KNOWLEDGE MEDIA
KMi
INSTITUTE

Mining Research Publication Networks for Impact

First Year Probation Report

Technical Report kmi-14-01
January 2014

Drahomira Herrmannova

The Open University

The Open University
Knowledge Media Institute

Mining Research Publication Networks for Impact

First Year Probation Report

October 2013

Drahomira Herrmannova

Contents

Abstract

The question of how to evaluate the quality of research publications is very difficult to answer and despite decades of research, there is still no standard solution to this problem. Particularly at present, with the amount of scholarly literature rapidly expanding, it might become very difficult and time consuming to recognise what is the key research that presents the most important contributions to science. Furthermore, this question is highly relevant not only to researchers, but also librarians, publishers, editors and promotion and grant committees.

Currently, the most widely used methods for evaluating research publications are based mainly on citations. One of the crucial problems of this approach is the time delay between the date of publication and receiving the first citations. This delay complicates the process of finding recent relevant research. Moreover, citations included in a publication are based solely on the choice of the author and they don't necessarily indicate the quality of the cited paper. There are also significant differences between citation patterns in different fields of science.

Within this area we are interested in finding new methods which use semantically richer information for assessing quality. An example of such information might include semantic similarity of publications and analysis of their full-text, citation network analysis, for example finding bridges between distinct clusters of publications, etc. Main goals of this research include evaluating how the measures based on citations represent the quality of a paper and designing new methods for measuring quality that will address the challenges in this area.

This report reviews the state-of-the-art methods for evaluating science and research. In particular it focuses on research publications and other recorded information related to science. The issues, challenges and gaps in the current research are reviewed. The research proposal presented in this report is based on this review and gap analysis. The final part of the report presents the pilot study and the detailed plan for the next two years.

Chapter 1

Introduction

The following report is concerned with one key question: "How to evaluate the quality of research publications?" This question is probably as old as scholarly publishing itself. Particularly at present, with the amount of scholarly literature rapidly expanding it might become very difficult and time consuming to recognise what is the key research that presents the most important contributions to science. This illustrates the need for indicators and measures of quality that would help to filter the literature.

Past decades have seen a steady growth of the research field known as *bibliometrics* and the birth and growth of several sister and sub-fields which all try to answer the question of how to evaluate the quality of research publications using different methods and data. However, in this report it will be demonstrated that the current most widely adopted and used methods are often criticised and alternatives to these methods are being sought.

This is the main aim of this research – to find new automatic methods for assessing the quality of scholarly publications, which will address the current issues and challenges. This report focuses on the state-of-the-art methods for evaluating the impact and quality of research publications, the gaps in the research and the need for new methods. It then presents the research proposal and questions, which are based on the literature review. Finally, the research plan for the next two years is outlined.

1.1 Motivation

Before getting to the literature review and the rest of the report, several motivating examples of use will be presented.

Example 1: How to select relevant literature for reading?

As the amount of research literature is steadily increasing, researchers often rely on various filters to help them to reduce the number of articles that they need to read. This is true especially now, when almost all research articles are published online and most research eventually gets published somewhere – if not in a peer reviewed journal, then for example in a self-archiving repository such as Arxiv.org. The Figure 1.1 well illustrates this growth.

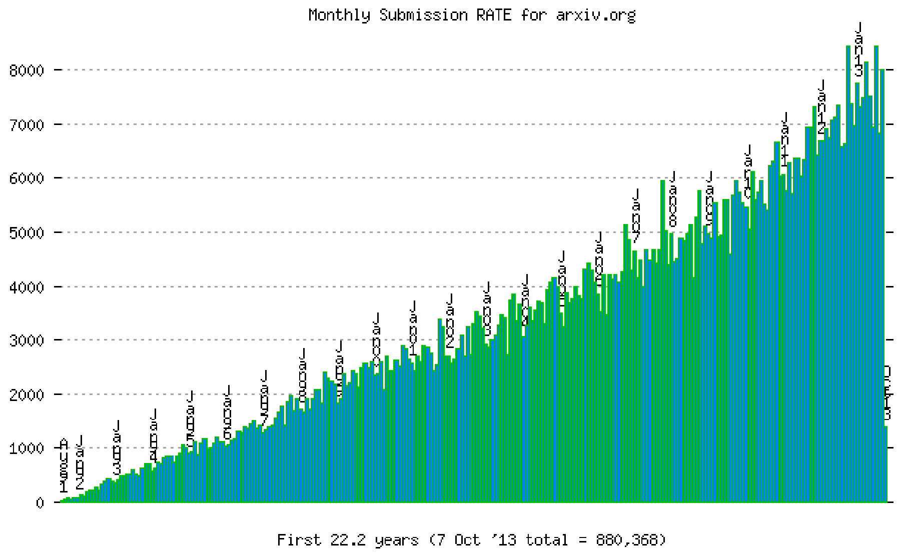

Figure 1.1: Monthly submission rate for the electronic preprints archive Arxiv.org. Figure is showing submissions for each month since 1991. Source: http://arxiv.org/

It is clear that with this amount of literature, researchers need to be able to select publications relevant to their work and to make decisions which publications to keep and read and which to skip. Quality and impact indicators can provide help with this decision process.

For example, Google Scholar, which is one of the major citation indexes, presents search results approximately in the order of decreasing citations. In addition it also offers listings of the most cited publications and authors in each area. The Open Access publisher PLOS allows sorting of articles by the number of views and downloads. Another option is subscribing to updates of individual journals which are of interest to the researcher. Nevertheless, it will be demonstrated, that using filters such as these might not necessarily lead to retrieving the most important literature and in addition, a significant portion of literature might be completely ignored.

However, the process of selecting literature is not the only way how researchers might benefit from various quality and impact indicators. The career progression of research employees is often dependent on how well they can demonstrate their productivity and the quality, importance and impact of their research [Seglen, 1997; Rossner et al., 2007; Arnold and Fowler, 2010].

Example 2: How to select journal subscriptions?

Another example of use is the question of selecting journal subscriptions by librarians. The prices of journal subscriptions between 1986 and 2003 have been growing more than three times faster than the consumer price index (CPI) [Panitch and Michalak, 2003], in 2009 the cost of journal subscriptions has grown to almost four times the CPI [Kyrillidou and Morris, 2011] (see Figure 1.2). The price growth has got to a point when universities started announcing they cannot any more afford the costs of journal subscriptions [Sample, 2012]. In a situation like this the possibility to compare journals based on the quality and importance of research published in them might be of help to librarians.

A well-known metric for evaluating journals is the journal impact factor (JIF). The JIF was proposed already in the 1972 by Garfield [1972] and is published yearly in the Journal Citation Reports[1]. The JIF rating of journals is based on the number of citations received by the journal and the number of articles published in that journal.

Provided that a citation is a demonstration of impact of the cited article, this measure should be sufficient for selecting the most influential journals in a research field. There are, however, many reasons why such metric is not sufficient, starting from the simple fact that many journals are by nature not cited very much (for example journals that are from a very narrow research field, or review journals), and ending with examples of purposely trying to manipulate and increase the JIF rating [Brumback, 2009; Arnold and Fowler, 2010].

Example 3: How to aid reviewers of funding and grant proposals, hiring committees etc.?

The question of evaluating the quality of research publications is also very relevant to universities, funding organisations and other institutions which need to make hiring decisions or distribute grants. The applications should generally be reviewed by an expert in the field who can objectively assess the

[1] http://thomsonreuters.com/journal-citation-reports/

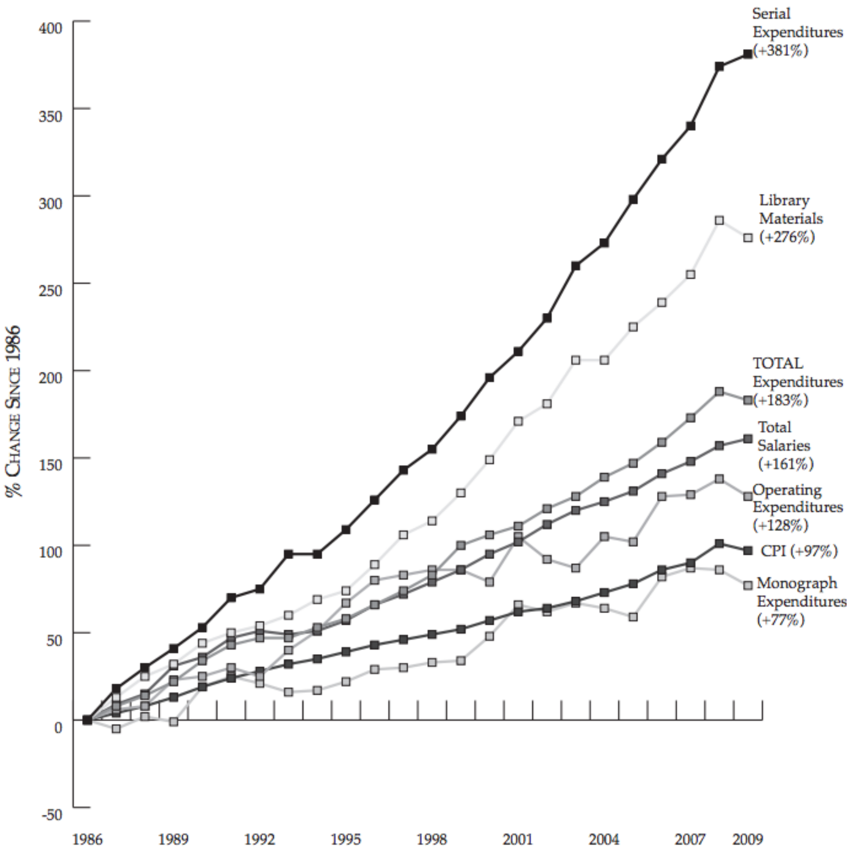

Figure 1.2: Expenditures in ARL libraries (Association of Research Libraries) between 1986 and 2009. Important lines are the Serials Expenditures (money spent on journal subscriptions) and the CPI (Consumer Price Index or inflation). Source: [Kyrillidou and Morris, 2011]

quality and importance of the work of the researcher, or novelty and applicability of the research project. Such objective and expert form of evaluation should be the ideal solution. Even expert peer review is however sometimes criticised [Priem et al., 2010; Holste et al., 2011]. Automatic research evaluation methods have been proposed as a support for the peer review process [Holste et al., 2011].

