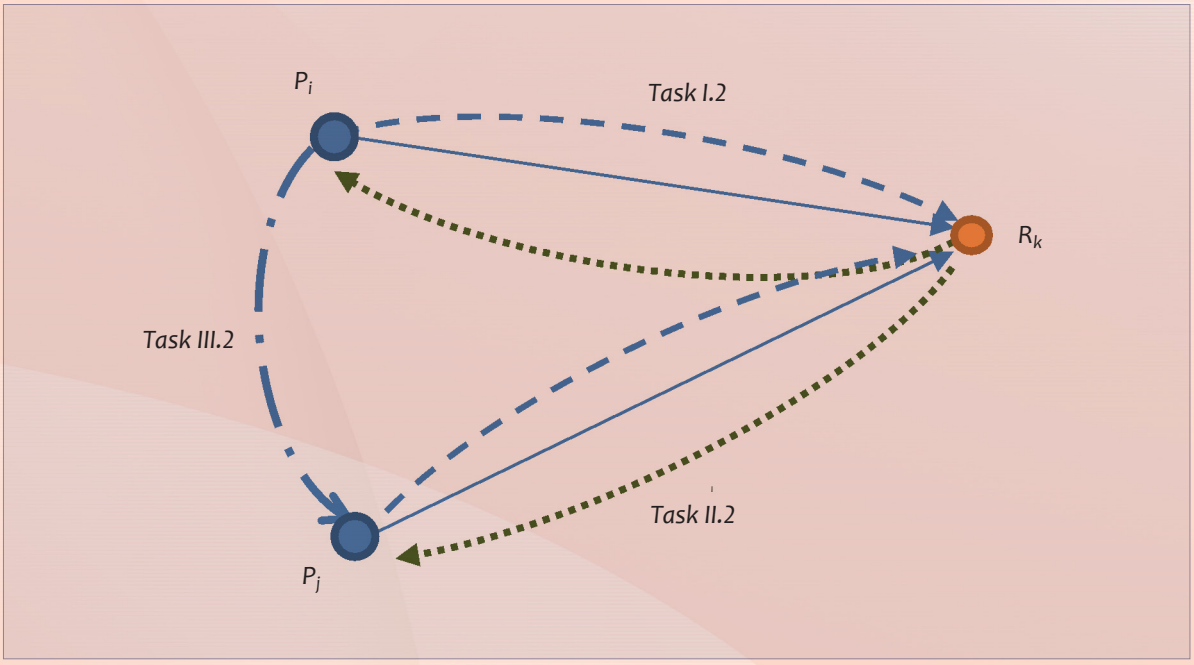


Figure 1. Schematic overview of message passing between two publications and one joint reference.

calculated based on the number of joint references, the out-degree of the current publication and the out-degree of the other publication being part of the message.

Step 2. Each publication can now send a message to its bibliographically coupled neighbour without actual edges being present. (Dash-dotted line in figure 1)

Step 3. Publications receive incoming messages and weighted edges are created and no further calculations are needed.

Several advantages are associated to this approach. Steps are performed in a sequential order and results are stored. Tasks within each step are suitable for distributed execution as they run independent from each other. References sending out messages in step 11.2 rely solely on the information already gathered by each individual reference. Consequently, each task can be performed in parallel.

But most important for this algorithm is the ability that this framework provides to define any function to be applied at the individual reference for the selection of publications receiving messages with the identifiers of their neighbour publications. This selection function could be completely randomized and thus be analogous to the selection of dimensions in a LSH procedure. But it also allows for more complex functions either deterministic or probabilistic for the selection of active references. In a state where the individual references have no or very limited information about the actual topology of the graph, it is the in-degree of each reference that is the most obvious parameter for the selecting function.