Example 4: How can publishers evaluate and promote their journals?

Editors and journal publishers benefit from evaluation methods for research publications as well. Quality and importance indicators and metrics are a way how journal publishers can promote their journals and on the other hand editorial boards can review the results of their decisions. Both tasks are at present fulfilled mainly by the above mentioned journal impact factor.

Example 5: How to evaluate the returns of research to the society?

Finally, research publication evaluation methods are beneficial to the society because they offer a possibility how to evaluate the returns of research investment. Publicly funded research is financed from the money of taxpayers and it is important to understand whether the money was worth spending. [Sutherland et al., 2011] summarised three main benefits that research might bring to the society:

- Improved life quality or sustainability. This includes for example research regarding health, the effectiveness of public services, policies, quality of life or the environment.
- Economical benefits which might come for example from linking research with industry and resulting financial profit.
- Contribution to knowledge, in case of research that is driven by curiosity.

Some countries have already realised the importance of evaluating the outcomes of research and started to implement research assessment exercises aimed at comparing different institutions and disciplines within the country as well internationally. Such exercises are for example the Research Excellence Framework[2] (REF) in the UK, Excellence in Research for Australia[3] (ERA) or the Performance Based Research Fund[4] (PBRF) in New Zealand. One of the main parts of these assessment exercises is the evaluation of research publications.

[2]http://www.ref.ac.uk/

[3]http://www.arc.gov.au/era/

[4]http://www.tec.govt.nz/Funding/Fund-finder/Performance-Based-Research-Fund-PBRF-/

1.2 Problem statement

The need for methods for evaluating the quality of research publications has been demonstrated in the previous section. Traditionally, expert peer review has been used as the main filter for controlling the quality of published research. This **qualitative evaluation** method might however also be facing some problems [Priem et al., 2010; Priem and Hemminger, 2010; Holste et al., 2011]. For example, it is not always possible to find and have the publication reviewed by a true expert in the area. Another problem is the timeliness and cost of peer review. Finally, the peer review process might also suffer from unfairness resulting from biased opinion, etc. A satirical example of the failure of peer review – a list of famous rejection letters – has been published online[5].

Another option are the classical **quantitative evaluation** methods, such as citation counts, the JIF and similar. These methods have several advantages, for example they are very clear and easily accessible (the JIF is calculated and published yearly by Thomson Reuters[6], citation counts received by individual papers can be freely obtained from many online citation indexes, such as from Google Scholar).

These methods however also suffer from some issues. Most importantly, citation-based metrics generally don't express the **quality** of the publication, but rather it's **impact** on the research field or it's **utility** [Seglen, 1992]. The traditional methods based on citations often ignore reasons for citation (which might be both positive and negative [Nicolaisen, 2007]), co-authorship, self-citation, article type (survey papers might receive more citations than primary research [Price, 1965]), etc. So called "citation half-life", which is the time required before a paper receives half of the citations it will ever receive, might be (depending on the research field) up to several years long [Arnold and Fowler, 2010]. For this reason metrics based on citations are not usable for newly published research. Many publications are never cited at all [Seglen, 1992; Garfield, 2005], but this doesn't necessarily mean that the research they represent is not significant. Journal impact factor is also sometimes incorrectly used to evaluate individual researchers [Seglen, 1997].

Many new methods have been proposed in the past decades with the aim of overcoming these issues. Thanks to the digitalisation of scholarly literature and creation of the Internet, many new fields have been born which focus on different data sources (for example usage logs and social media).

[5] http://www.fang.ece.ufl.edu/reject.html

[6] http://wokinfo.com/products_tools/analytical/jcr/

Nevertheless, none of the new methods have yet replaced the classical evaluation metrics. Both the classical methods as well as the new research areas will be reviewed in Chapter 2.

1.3 Research Objectives

In the previous two sections the need for evaluation methods for research publications was demonstrated and some issues and challenges were mentioned. It can be seen that the area of evaluating research publications faces one main challenge: There is no widely used and also generally accepted method or set of methods for evaluating the quality of research publications. I believe there is one additional challenge that needs to be overcome: There is no clear definition of what influences the quality of a publication or how does the quality reflect in different attributes of different publication types.

My main research questions can therefore be formulated as follows:

Question 1: What factors influence the quality of a research publication (with regard to the publication type and to the research field)?

Question 2: What is the relationship (if there is any) between the impact of a publication, measured by the classical bibliometric methods, and the quality of a publication?

Question 3: How can we detect the factors influencing quality in order to evaluate the quality of a research publication?

Question 4: How can this evaluation be used in other disciplines, such as in information retrieval, and to detect research fronts and trends in development of research disciplines?

1.4 Preliminary work

Literature review: During the first months of my PhD studies I have conducted a literature survey of the fields of scientometrics, bibliometrics and related fields. I have first focused on the historical developments that led to the birth and growth of this research discipline, such as the creation of the first citation index of science literature and the first metrics that could be developed thanks to the existence of this citation index. I believe these historical developments are quite important for understanding this research area, especially because many of the classical metrics and methods created decades ago are still being used in research evaluation. Starting from the history I have looked at the main changes and the birth of various sister and sub fields whose creation was catalysed by the creation of the Internet and later the Web 2.0. Finally I have studied the "modern" bibliometrics and the recent developments such as the growth of Open Access publishing. During this literature survey, I was guided by three excellent reviews: a book by Bellis [2009], which mainly focuses on the history of bibliometrics, and two review articles [Thelwall, 2007; Bar-Ilan, 2008].

Data collection: In order to perform any kind of experiments and analysis the main thing needed are the data, in this case the research publications. At present there are several datasets available but unfortunately none is quite complete (in terms of dense enough network of citations within the dataset, full texts of the publications, coverage of various research disciplines, etc.). The examination of existing datasets was my first goal.

1.5 Structure of report

This report is organised as follows. First, Chapter 2 is devoted to a review of existing literature. It also describes the limitations of the current methods and the gap created by these limitations. Chapter 3 describes the plan for the next two years of my PhD studies. Finally, Chapter 4 reviews the preliminary work that has been done.

Chapter 2

Literature review

This chapter is divided into three main sections. Section 2.2 aims at summarising the main directions and approaches which relate to the problem and objectives mentioned in the Introduction (Chapter 1). Section 2.3 then explains what are the limitations of the existing approaches and why an ideal solution is still missing. Finally, Section 2.4 provides the gap analysis and describes the selected approach for the research proposal.

2.1 Background

The following section presents a brief explanation of some concepts and terminology, which will provide background for better understanding the literature review. In particular, this section aims at explaining how can research be evaluated at different stages, what are the different levels of granularity at what we can evaluate research (articles, people, groups, ...) and what purposes can research evaluation serve. It will also be explained how does this project fit into this big picture.

2.1.1 Evaluation levels

There are several levels of granularity at which we can evaluate research. These levels, starting from the most fine-grained one, are:

- **Publications** and any other types of research output (such as measurement data and results, plots, figures etc.). Methods used at this level include citation counting, citation network analysis and text analysis.
- **Journals** stand at a higher level of granularity as they group together research papers concerned with a similar topic. Probably the best know evaluation metric for journals is the journal impact factor.

- **Individual researchers**, who are represented by the set of papers which they published. The evaluation of researchers however does not rely only on their publication record, other useful data include academic affiliation, teaching record, etc.). A well known metric for evaluation of researchers is the Hirsch index.

- **Groups of researchers** belonging to firms, research institutions, universities and countries. Methods used at this level include for example analysis of international collaboration.

The higher (or more general) levels are dependent on the lower levels. For example the rating of journals is dependent on the quality of articles published in them, while research groups are evaluated based on individual researchers belonging to the group (and in turn also based on publications). This research project is concerned with the evaluation of research publications, because of this dependency it will also focus on the question of how can document-level indicators be applied to the higher levels.

2.1.2 The linear model of innovation

Research can be evaluated at different stages. The linear model (LM) of innovation illustrates, how a technical change can happen, starting from pure research and ending with societal benefits. Figure 2.1 shows the LM along with evaluation methods available at each stage. Because the aim of this research project are research publications, it covers the pure and applied research stages.

2.1.3 Four point of view

The quantitative analysis and evaluation of scholarly literature can serve many purposes. In the introductory chapter (1 some examples of how it might serve different groups of people were provided. The goals and aims of the analysis will influence which tools and methods will be utilised and how the results will be used. The main goals that research evaluation can serve include (first three were listed in [Bellis, 2009, Chap. Introduction]):

1. Information retrieval; for example searching for similar literature, key publications, etc.

2. Research quality control; for example evaluation of research publications and other research outputs, evaluation of individual researchers, institutions, countries, etc.

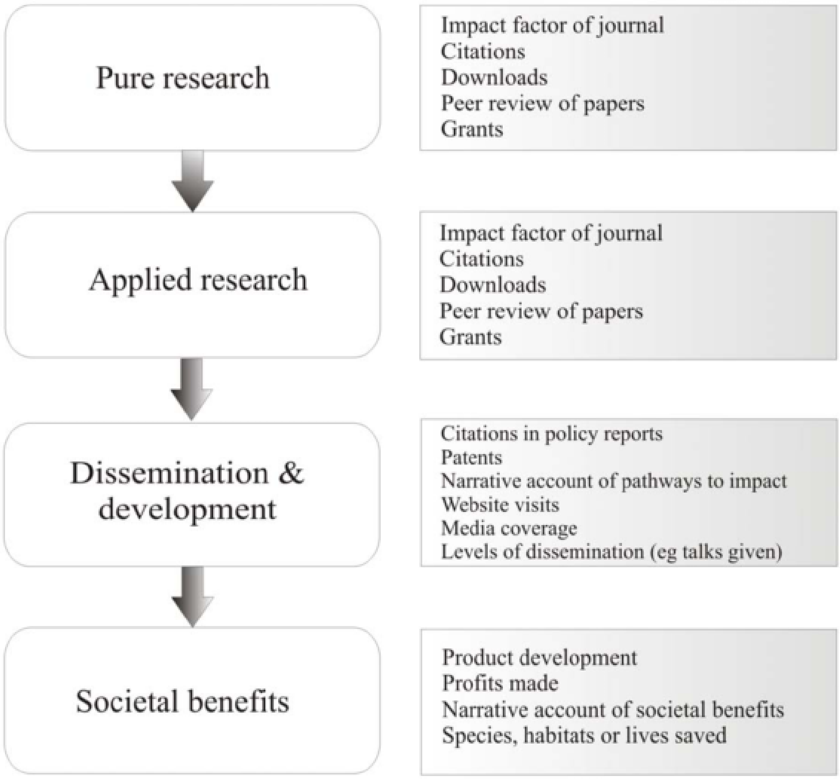

Figure 2.1: The linear model of innovation. Source: [Sutherland et al., 2011]

3. Study of the history and sociology of science; for example the structure and evolution of scientific disciplines, collaborations between authors, etc.

4. Detecting research fronts and emerging topics; for example current key publications and personnel.

This research project focuses on the item 1 and the possibilities of application to the items 2 and 4.

2.1.4 Terminology

Before moving on to the literature review some important terminology will be explained. This relates to the naming of the research field. Several names are being used with respect to the methods and data being referred to. A brief description of each of these research areas follows.

Scientometrics is, paradoxically, a science which is devoted to the study of science and research, or in other words it is a science of science. The term *naukometriya* or scientometrics was created by Nalimov and Mulchenko [1969]. Scientometrics is concerned with scientific productivity, with the relations, history and evolution of scientific disciplines, their structure, etc. Bibliometric indicators are often used in scientometric evaluations, but these are not the only methods and data available – researchers' background (affiliation, teaching, ...), research inputs and outputs (other than publications, for example financial inputs and outputs) etc. can also be considered.

Bibliometrics is concerned with any kind of scientific literature or more generally with any kind of written information. The term bibliometrics was first introduced by Pritchard [1969]. Pritchard defined bibliometrics as "the application of mathematics and statistical methods to books and other media of communication". The methods used by bibliometrics include counting of articles, books, patents and other publications and also citation counting. Bibliometrics is commonly used to asses scholarly impact, but for example also to study the evolution of scientific disciplines. Bibliometric methods are also often used in scientometrics, bibliometrics and scientometrics thus overlap to a considerable degree.

Informetrics was according to Hood and Wilson [2001] first proposed in 1979 by Otto Nacke. Simply put, it is a quantitative study of any type of information (including research publications and other outputs). Informetrics applies bibliometric techniques also to non-scientific publications and written records, it can therefore be viewed as an extension or superset of bibliometrics.

Webometrics takes the informetric methods and models and adapts them for use on the web. The term webometrics was introduced in 1997 by Almind and Ingwersen [1997]. Webometrics is based on the idea that it's possible to view the web as a citation network where nodes are web pages. [Bjrneborn and Ingwersen, 2004] divide webometric studies into four main areas: (1) analysis of page content, (2) analysis of link structure, (3) usage analysis and (4) analysis of web technologies (such as search engine performance).

Cybermetrics has first appeared in a title of a new journal in the same year as webometrics (1997). Cybermetrics and webometrics are related terms which are used to describe the same research area. This allows them to be used interchangeably. Bjrneborn and Ingwersen [2004] distinguish between the two terms and propose to use webometrics to describe informetric studies of the web, while cybermetrics describe informetric studies of the whole Internet (that means not just web pages and documents but all Internet communication and technology).

Altmetrics is the newest research area of the previously mentioned. The term and the vision of altmetrics (originally alt-metrics, short for alternative metrics) was first introduced by Priem et al. [2010]. The goal of altmetrics is to study science and research by using data from the social web. This includes online bookmarking services, discussion forums, blog and microblog posts, etc. Altmetrics were created as an alternative to the traditional citation counting, the extensive use of which has been criticised by many authors.

Figure 2.2 well illustrates the relations between the research areas.

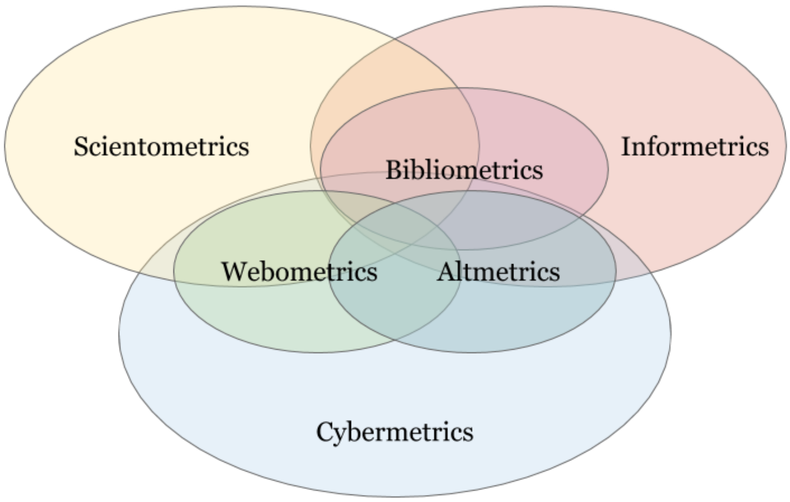

Figure 2.2: The relations between the research areas

2.2 Research evaluation

The following section presents the history and the main developments in the field of research evaluation. The few past decades have witnessed several important events and changes. These have been the computerisation of the scholarly literature and later the transition of the literature (and the whole publishing process) to the Internet, the creation of the first citation index of scholarly literature and finally one very recent change – the growth of Open Access publishing. As will be seen, each of these events reflects in the evolution of this research field.

2.2.1 Foundations of bibliometrics

As mentioned earlier, the term bibliometrics was first introduced in [Pritchard, 1969]. Authors intention was to create a more suitable name for the field which was previously sometimes referred to as "statistical bibliography" (this term comes from the fact that the research area was mainly concerned with statistical analyses of bibliographies and scientific journals). Bibliometric methods existed and were used decades before the term bibliometrics was introduced. According to [Bellis, 2009] the bibliometric study done by Cole and Eales [1917], in which the authors examined for example the amount of literature published in each European country, is often regarded to be one of the first bibliometric studies.

Bibliometric laws

During the 1920s and 1930s three important bibliometric studies were published which revealed some important patterns. In the 1926, Lotka [1926] observed that the distribution of productivity among scientists is very skewed and created a formula, which is now called *Lotka's law*. Later, in 1934, Bradford [1934] first described a pattern of scattering of literature over different journals, called *Bradford's law*. Zipf [1935] observed another skewed distribution, unlike Lotka's and Bradford's not related to publishing productivity, but to the content of publications. *Zipf's law* is concerned with the frequency of words in text. These three laws can be used to describe many datasets and similar skewed distributions have later been found for example also within the citation networks formed by references found in scientific publications [Price, 1965; Seglen, 1992] or on the Web [Bjrneborn and Ingwersen, 2004].

The Science Citation Index

The main event that helped to speed up the growth and popularisation of bibliometrics was the creation of the first citation index for science. The creation of this index was proposed already in 1955 by Garfield [1955] and it came to existence in 1960s as Institute for Scientific Information (ISI) Science Citation Index (SCI) [Garfield, 2006]. Garfield [1955] suggested, that a citation index of scientific literature is a way how to cope with information overload – it simplifies finding relevant literature by tracing citations, improves communication between scientists and enables them to see the consequences of their work. The Science Citation Index however also made it possible to calculate new statistics regarding scholarly literature, such as number of citations received by an article or publication counts aggregated by scientists, institutions etc.