EXPERIMENTAL SETUP

For a valid testing of the different selection scenarios, I use the amount of information that is being sent after the application of the selection function in step 2. This amount can be calculated based on the in-degree of the reference and the total distribution of indegrees across the network. The next section introduces the required definitions and formulas for the calculation of the amount of passed information. At first, a bipartite graph G is defined by the sets of publications P, cited references R and edges E, with p, r and e their respective cardinality.

$$G = (P, R, E)$$
$$p = |P| \quad \text{and} \quad r = |R| \quad \text{and} \quad e = |E|$$

The number of outgoing and incoming links is calculated as the out- and indegree of publication and reference.

$$outdeg_i = outdegree\ of\ publication\ P_i \in P$$
$$indeg_j = indegree\ of\ reference\ R_j \in R$$
$$\sum_{i=1}^{p} outdeg_i = \sum_{j=1}^{r} indeg_j = e$$

The indegree for the references in the graph ranges from 1 to some highest value $n_{\max}$. References not cited by any publication are not included in the graph. Each reference can be assigned to a set of references with the same indegree.

$$n = 1..n_{\max} : \forall R_j \in R : 1 \le indeg_j \le n_{\max} \wedge \exists R_j \in R : indeg_j = n_{\max}$$
$$\forall R_j \in R : R_j \in D_{(n)} \Leftrightarrow indeg_j = n$$

The amount of information to be sent by all the references in a set of same indegree n is equal to the product of cardinality of this set and the number of possible 2-combinations in a set of size n. One unit of information is the pair of the identifier of the citing publication and its outdegree as it is send out at step 1.2. References with a degree of 1 will not send out any information as it is not possible to make any 2-combination in a set of size 1

$$I_{(n)} = |D_{(n)}| \frac{n(n-1)}{2}$$
$$I_{(1)} = 0$$

The total information in the publication-reference network is equal to the sum of information over each of the indegree sets.

$$I_{tot} = \sum_{n=2}^{n_{\max}} I_{(n)}$$

It is not only possible to calculate the total amount of transmitted information but also to sum over a range of indegree values.

$$I_{(n..m)} = \sum_{i=n}^{m} I_{(i)}$$

The range can be chosen with such boundaries that it accounts for a given share of the total information. The upper bound for the range of indegrees accounting for up to 25% of the total information can be defined as follows:

$$n = n_{25\%} \Leftrightarrow \left|D_{(n)}\right| > 0 \wedge \forall m < n_{25\%} : \frac{I_{(2..m)}}{I_{tot}} \leq \frac{I_{(2..n_{25\%})}}{I_{tot}} \leq 0.25$$

The definition of these ranges associated with some share of information provides the mechanism to choose different testing scenarios with equal amount of information contained in the messages being transferred from reference back to publications. These ranges can not only be taken from the lowest end of the indegree distribution, but also from the top, in the middle or a combination of bottom and top end.

As table 1 shows, a combination of these four types with four different levels of shares of information to be transmitted defines the first sixteen scenarios to be tested. The table 1 specifies the indegree ranges. This approach defines deterministic binary functions solely based on the indegree and the relevant range. References do or do not send their compiled list to each of their citing publications;

However, it is also possible to define probabilistic functions. The first four probabilistic scenarios apply a simple random function to each message to be transmitted. The probability to be transmitted is equal to the given share of information and independent of the indegree at the level of the cited reference and independent of the size of the compiled list to be sent. Analogous to the deterministic functions the shares are set to 25%, 40%, 50% and 66%

A last series of four 'tailed' scenarios combines a deterministic upper limit threshold with a probabilistic function for those references where the indegree exceeds the threshold. This means that these references randomly select a limited set of citing publications that receive the compiled list. The probability to be selected can be defined as

$$p = \frac{k}{n}$$

where n is the indegree of the reference and k is equal to

$$k = \begin{cases} n \mid n \leq l \\ \left\lceil \frac{l(l-1)}{2(n-1)} \right\rceil \mid n > l \end{cases}$$

with l being the threshold.