The journal impact factor

One of the new metrics which were developed thanks to the existence of the SCI is the well-known journal impact factor (JIF) and the related metrics such as the cited half-life (median age of articles that were cited in the given year) and the immediacy index (the frequency of citations that one article receives within specific time period). The idea of JIF was first presented in 1972 [Garfield, 1972, 2005]. Since the 1970s, it is published yearly in the Journal Citation Reports (JCR). The formula for calculating the JIF is as follows:

$$JIF_x = \frac{Y}{Z}$$

where Y = number of citations to articles published in a journal in years $x-1$ and $x-2$; and Z = number of "citable items" published in the journal in the years $x-1$ and $x-2$. For example, the 2013 JIF is calculated using the data from years 2012 and 2011. The two year window was selected by Garfield [1972] based on the distribution of age of citations to articles, which has shown that typical article is cited most heavily during the first two years after it is published. The JCR however contains also the five-year impact factor.

Citation networks

The arrival of the SCI also allowed the analysis of the science citation networks. The first study of this kind was probably the one by Price [1965] [Bellis, 2009]. In this study, Price developed models to represent the distribution of citations received by a paper and used this distribution to describe the "active research front" in science. Since then, many researchers have applied bibliometric and other methods to citation networks in order to analyse and obtain impact and importance of publications. These studies include for example [Dieks and Chang, 1976; Seglen, 1992; Redner, 2005; Shi et al., 2010].

Co-citation analysis

The availability of citations and citation networks has also allowed the creation of measures of correlation based on specific citation patterns. These measures include bibliographic coupling (two documents are "coupled" if they contain the same reference or references) [Kessler, 1963] and co-citation analysis (two documents are co-cited if they are referenced by the same document) [Small, 1973]. Particularly co-citation is an interesting measure of correlation or closeness because it expresses the fact that other authors perceive selected work as similar or related. A similar measure based on co-citations is the analysis of co-cited authors [White and Griffith, 1981] which measures similarity of authors based on how frequently is their work cited together. Co-citation analysis has also been applied to journals, the first such study applied this method in the field of economics in order to map the relations among economics journals [McCain, 1991; Bellis, 2009].

Co-word analysis

Bibliometrics is however not restricted to using only citations. Research similar to the bibliographic coupling and co-citation analysis has been done using also the text associated with articles, such as article titles, keywords, abstracts and index words. Co-word analysis has been proposed and used as a tool for mapping the structure of scientific disciplines [He, 1999; Whittaker, 1989] and as an alternative to direct citations and co-citations [Leydesdorff, 1989]. Using words instead of citations has several advantages. Words are meaningful and ubiquitous and unlike citations they don't take time to accumulate [Leydesdorff, 1989]. Using words also requires less assumptions than using citations [Bellis, 2009]. Co-word analysis has on the other hand been criticised for example because of the varying quality of used keywords and index words [He, 1999] or because single words used in the analysis lack the meaning of the context [Leydesdorff, 1989].

2.2.2 Bibliometrics today

The recent developments in bibliometrics can be described by two main ideas. The journal impact factor which was mentioned in the previous section was developed to be an evaluation tool for comparing different journals in a research field [Garfield, 1972] but has been used to evaluate individual researchers as well, for which it has been repeatedly criticised [Seglen, 1997; The PLoS Medicine Editors, 2006; Brumback, 2009]. Many researchers have also voiced concerns about the reliability of the JIF for any kind of evaluations [Moed and van Leeuwen, 1996; Rossner et al., 2007; Arnold and Fowler, 2010]. New metrics and methods have been developed in the recent years for evaluating individual researchers and journals. These metrics aim at improving or replacing the JIF. Some will be mentioned in this chapter, including probably the best known *h-index* for evaluating researchers and *Eigenfactor* for evaluating journals.

The second important idea that drives the recent developments in bibliometrics is the question of the validity of using citations for research evaluation. The reasons for this are for example the fact that the distribution of citations to scientific articles is very skewed and a significant part of scientific articles is cited very few times or even not cited at all [Seglen, 1992; Campbell, 2008] Other reason is the fact that motives for citing vary and might not always be positive [Nicolaisen, 2007]. Scientometrics and bibliometrics of course don't rely only on citation data. Previous section has introduced co-word analysis which is based on words and texts found in scientific publications. Other data that might be used in research evaluation are for example authorship (productivity), funding, author's association with prestigious institute, qualitative indicators such as peer review, etc. Some additional indicators can be found for example in [Holste et al., 2011]. Moreover, additional data became available in the recent years which weren't available in the past. These include web usage statistics, full-text content of publications, etc.

Evaluation of researchers

Some issues of JIF have been discussed above, particularly it's inappropriateness as a tool for evaluation of individuals. These reasons have pushed the development of similar metrics which would be more suitable for evaluation of individuals.

Probably the best known example is the *h-index* [Hirsch, 2005]. A researcher will be assigned the value h if h of his or her publications have each received $\geq h$ citations and all the remaining publications have received $\leq h$ citations [Hirsch, 2005]. H-index thus expresses the number of core highly cited publications of a researcher. This metric has several advantages, for example it overcomes some issues of simple metrics such as citation counts, average citations per paper, etc. [Hirsch, 2005; Rousseau, 2008]. It however also has several drawbacks, for example it is field dependent and it also creates a disadvantage toward young scientists [Rousseau, 2008].

For these reasons several variants and replacements of the h-index have been proposed. For example the m quotient divides h-index by the number of years that the scientist has been active [Hirsch, 2005; Bornmann and Daniel, 2009] and the *g-index* is calculated as the highest number g of papers that receive in total at least g^2 citations [Egghe, 2006]. Many more alternatives are listed in [Rousseau, 2008] and [Bornmann and Daniel, 2009].

Evaluation of journals

In the recent years there have been many attempts at creating new, more robust metrics that would complement or replace the JIF. A well known example is the Eigenfactor[1] [Bergstrom, 2007]. The Eigenfactor algorithm works similarly as Google's search algorithm PageRank, the citations pointing to a journal are counted and weighted based on the ranking of the source journal. The source journal ranking is further normalised by the total number of citations that appear in that journal [Bergstrom, 2007]. The Eigenfactor metric thus overcomes one major issue of the JIF – the fact that JIF treats all citations as equal.

A similar metric to the Eigenfactor is the SCImago Journal Rank (SJR). Just as the Eigenfactor, SJR weights the incoming citations based on the rank of the source journal, so that citation from prestigious journals contribute more to the final rank than citations from less significant journals [Butler, 2008]. The difference between the two metrics are the underlying data, the Eigenfactor uses the Thomson Reuters Web of Science database [University of Washington, 2012] (which includes the Science Citation Index originally created by Eugene Garfield) while the SJR uses Elsevier's Scopus database [Scimago Lab, 2007].

Field normalisation of indicators

One concern about citation-based evaluation measures is the fact that citation patterns differ significantly across fields. Such measures can then produce unfair evaluations. For example, biochemical papers often contain much more references than mathematical papers which in turn leads to higher citation counts of biochemical papers [Moed, 2011]. This however doesn't mean that research in biochemistry is more important than mathematical research. These reasons led to the development of various normalisation methods which aim to decrease or even eliminate the differences between fields.

Li et al. [2013] divide the normalisation approaches into two main categories: **target-based normalisation** approaches which are functions of the cited papers and **source-based normalisation** approaches which are functions of the citing papers. The target-based normalisation approaches include for example normalisation using the average value for a given field

[1] http://www.eigenfactor.org/

[Li et al., 2013], while the source-based approaches include methods such as fractional citation counting (FCC) [Leydesdorff and Opthof, 2010] and source normalised impact per paper (SNIP) [Moed, 2010]. A comparison of several different normalisation approaches can be found in two recent studies [Waltman and van Eck, 2013; Li et al., 2013].

Patents

An open question is the analysis of interactions between science and technology (S&T). At the forefront of this research are patents as valuable data for mapping the connections between S&T. A patent is a legal document which, for a limited period of time, provides its owner with exclusive rights to an invention. It is granted by a government patent office in exchange for the public disclosure of the technical details of the invention. Patents can be considered a special type of scientific publications. Just as other research articles they contain a list of references which can be used to trace previous work that the invention is built upon. The similarities between patent citations and paper citations have been studied [Narin, 1994; Meyer, 2000] and patent citations have been utilised for assessing usefulness of scientific contributions [Meyer, 2003; Callaert et al., 2006]. Patents have also been used in evaluation studies to compare firms, institutions and countries, or to map evolution of disciplines [Milanez et al., 2013].

Bibliometrics online

The popularisation of the Internet in the 1990s brought many new possibilities to bibliometric research. First and foremost, much of the scholarly literature (if not all) is now published and disseminated online. This change has brought about the creation of online databases and automatic indexing services, which aim at collecting and storing data from various online journals, repositories and libraries and making this data easily accessible and searchable.