The amount of information to be sent by a set of references with the same indegree n can be calculated by substitution of n by k

$$I_{(n|l)} = \left|D_{(n)}\right| \frac{k(n-1)}{2}$$

and the amount of information transmitted in the network with a given threshold l is then the sum over all the indegree values in the

$$I_{tot|l} = \sum_{n=2}^{n_{max}} I_{(n|l)}$$

These definitions allow us to set the threshold to such a value that only a given share of information is used for the creation of the bibliographic coupling networks. In line with all the previous scenarios, thresholds are set to create tailed scenarios ac-

	BOTTOM	MIDDLE	TOP	BOTTOM + TOP
20%	$2..n_{20\%}$	$n_{40\%}..n_{60\%}$	$n_{80\%}..n_{max}$	$2..n_{10\%}$ or $n_{90\%}..n_{max}$
40%	$2..n_{40\%}$	$n_{30\%}..n_{70\%}$	$n_{60\%}..n_{max}$	$2..n_{20\%}$ or $n_{80\%}..n_{max}$
50%	$2..n_{50\%}$	$n_{25\%}..n_{75\%}$	$n_{50\%}..n_{max}$	$2..n_{25\%}$ or $n_{75\%}..n_{max}$
66%	$2..n_{66\%}$	$n_{17\%}..n_{83\%}$	$n_{34\%}..n_{max}$	$2..n_{33\%}$ or $n_{67\%}..n_{max}$

Table 1. Indegree ranges used for the specified share of transmitted information

counting for 20%, 40% 50% and 66% of the total amount of information.

The results from these scenarios are gauged against the original bibliographic network using all the publication-reference links.

DATA SOURCE AND PROCESSING

1.39 million Publications of type Article or Review indexed in the 2013 volume of Clarivate Analytics Web of Science (WoS) were used. In WoS, references in these publications get a specific R9-code. References to the same cited work in different publications are labelled with the same R9-code. Consequently, a co-occurrence of R9 –codes in the reference lists of two publications indicates a bibliographic coupling. Both publications and cited references are considered to be nodes in a large network. The reference to a cited document is recorded as a directed edge in the bipartite network. The final dataset consists of pairs of identifiers where the first refers to the citing publication and the second to the cited reference.

The processing is done using the Elastic MapReduce service offered by Amazon in their AWS Cloud Compute environment. Several Hadoop clusters running Spark with one master and from five up to ten memory optimized worker instances were created. The bipartite network is processed by using the GraphX library which is the graph computation API within Apache's Spark. This library provides the required methods for the development of a bulk-synchronous messaging system. *mapVertices* and *joinVertices* are the two methods that can be used in the first (calculation) task in each superstep. The *aggregateMessage* method combines second and third task which passes relevant information across existing edges and combines all the incoming messages using the provided function.

RESULTS

The analysis started with the calculation of distribution of the indegree and the amount of information to be transmitted at step II.2 associated with each value of indegree found in the graph. As mentioned before, the unit of information to be sent is the pair of identifier and outdegree of the citing publication (a pair of a *Long* and *Int* values in the Spark implementation). Figure 2 plots this amount of information in a logarithmic scale for the obtained indegree values. The highest indegree value found was 5927 and occurred only once. The horizontal axis is not truly interval scaled but merely ordinal as only those indegree values that occur in the dataset are included. Consequently, the figure shows a steep increase of information near the end of the distribution for those values that are only observed once.

The particular shape of the figure resembles the typical design of Viking ships (drakkar called in Dutch, hence the name of the algorithm) with symmetrical ends and justifies the selection of a given amount of information from both sides of the distribution. The thresholds of the indegrees are given in table 3 and allow the creation of the intervals required for the definition of the sixteen scenarios as presented in table 1. In the tailed scenarios, the thresholds are respectively set to 20, 93, 193 and 600 to obtain the same shares.

The first test measures the recall of each scenario. This is calculated by comparing the final number of bibliographic coupling links of each scenario with the selected share of in-

Number of publications	p	1,391,192
Number of cited references	r	17,248,290
Number of publication-reference pairs	e	49,156,442
Average number of references per publication	p/e	35.34
Average number of citations to references	e/r	2.85

Table 2. Descriptive statistics for the bi-partite network [Data sourced from Clarivate Analytics Web of Science Core]

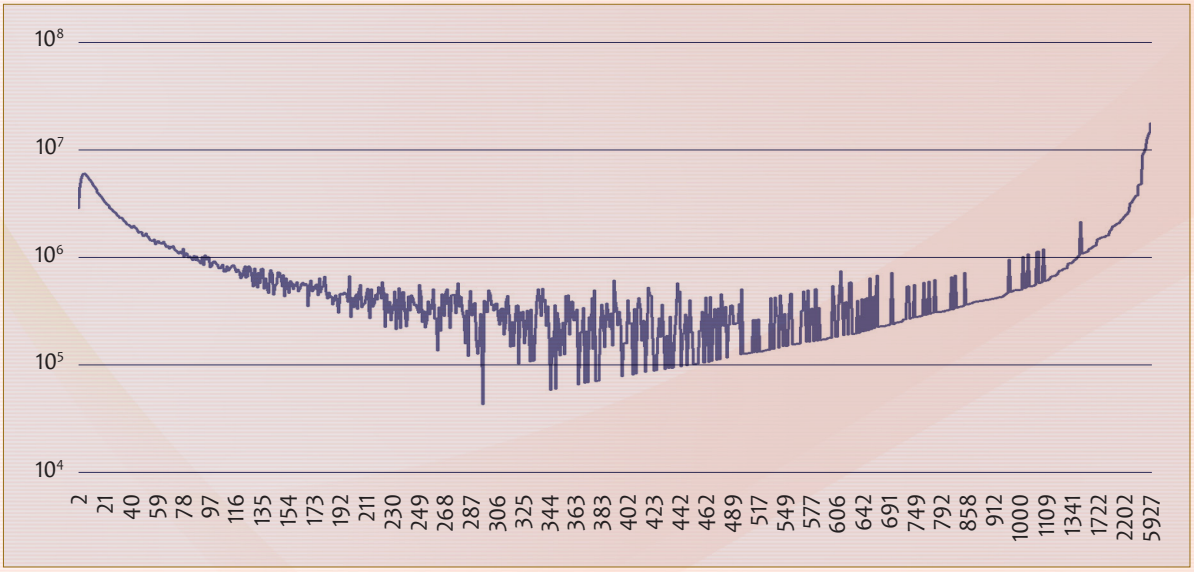


Figure 2. Amount of information to be transmitted by each value of observed indegree (x-axis: observed indegree; y-axis: amount of information) [Data sourced from Clarivate Analytics Web of Science Core Collection]

	THRESHOLD		THRESHOLD		THRESHOLD
$n_{10\%}$	13	$n_{34\%}$	98	$n_{70\%}$	1260
$n_{17\%}$	26	$n_{40\%}$	158	$n_{75\%}$	1760
$n_{20\%}$	35	$n_{50\%}$	326	$n_{80\%}$	2165
$n_{25\%}$	53	$n_{60\%}$	657	$n_{83\%}$	2725
$n_{30\%}$	75	$n_{66\%}$	1026	$n_{90\%}$	4333

Table 3. Upper thresholds for the selection of the associated amount of information. [Data sourced from Clarivate Analytics Web of Science Core Collection]

formation with the number of bibliographic coupling links when using the complete citation graph. Using all the information present in the original publication-reference network of all 2013 publications resulted in 496 million weighted links between publications. This recall could also be rephrased as the ratio between the density of the bibliographic network after the application of the selection function and the density of the BC network without any selection. The density of the latter network is about 0.05%. Table 4 presents these recall values for each of the twenty-four versions. The columns present the amount of information that is transmitted and the rows refer to the different scenarios being applied for the selection function.

The first observation is that when using a pure random function the recall is almost the same as the selected information share. This can be observed in the sixth row. Next, the recall of the two scenarios that focus on those references with a low indegree ('Bottom' and 'Tailed') is below the value set by the pure random selection and the share of used information. The highest density is obtained when choosing those references with the highest indegree to build the bibliographic coupling network.