Many of these databases and indexing services are field specific, these include for example the NASA Astrophysics Data System[2] for astronomy and physics, Chemical Abstracts Service[3] for chemistry, PubMed[4] for life sciences and biomedical research, INSPIRE HEP[5] for high energy physics,

[2]http://adswww.harvard.edu/
[3]https://www.cas.org/
[4]http://www.ncbi.nlm.nih.gov/pubmed/
[5]http://inspirehep.net/

ArXiv[6] mainly for physics and mathematics and CiteSeerX[7] for computer science.

The recent years have also seen the birth of big multidisciplinary indexes (and thus competitors of the Thomson Reuters Science Citation Index). The three big names in this area are the free services Google Scholar[8] and Microsoft Academic Search[9], and commercial Scopus[10] (maintained by Elsevier). These new services provide, along with citation data, also a set of bibliometric indicators. Google Scholar and Scopus for example both calculate the h-index.

Because the algorithms and the source databases used by these services for indexing might differ (and thus also the citation data), it makes sense that the bibliometric indicators provided by these services often yield different results. A very brief comparison of Google Scholar and Microsoft Academic Search has recently been published in *Nature* [Butler, 2011]. Other studies have analysed Google Scholar and Microsoft Academic Search as free alternatives to the commercial SCI and Scopus [Meho and Yang, 2007; Aguillo, 2011; Jacso, 2011].

Open Access publishing

Open Access (OA) is the practice of providing free unrestricted access to scholarly literature. In contrast to the traditional subscription based journal literature, OA removes fees for accessing the literature as well as most copyright and licensing restrictions. OA was defined in three public statements, the Budapest Open Access Initiative [2002] (BOAI), the Bethesda Statement [2003] and the Berlin Declaration [2003].

The BOAI declaration states that OA means "free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose." An important part of the statement is that it's possible to harvest the full text of the articles and process them automatically using computer software. This means the article full texts can be used in bibliometric analyses. Furthermore, the UK government has recently accepted a policy which states that by 2014, all publicly funded research has to be published as OA [Research Councils UK, 2012] using either gold (publishing in OA journals) or green

[6] http://arxiv.org/

[7] http://citeseerx.ist.psu.edu/

[8] http://scholar.google.com/

[9] http://academic.research.microsoft.com/

[10] http://www.scopus.com/

(self-archiving in author's institutional repository) OA options.

The proportion of OA literature has been found to be around 20% in 2009 [Björk et al., 2010] while a report from 2013 states, that the proportion of OA articles from 2011 is almost 50% [van Noorden, 2013]. I believe this fast progress in OA publishing is good news for bibliometrics and a great opportunity. The idea of analysing the content of publications and development of quality indicators based on the content has already been explored by researchers, however the growth of OA literature will allow and simplify further analyses.

Full text of articles and document level indicators

The idea of using word and text analysis for evaluation of articles is not new and has roots probably in the co-word analysis. The early research in this area has typically utilised titles, abstracts and keywords. However, the recent growing availability of full texts of articles has enabled researchers to perform more complex analyses of the content.

One visible research topic in this area is analysis of citations with regard to the full text context. For example, a recent study analysed the distribution of citations within scientific documents with regard to the traditional IMRaD structure (Introduction, Methods, Results and Discussion) [Bertin et al., 2013]. Research in the citation context analysis can be used to assign citations different weights. Other researchers have used the full text to predict future influence of papers [Yan et al., 2012] or to evaluate research proposals [Holste et al., 2011]. Glenisson et al. [2005] have compared full texts and abstracts of articles for clustering tasks and developed a hybrid approach for mapping scientific articles using both full texts and classical bibliometric indicators.

2.2.3 Webometrics

Webometrics is a relatively new research area which has, since its birth, attracted much attention. Webometrics has first been formally described in 1997 as the use of informetric and bibliometric approaches with online data [Almind and Ingwersen, 1997] in order to map the structure and usage patterns of the web.

The underlying idea behind webometrics is that it is possible to replace papers and citations in the traditional citation networks with web pages and links between them [Almind and Ingwersen, 1997] (because of the similarity to citations, the links between web pages have sometimes been called "sitations" [Rousseau, 2003]). This then allows to use traditional methods, such as analyses of co-citation and bibliographic coupling, etc. In addition, the web allows to track the online scholarly communication which offers new ways of assessing how research results are used by scientists, in teaching and by public [Bjrneborn and Ingwersen, 2004].

A famous application of webometrics is the Webometrics Ranking of World Universities[11]. The Open Access publisher PLOS collects and publishes the online usage data of their articles [Patterson, 2009].

Bjrneborn and Ingwersen [2004] listed in their article four main areas of webometric research. Of these four, one (web technology analysis) is concerned mainly with the study of the web itself rather than with its applications in research evaluation. This area therefore won't be mentioned here. The description of the three remaining areas of webometric research follows:

1. **Analysis of the content of Web pages.** This research area includes for example previously mentioned co-word analysis applied on Web pages. This approach has been used by Leydesdorff and Curran [2003] to identify the online connections between industry, universities and government.

2. **Analysis of the link structure.** For example, a simple idea based on the link structure of the web is to evaluate the importance of a web page based on the number of links pointing to this site (so called *inlinks* [Bjrneborn and Ingwersen, 2004]). This idea has been used to design a metric called the web impact factor (WIF) [Ingwersen, 1998] or to compare health web pages [Cui, 1999]. One of the drawbacks of using this idea in evaluation studies is the fact that some research areas might be by nature more online based then others (such as those where production of web pages and services is part of research) [Thelwall, 2007]. The link structure might however be useful for example for science mapping purposes [Harries et al., 2004].

3. **Web usage analysis.** This area includes analysing log files of users' online behaviour. A significant correlation has been found between download counts of research articles and later citation impact [Brody et al., 2006]. Another study has compared 39 scientific impact measures for evaluating journals (based on both citation and online usage data) in order to evaluate how they relate to each other and how well they represent scientific impact [Bollen et al., 2009].

[11] http://www.webometrics.info/

2.2.4 Altmetrics

Altmetrics is the newest research area, which was proposed in 2010 [Priem et al., 2010]. The main trigger for the proposition of the new research area was the growing concern about the use of citation data and unreliable indicators such as the JIF for the evaluation of research, as well as in daily work for information retrieval purposes. The authors proposed to use the data from the web 2.0 tools, such as social networks, online bookmarking services, forums and discussion groups, etc. In contrast to citations, which might take years to accumulate, mentions in social media normally start appearing immediately after publication. Compared to peer review, social media are fast and free. Several services have already been created that track and publish altmetric indicators, for example commercial the Altmetric[12] and the free service ImpactStory[13].

Many altmetric studies have focused on analysing the available sources of data and the potential of their use. Priem and Hemminger [2010] have summarised some data sources which might be mined for impact. These sources include social bookmarking services, reference managers, blogs, microblogs and comments on articles. A review by Taraborelli [2008] focuses on social bookmarking services and reference managers and on how the data from these tools could be used for assessing semantic relevance, popularity or hotness of publications. Other researchers have focused on the coverage of scholarly literature in online social media and on how the online mentions and bookmarks correlate with citations [Thelwall et al., 2013; Haustein et al., 2013]. The study by Haustein et al. [2013] has also analysed the adoption of social media by scientists. Both the coverage of scholarly publications in social media (except for Twitter and Mendeley) and the adoption of social media among scholars has been found low.

2.3 Limitations

The previously mentioned approaches and metrics each provide certain benefits but also suffer from some limitations. In this section I will focus on these issues. Being aware of the limitations is crucial when deciding about which methods to use for research evaluation.

Particularly the drawbacks and limitations of citation-based methods have been already addressed by many authors. An excellent overview of the main problems of using citations and citation-based methods can be found

[12]http://www.altmetric.com/

[13]http://impactstory.org/

in [Seglen, 1992]. In this article, Seglen focuses on various citation distributions: distribution of citations to articles from a single journal, of citations to articles written by the same author, etc. He found all citation distributions to be skewed (in agreement with the bibliometric laws). He argues, that as a consequence of this skewed distribution, citations aren't suited for research evaluation and mentions four main categories of problems:

- **Citation bias** and reasons for citing. In contrast with the basic assumption that researchers refer to previous publications in order to acknowledge work of their colleagues which inspired them or which they are building on, citations might in fact not always serve as a fair evidence of impact or influence. For example, how do researchers select one reference to be cited from multiple possible ones? Do citations serve as a link to the underlying ideas and methods behind or simply as a support for claims?

- **Incomplete journal coverage** and other database and data collection problems. This issue is particularly visible now, when beside the Science Citation Index there are other indexing services available. In Section 2.2.2 the issue of varying coverage of these databases was mentioned. Other problems might be for example misspellings and other database errors, such as incorrect author disambiguation.

- **Author variability.** This includes the issue of a long (particularly in some fields) citation half-life which makes it difficult to evaluate individual authors and their works in a timely manner, or the question of how to distribute citations among the co-authors of the same paper. Furthermore, Seglen [1992] has shown, that the distribution of citations to articles of individual authors is also skewed, which means few publications of one given author will receive many citations while the remaining publications will receive none or very few citations. When assessing the importance and impact of a publication or a researcher a question presents itself, which is how to deal with these differences – should uncited items (or items cites less than a given threshold) be removed from any evaluation?

- **Field effects.** This issue has been briefly discussed in Section 2.2.2. Research has shown that there are (sometimes quite significant) differences between citation patters in different fields. This reflects in varying citation half-life, immediacy index, average number of received citations in a given field, etc. Moreover, larger research fields provide better opportunity for an individual article to get cited.