The scenarios where either the references located at the centre of the information distributions or at the outer bounds are selected still perform better than the random scenarios. The largest difference between top and bottom can be observed when half the information is used. Cutting the set of references into two subsets associated with an equal amount of information results in a recall of 46.8% for the lower end compared to 57.6% for the upper end. But relatively, when only using 20% of the available information, the use of the most cited references results in an BC-network with a 30% higher recall than based on the least cited references. Based on these observations, it would be

	20%	40%	50%	66%
Bottom	18.5%	36.8%	46.8%	64.2%
Between	23.4%	44.8%	54.4%	70.1%
Top	24.2%	45.9%	56.7%	73.8%
Bottom & Top	22.9%	42.5%	52.1%	68.8%
Tailed	17.9%	36.1%	46.0%	63.3%
Random	20.1%	40.0%	50.0%	66.0%

Table 4. Recall of each scenario with the associated amount of selected information. [Data sourced from Clarivate Analytics Web of Science Core Collection]

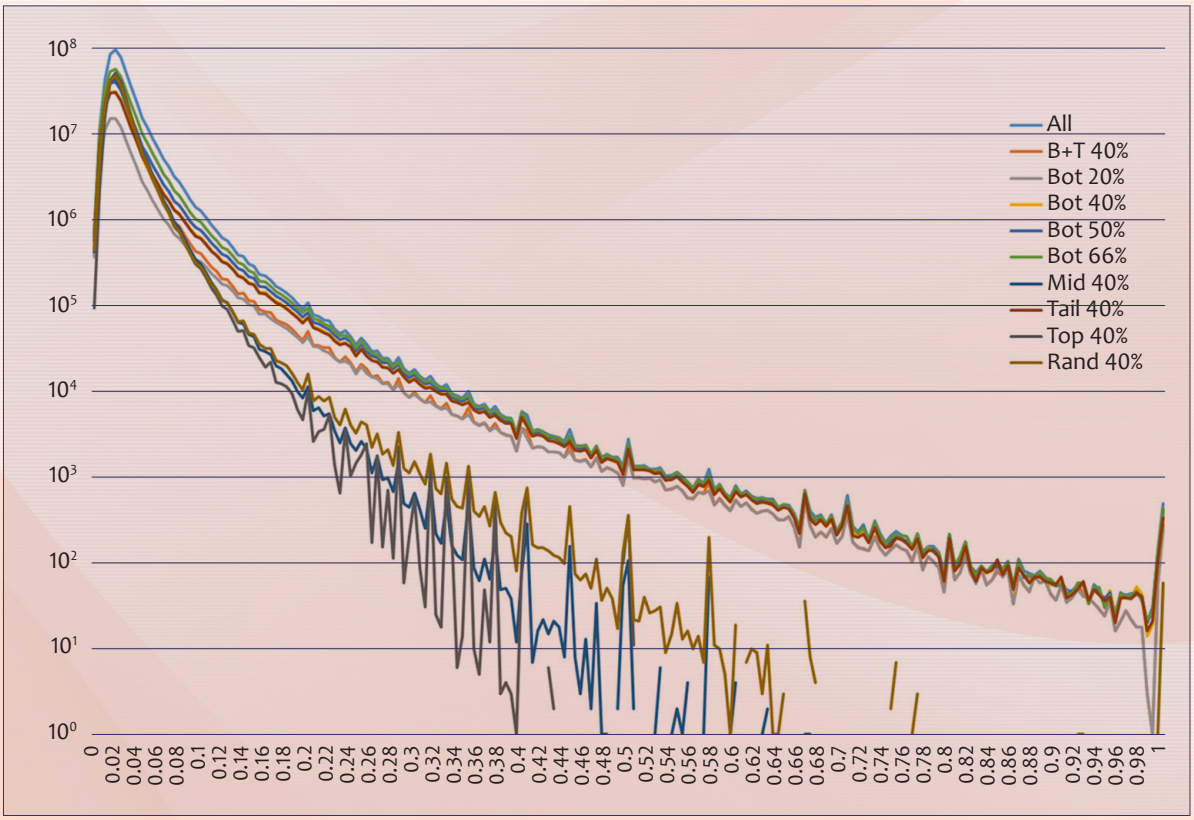


Figure 3. Comparison of distribution of weighted links across different scenarios. (x-axis: Strength of BC-link; y-axis: count of observed links) [Data sourced from Clarivate Analytics Web of Science Core Collection]

justified to say the selection of the top references is a better approach.

It should be noted that this messaging based algorithm does not result in false positive BC-links between two publications as messages are only passed along truly existing citation links. When using a binary approach, the precision would then be equal to 100%. However, as Bibliographic Coupling results in a weighted network, we can measure the ability of each scenario to approximate the actual strength or weight of the link. It is in step III.1 that the cosine similarity between publications is calculated. Given the fact that false positives are absent and only false negatives can occur, the weight of the link can never be overestimated. The effect of a selection of different scenarios on the distribution of link weights can be seen in figure 3.

The top line refers to the BC-network without any selection function. None of the other scenarios has a distribution that surpasses this at any weight value. Two clear phenomena can be observed from this graph. First those scenarios that do not include systematically those references with a low indegree underestimate the strong links. The top scenario with 40% of the available information is at the bot-

tom of the graph from a weight of at least 0.1. But also the scenario which selects the references in the middle of the information distribution performs lower. And a pure random function is not much better. But those scenarios that focus on the references with low indegree approximate the distribution of strong links. As expected, adding more information to these scenarios improves the results slightly at the upper end of the weight distribution.

The second observation is that the scenarios selecting the bottom references fail primarily in detecting the lower weighted links. It is in this area that the explanation can be found for the results presented in table 4 with respect to the lower recall of those scenarios.

It seems that closely related documents share references to poorly cited documents while highly cited documents receive citations from a broader range of papers covering multiple topics. These observations have strong implications on the choice of selection function. When the objective of the creation of a large scale bibliographic coupling network in a computationally constrained environment is to find pairs of closely related paper then the selection function should focus on the lesser cited references. Opposed to this, the objective could also be the clustering of the complete network in which case the selection functions can be restricted to the upper end of the indegree distribution.

CONCLUSIONS

The application and use of large scale bibliographic coupling networks has been hindered by the computational and storage resources required for the creation of these networks. Alternative networks based on direct citations have been used in large scale analysis. The new graph messaging algorithm proposed in this paper provides an opportunity to produce the large scale networks through the application of different selection functions at the level of individual cited references. The experiments with different functions show that references at the lower or higher end of the indegree distribution play a different role in the citation network. Focussing on the bottom results in a network that approximates most of the strong links but is more likely to ignore the weaker ones. Shifting the focus to the other end creates the inverse effect: a higher recall but worse for the identification of strong links. The choice for a particular set of selection function thus depends on the actual objectives for the creation of these BC-networks. If global clustering is the goal then the upper end of the distribution is the right path while if the objective is only to delineate a set of documents closest related to a particular sample the lower end of the indegree is most relevant. Future research will investigate the applicability of this graph based nearest neighbour search algorithm for lexical similarity between scientific documents.

REFERENCES

1. Johnson, W. B. & Lindenstrauss, J., (1984). Extensions of Lipschitz mappings into a Hilbert space. Contemporary Mathematics. 26, 189–20
2. Karapiperis , D. & Verykios, V.S. (2016). A fast and efficient Hamming LSH-based scheme for accurate linkage. Knowledge and Information Systems, 49 (3), 861-884.
3. Kessler, M.M., 1963, Bibliographic Coupling between scientific papers. American Documentation, 14 (1), 10-25.
4. Malewicz, G., Austern, M.H., Bik, A.J.C., Dehnert, J.C., Horn, I, Leiser, N. & Czajkowski, G. (2010). Pregel: a system for large-scale graph processing. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 135-146
5. Rajaraman, A. & Ullman, J. (2010). "Mining of Massive Datasets, Ch. 3." URL: http://infolab.stanford.edu/~ullman/mmds.html
6. Ravichandran. D., Pantel. P. & Hovy. E. (2005). Randomized algorithms and nlp: using locality sensitive hash function for high speed noun clustering. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 622-629.
7. Valiant, L.G. (1990). A bridging model for parallel computation, Communications of the ACM, 33 (8), 103-111.