There are several other aspects that should be considered. I would like to mention particularly the following three:

- **Uncited publications and Matthew effect.** The skewed distribution of citations has already been mentioned several times. The consequence of this skewness is that many publications are never cited at all. The number of uncited publications has been found to be as high as half of all publications found in the SCI [Seglen, 1992; Garfield, 2006]. It has been suggested that the reason behind the skewed distribution might be the so-called *Matthew effect* Merton [1968]. Matthew effect is a phenomenon where rich people get richer while poor people get poorer. In citation terms this means that authors who were already cited in the past accumulate even more citations, while authors who received few or no citations get ignored even though they might be more productive and have higher impact in their research area than their more acknowledged colleagues. Although the evidence of this phenomenon in science might be controversial [Seglen, 1992], it is still an argument worth considering.

- **Manipulation of metrics.** Arnold and Fowler [2010] points out that "when a measure becomes a target, it ceases to be a good measure". An example of this effect is the journal impact factor. It has become one of the key means of evaluation of journals (as well as publications published in them and subsequently also their authors) which has led to various attempts at manipulating it [The PLoS Medicine Editors, 2006; Rossner et al., 2007; Brumback, 2009; Arnold and Fowler, 2010]. Several authors have argued out that the whole research profile of an author, journal or a publication cannot be expressed with a single number [Hirsch, 2005; Neylon and Wu, 2009]. A better approach might be a collection of indicators, each of them pointing to a different aspect of quality or impact.

- **Using journal impact factor for research evaluation.** This issue has been in detail analysed in [Seglen, 1997]. It has been shown that JIF correlates very poorly with the actual citations of individual articles. Moreover, some researchers have reported that they weren't able to replicate the JIF calculation results [Rossner et al., 2007] or the process of selecting citable items for the JIF calculation [The PLoS Medicine Editors, 2006]. A recent review article has listed many other papers criticising the JIF [Bar-Ilan, 2008].

In sections 2.2.3 and 2.2.4 two research areas which aim at utilising data from the web were mentioned. These new approaches work with data that avoid many drawbacks of citations. For example, one significant disadvantage of citations is the time they need to accumulate. A newly published paper therefore cannot be evaluated using the traditional citation-based methods. In contrast to citations, the online data, such as accesses, downloads and mentions on microblogs are often available from the moment that the article is published. They also provide an evidence of how the publications are used. Priem and Hemminger [2010] suggest that "scholars who would not cite an article or add it to their Web pages may bookmark, tweet, or blog it." These new approaches are however also not without problems. The main issues include:

- **Gaming web-based and social metrics.** Just as with citations, it can be expected that once there are new metrics, there will be some attempts at gaming them, both by authors and publishers. This can be achieved quite easily on the web, as it is a matter of few additional downloads, bookmarks or comments. Priem and Hemminger [2010] however point out, that similar attempts have been done in the past, for example to improve search engine results, and have been successfully controlled (although not completely removed).

- **Problems of data collection.** One theoretical problem is the timely collection of data. A large number of Web 2.0 tools which could be potentially mined for impact exists. Furthermore, many new tools are being created and some old have already ceased to exist. Also the collection of data from web search engines poses several problems, for example coverage issues and the question of how to ensure that all potentially relevant data have been retrieved.

- **Adoption of social media by users.** The adoption of social media by users has been addressed in a recent study [Haustein et al., 2013]. The study reported quite low proportion of social media users. As Priem and Hemminger [2010] point out, the proportion of users could potentially also be skewed toward younger users as well as toward more technical disciplines.

- **Accumulated advantage.** The previously mentioned "Matthew effect" could work on the web as well [Priem and Hemminger, 2010].

The main limitation of the approaches utilising full text of articles is in the fact that not all articles have their full text available online. If we want to compare research publications in a given area for evaluation purposes or for example to detect key publications, it's important to have the full set available. This issue will however diminish in the future as the proportion of OA publications increases.

2.4 Defining the gap

This chapter presented an overview of the main approaches and directions in the area of science and research analysis and evaluation. Limitations and drawbacks of the current methods have been mentioned. It can been seen that despite an active research in this area, a widely accepted solution for evaluating the quality of research publications (as well as researchers and journals) is still missing. Also, citations have been for decades used as a proxy to impact, but it is unclear what is the relationship between the impact of a publication (measured by received citations) and the quality of the publication. The gap analysis, identifying the open problems in this area is presented in this section.

As presented in this chapter, many approaches (particularly the citation-based, but also the new web-based) have focused on developing a single number that would represent the publication (e.g. number of received citations, number of downloads, number of bookmarks), the researcher (e.g. h-index, number of patents) or the journal (journal impact factor). As mentioned in Section 2.3, some researchers have however already argued that the complex profile of a publication, researcher or journal cannot be expressed by a single number. This research will therefore focus on creating a set of quality indicators rather than on developing a single measure.

In addition, this chapter has presented various approaches focusing on different data. The data used include authorship, citations, web usage, social media, words and texts. In the past, much attention has been paid particularly to authorship and citations and in the recent years also to web data. One area which hasn't been explored in such detail is the use of words and full text of publications for evaluation purposes.

The reason for this is that until quite recently the full texts haven't been so easily available. In the past many text-analyses have therefore focused on the titles, abstracts and keywords of publications. This has however changed in the recent years with the growth of the online databases and citation indexes and especially now with the rapid growth of Open Access publishing. Particularly this latest change represents a start of a new era of scholarly publishing which will enable further analyses of the article content and in turn also the development of new evaluation methods. Approaches utilising the full-text of articles have the potential of addressing the actual **quality** of a publication, rather than it's impact, for a simple reason: How else can I assess the quality of a publication than by looking at it's content?

Aside of the text, there are definitely certain indicators, that can help in determining the quality of a publication. These indicators include for example:

- A paper is written by a recognised researcher or by a researcher collaborating with other recognised researchers.
- A paper is citing or is being cited outside of its research area.
- A paper is published in a field-specific prestigious journal.

If a research paper satisfies such criteria, one can expect that this paper will probably be of good quality. The full-text analysis can therefore be complemented by these additional indicators.

To summarise this section, the application of full-text analysis in research evaluation represents an interesting but also under-researched area. This research will therefore focus on the application of natural language processing in research evaluation. Some examples of using full-text analysis in this area include:

- Co-word analysis using the publication full-text.
- Analysis of citation context (for example sentiment analysis and analysis of distribution of citations in text).
- Clustering and comparing publications based on their semantic similarity.

To fully represent all quality aspects of research publications, the full-text analysis will be combined with bibliometric indicators, network analysis and other methods.

Chapter 3

Research plan

The main goal of this PhD project is to explore the possibilities of detecting various factors influencing the quality of research publications using methods such as full text analysis (natural language processing), graph and network analysis and bibliometric indicators. In order to achieve this goal the work has been divided into two main phases: (1) finding and selection of the factors influencing the quality of research publications and (2) evaluation of these factors using the above mentioned methods.

The following chapter presents the research proposal and describes the approach that will be used in answering the research questions. The research questions are based on the limitations and gaps identified in the literature review (Chapter 2) as well as on the motivating examples of use presented in the introduction (Chapter 1). Aside of the two main phases the work has been further broken down into eight technical tasks, which will need to be completed in order to answer the research questions. Both the two phases and the eight planned tasks are presented in this chapter.

3.1 Research questions

The research questions have already been presented in Chapter 1 and are repeated here for completeness. They have been formulated into the following four points:

Question 1: What factors influence the quality of a research publication (with regard to the publication type and to the research field)?

Question 2: What is the relationship (if there is any) between the impact of a publication, measured by the classical bibliometric methods, and the quality of a publication?

Question 3: How can we detect the factors influencing quality in order to evaluate the quality of a research publication?

Question 4: How can this evaluation be used in other disciplines, such as in information retrieval, and to detect research fronts and trends in development of research disciplines?

3.2 Approach

The proposed approach is to find a set of features of research publications that contribute to or influence the quality of the publications and to utilise analysis of the full text of the publications, network analysis of their citation and collaboration structure, bibliometric indicators and other methods in order to detect, measure and evaluate these features. There are three key points of this approach which make it a novel solution.

One key point of the selected approach is the fact it aims at developing a collection of metrics and indicators focused at different features instead of developing a single number for the evaluation (which has been indicated as one of the issues of many of the currently used research publication evaluation methods, see Chapter 2.3). The second key point of this approach is the focus on the full text of the publications. The increasing availability of full text of research publications is a quite recent change, which, in our opinion, is one of the reasons why the use of full text for the automatic evaluation of research publications is a relatively under-research area. Finally, it is important to note that our focus is on evaluating and measuring the quality of research publications rather that on other factors, such as their impact on the research field or their utility.

The envisioned research work is divided into two main phases (Phase 1 and Phase 2) with an additional preparatory phase (Phase 0). These three phases have been broken down into eight tasks in total.

Phase 0: Data collection and preparation. This phase constitutes the collection of data needed for the further analyses. It has been broken down into two tasks:

Task 1: Identifying the information sources such as publication repositories, research data repositories and databases of bibliographic information, that may provide relevant publication data. This task has been addressed during the first year of this PhD and the results are presented in Chapter 4.

Task 2: Using these information sources, developing various relevant data structures such as:

- collaboration networks
- citation, co-citation and bibliographic coupling networks
- clusters of semantically related publications
- clusters of publications corresponding to different topics

Phase 1: Feature collection and validation. Analysis factors influencing the quality of research publications. The hypothesis in this phase is that the quality of a research publication can be defined as a set of features. This is one of the two main phases of the research work and consists of a single task.

Task 3: Finding and selecting the features of research publications which contribute to or influence the publication quality. The goal of this task is to develop a set of features which will represent the quality of research publications in an understandable and meaningful way. These features could be for example novelty and applicability. Because the outcome of this task could potentially significantly differ based on the research field, only one or two (for comparison purposes) research fields will be taken into account. Furthermore this should simplify the selection of a smaller set of features suitable for the further analysis. Several steps will be taken in order to accomplish this task:

- Analysis of state of the art directly related to the evaluation of research publications based on their full text. Because peer review is based on the full text, this step might include for example the analysis of the peer review process of different journals and conferences. Research evaluation frameworks of different countries and the criteria used in their grading will also be analysed.
- Based on this analysis collect a set of features of research publications related to the quality of the publications, i.e. candidate features for further analysis. These features might be for example originality (novelty), significance, applicability, interdisciplinarity, etc.
- Identify a set of the main features which could potentially be used for assessing the quality of research publications. The selection and validation of the main features and requirements for these features (the perceived meaning of these features)

will be done through an interview with a sample of researchers representing different career stages (from PhD students to senior researchers). The outcome of this task will be a clearly defined set of features related to quality of research publications.

Phase 2: Feature analysis and evaluation. Development of methods and indicators for measuring and evaluating the selected features. The goal of this phase is to find methods which could be used for assessing individual features. Rather than developing entirely new methods from scratch, the focus will be on utilising and building on existing methods and techniques. This phase is based on the premise that some or all of the quality features found in the previous phase can be assessed using automatic analysis of the full text of publications in combination with other methods, such as analysis of collaboration and citation networks and classical bibliometric indicators. Several tasks will be accomplished in this phase.

Task 4: Studying the possibilities of application of natural language processing, text analysis and computational linguistics for the evaluation of research publications and the selected features. An example application of these methods is the use of semantic similarity of publications within a given area to determine originality. The main goal of this task will be the study of the state of the art methods of text analysis and related areas related to the evaluation of the selected features.

Task 5: Investigating the structures developed in Phase 0 using graph and network theory as well as bibliometric indicators, in order to determine whether the evaluation of the selected features could be supported or enhanced using these methods. For example, interdisciplinarity could be evaluated based on the citation and/or co-authorship networks of the publication.

Task 6: Analysing the possibilities of combining the methods investigated in the Tasks 4 and 5 in order to design a set of new methods for estimating quality of research publications based on the evaluation of the individual features. The goal of this task is to test, select and adapt the methods found in the previous two tasks. The outcome of this task will be a set of methods for evaluating the selected features.

Task 7: Evaluating the proposed methods against current standards. Peer review is often regarded to be probably the most reliable evaluation method, therefore the proposed methods will be evaluated on a selected set of publications against an expert review. It will also be compared with the currently most widely used evaluation metrics such as the journal impact factor of the publications and their citation counts.

Task 8: Analysing the use of the new methods in other disciplines, for example their application in information retrieval, to detect trends in the evolution of scientific disciplines or to detect key publications that trigger important changes.

To the best of our knowledge, we are not aware of any similar approach that would combine text analysis, network analysis and bibliometric indicators in order to assess the quality of research publications.

3.3 Contribution

The expected contributions of this research project are:

- New methods for the evaluation of research publications. These new methods will simplify the process of searching for relevant literature and will aid in determining the key literature in a given area. They will also potentially help in the evaluation of researchers and journals.
- An analysis of the quality features of research publications. This analysis will be useful in any future research evaluation studies.
- A potential application of the new methods in other areas, for example in information retrieval, in the mapping of scientific disciplines, etc.

Taraborelli [2008] lists several requirements for science evaluation methods. These requirements are:

1. The method/metric has to be reliable and accurate, comparable or better than the peer review system.
2. It should be easy to understand.
3. It should be economical in terms of development and maintenance, but also in terms of time required to understand it, etc.
4. It should be faster than citations, at least comparable to the speed of peer review.

5. It should be resistant to manipulation and gaming.

During the development of the new evaluation methods, we would like to follow these requirements.

3.4 Current progress

Research paper relevant to this project has been accepted to and presented at the following conference:

- Knoth, P. and Herrmannova, D. **Simple Yet Effective Methods for Cross-Lingual Link Discovery (CLLD) KMI @ NTCIR-10 CrossLink-2**. NTCIR-10 Evaluation of Information Access Technologies, 2013, Tokyo, Japan

This research project has also been presented at the CRC PhD Student Conference at the Open University:

- Herrmannova, D. **Mining Research Publication Networks for Impact**, CRC PhD Student Conference, 2013, Milton Keynes

In addition, the preliminary work has focused on the completion of the **Task 1**. First, a summary of types of research publications (journal articles, conference proceedings, patents, ...) and a summary of types of relevant available data (citations, library and web usage data, ...) has been created. Following that, an identification, analysis and comparison of the possible data sources has been conducted. This study will be extended in the future to provide an overview of the contents of each dataset in terms of number of publications with full-text, number of citations per publication, publication type statistics, etc.

3.5 Future plans

This section presents work planned for the future which will need to be done in order to answer the research questions. This work consists of three phases and eight tasks that were presented in Section 3.2. In addition, the work includes the following activities:

1. **Literature review**, which will focus on literature relevant to the selected approach and the eight tasks.

2. **Research work** done while completing the eight tasks.

3. **Writing**, which will include preparing the second year progress report, disseminating the research outcomes by publishing at relevant conferences and thesis writing.

The following conferences were recognised as the most suitable for disseminating this research work:

- International Society of Scientometrics and Informetrics conference (ISSI) – a bi-annual conference, the next will be organised in 2015
- Conference on Webometrics, Informetrics and Scientometrics (WIS) – every August/September
- International Conference on Science and Technology Indicators (STI) – every September
- International Conference on Theory and Practice of Digital Libraries (TPDL) – every September
- Joint Conference on Digital Libraries (JCDL) – every June/July
- Open Repositories Conference (OR) – every June/July

Table 3.1 presents the progress plan for the next two years my PhD. The planned work has been divided into several packages according to the three main activities (literature review, research work, writing) and according to the seven tasks to be completed.

Second year	Q1	Q2	Q3	Q4
1. Literature review	X	X	X	X
2. Research work	X	X	X	X
2a. Task 1 (data collection)	X	X	X	
2b. Task 2 (data preparation)	X	X	X	
2c. Task 3 (feature selection)	X	X	X	X
2d. Task 4 (full text analysis)		X	X	X
2e. Task 5 (data analysis)		X	X	X
2f. Task 6 (development of new methods)			X	X
2g. Task 7 (evaluation of the proposed methods)				X
3. Writing	X	X	X	X
3a. Second year progress report				X
3b. Conference submission	X			
3c. Conference dissemination		X	X	
Third year	Q1	Q2	Q3	Q4
2. Research work	X	X		
2c. Task 4 (full text analysis)	X			
2d. Task 5 (data analysis)	X			
2e. Task 6 (development of new methods)	X	X		
2f. Task 7 (evaluation of the proposed methods)	X	X		
2g. Task 8 (application in other disciplines)	X	X		
3. Writing	X	X	X	X
3b. Conference submission	X			
3c. Conference dissemination		X	X	
3d. Thesis writing		X	X	X
3e. Thesis validation and correction				X

Table 3.1: Progress plan for the second and third year.

Chapter 4

Pilot study

The following chapter presents the work that has been done so far. Apart of the literature review, the previous work has been concerned with the **Task 1** (presented in Chapter 3.1). To assess the performance of research evaluation methods, several databases and datasets are available. This chapter provides an overview of eleven databases and datasets. The overview includes information about APIs and data downloads, size of the database or dataset and information whether the source contains citations and full-text.

4.1 Types of scientific publications

When talking about scientific publications one might think about two questions:

1. What is a scientific publication?
2. What types of scientific publications exist?

Scientific publication is a type of publication, aim of which is to present and distribute some original scientific work, for example academic research and scholarship. The following list summarises the main types of scientific publications. This list was compiled with the help of BibTeX entry types documentation [Patashnik, 1988].

- **Journal articles** and magazine articles are peer reviewed publications appearing in online or printed journals and magazines. In many research areas journal literature is the most important mean of communication and dissemination of research. There are three main types of journal articles:

1. **Research article** is a so-called *primary source*, it typically reports results from one or more studies and it's normally written from the perspective of the person who conducted the experiment.
2. **Review article** is a *secondary source*, it provides an overview of recent advancement in science, but not any original research.
3. **Editorials, letters, etc.** are as well a *secondary source*. These articles provide expert opinions, observational studies, comments, discussions, etc.

- **Conference proceedings** is a collection of peer reviewed research papers presented at a conference. In some research fields conference proceedings are the main way of communication (particularly in computer science).
- **Books & book chapters** include scientific and research books written by one or few authors or books where each chapter is written by a different author.
- **Theses** include both Master's and PhD theses. These documents represent authors research and findings conducted in pursuit of an academic degree.
- **Patents** are legal document which describe an invention (a product or a process) and which provide its owner with exclusive rights to the invention.
- **Government reports** are documents published by a government agency which provide for example details of an investigation.
- **Project proposals, technical reports and working papers** issued either by individual researchers or by organisations.
- **Presentations** presented at workshops, seminars or academic conferences.
- **Online scientific publications**, for example preprints and other research articles published online, for example on a personal web page or in an online self-archiving repository.
- **Blogs** are short articles published in online blogs which might contain opinions and ideas as well as research.

Source	multidisc.	API	OAI-PMH	data dumps	cit.	full-text
CSX	-	-	X	-	X	X
MAS	X	X	-	-	X	*
JSTOR	X	-	-	X	X	*
DBLP	-	X	-	X	-	*
CORE	X	X	X	X	X	X
ArXiv	X	X	X	X	-	X
KDD	-	-	-	X	X	X
iSearch	-	-	-	X	X	X
DBLP+C	-	-	-	X	X	-
ACM	-	-	-	X	X	-
OCC	-	-	-	X	X	-

Table 4.1: Overview of research evaluation data sources. The stars (*) in the table represent sources, which don't store full-text but provide links to the full-text of articles where available. The first column provides information whether the data source covers multiple disciplines.

4.2 Research publication data sources

In this section twelve data sources which can be used for research evaluation studies will be presented. The data sources for this study were selected according the following criteria:

1. The data source has to be publicly available to the research community. This excluded both major databases, Thomson Reuters Science Citation Index and Elsevier Scopus, as these are both commercial.

2. The data should be available for download, for example through an API, an OAI-PMH endpoint or through data dumps. This excludes the third major database Google Scholar, which is publicly available for searching, but there is no API or data dumps available for download and Google forbids crawling of the search service.

In the following subsections, an overview of each of the data sources is provided. The section is divided into two subsections, the first subsection lists several databases and autonomous indexes, while the second subsection provides a list of prepared dataset of fixed size which can be downloaded and directly used. Table 4.1 provides a summary of the main features of the data sources.

4.2.1 Databases

CiteSeerX

CiteSeerX[1] is an public online search engine, digital library and a citation index, which focuses mainly on computer and information science. It crawls and harvests publicly available documents from the web and automatically extracts metadata and citations from these documents. CiteSeerX provides an OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) endpoint through which the CiteSeerX data can be harvested.

Microsoft Academic Search

Microsoft Academic Search[2] (MAS) is a multidisciplinary digital library, search engine and citation index. It provides an API for accessing its data (for non-commercial purposes). The API documentation can be found online[3]. The use of the API is restricted to a limited number of queries per minute. The API also cannot be used to crawl the entire database. Citations and references can be retrieved through the API but MAS doesn't store the full-text of publications. It however provides a link to the full-text (where available). According to the MAS About page[4] the corpus currently contains more that 45 million publications. MAS data were used for developing the Eigenfactor project[5]. A review of the use of MAS for bibliometric analyses was presented in [Jacso, 2011].

JSTOR

JSTOR[6] is a multidisciplinary digital library which provides access to books and journals. JSTOR provides it's data for non-commercial purposes through bulk data downloads. The data downloads can be requested via an online tool Data For Research[7] (DFR) which allows querying the JSTOR corpus and defining the content to be downloaded. Initially the downloads are limited to 1000 items per download, but larger downloads can be requested. The data downloads contain citations and additional information, such as key terms, however not full-texts. At present the size of JSTOR dataset is almost 9

[1] http://citeseerx.ist.psu.edu/

[2] http://academic.research.microsoft.com/

[3] http://academic.research.microsoft.com/about/Microsoft Academic Search API User Manual.pdf

[4] http://academic.research.microsoft.com/About/Help.html

[5] http://mas.eigenfactor.org/

[6] http://www.jstor.org/

[7] http://dfr.jstor.org/

million publications. The JSTOR dataset was used for example in study by Shi et al. [2010].

DBLP Computer Science Bibliography

DBLP[8] (or The DBLP Computer Science Bibliography) is an online bibliography of computer science research. It indexes metadata of books and documents from journals, conferences, etc. It doesn't store citations or full-texts, however the metadata contain links to the articles. The DBLP data are released under the ODC-BY 1.0 license, which means they can be freely used as long as their public use is attributed. The DBLP database can be accessed in two ways, one option is a simple API (described in [Ley, 2009]). Alternatively the whole DBLP database can be downloaded in one XML file[9]. According to the DBLP website, the size of the dataset is now more than 2.3 publications[10].

CORE

CORE (COnnecting REpositories) is an aggregator of content stored in Open Access repositories. Besides harvesting and storing the content it provides additional services, such as a citation index and calculation of semantic similarity of publications. All CORE data are available under some Open Access compatible license. CORE data include publication full-texts (where available) in both PDF and text formats. The data can be accessed via an API[11] which allows retrieving single documents, keyword searching, searching for similar documents and other methods. Furthermore the CORE dataset can be obtained through a bulk download[12] and through an OAI-PMH endpoint (currently in beta version). At present, the CORE dataset contains over 15 million metadata records out of which almost 2 million documents contain a PDF.

ArXiv

ArXiv[13] is an online self-archiving repository for research articles. It covers many fields including physics, mathematics, biology and computer science.

[8] http://www.informatik.uni-trier.de/~ley/db/

[9] http://www.informatik.uni-trier.de/~ley/faq/How+can+I+download+the+whole+dblp+dataset.html

[10] http://www.informatik.uni-trier.de/~ley/faq/What+is+dblp.html

[11] http://core.kmi.open.ac.uk/api/doc

[12] http://core.kmi.open.ac.uk/intro/data_dumps

[13] http://arxiv.org/

The ArXiv data are available under various licenses (depending on the choice of the author), the most common one states that ArXiv is only permitted to distribute the articles but grants no additional rights[14]. ArXiv data can be accessed through various methods. ArXiv provides an OAI-PMH endpoint, an API and an RSS feed for accessing metadata of articles[15] (which include a link to the article full-text), and bulk download of the PDF files[16]. The size of the ArXiv dataset is now almost 900,000 documents.

4.2.2 Datasets

KDD cup Dataset

KDD cup dataset[17] is a subset of ArXiv documents from 1992 until 2003. The dataset contains documents from the high energy particle physics (HEP) section of ArXiv. The dataset includes both article full-texts and extracted citations, there are around 29,000 documents in the dataset.

iSearch collection

The iSearch dataset[18] contains physics documents collected from open archives and from public library. The dataset contains both full-texts and extracted citations. There are around 450,000 documents in the collection out of which 150,000 contain full-text in PDF.

DBLP+Citation

DBLP+Citation[19] is a dataset released by ArnetMiner.org[20], which consists of DBLP articles enriched with abstracts and citations from ArnetMiner. The latest version of the dataset (from September 2013) contains over 2 million articles and can be freely downloaded (for research purposes only) from the website.

[14]http://arxiv.org/help/license
[15]http://arxiv.org/help/bulk_data
[16]http://arxiv.org/help/bulk_data_s3
[17]http://www.cs.cornell.edu/projects/kddcup/datasets.html
[18]http://itlab.dbit.dk/ isearch/
[19]http://arnetminer.org/DBLP_Citation
[20]http://arnetminer.org/

ACM dataset

The ACM dataset[21] contains information about publications from the Association for Computing Machinery (ACM) digital library. The dataset is available in RDF for download[22] and can also be browsed using an online interface for navigating RDF triples[23]. It contains information about publications including citations, it doesn't however contain publication full-texts. The ACM dataset for utilised in the study by Shi et al. [2010].

Open Citation Corpus

The Open Citation corpus[24] contains citations extracted from open access articles in the PubMed central database (biomedical domain). This corpus can be freely downloaded (either in RDF or in BibJSON format) and reused or browsed online[25]. The dataset doesn't contain article full-texts.

4.3 Conclusion

This chapter reviewed several available data sources. Each data source provides data covering different disciplines and the data sources also differ in terms of whether citations and full-texts are offered. As a next step we would like to perform a comparative study in terms of the citation density and the proportion of full-text articles provided by each data source.

[21] http://datahub.io/dataset/rkb-explorer-acm
[22] http://acm.rkbexplorer.com/models/dump.tgz
[23] http://acm.rkbexplorer.com/
[24] http://opencitations.net/
[25] http://opencitations.net/source-data/

Chapter 5

Summary

The growing amount of scholarly literature and the limitations of current solutions for evaluating research publications require the development of new evaluation methods. Furthermore, the recent changes, such as the growth of Open Access publishing, create an opportunity for finding new methods utilising semantically richer data than the traditionally used citations.

This report reviewed the classical as well as the new methods used in this area, including their advantages and drawbacks. Following this review, a gap analysis was presented and research questions were outlined. Along with the research questions, several tasks were listed. Completion of these tasks will help in answering the research questions. Finally, a pilot study was introduced, which summarises the available datasets and data sources.

My work during the next two years will focus mainly (1) on understanding which factors influence the quality of a publication and what is the relationship between citation impact and the quality and (2) on finding a way to express these factors in a measure or a set of measures that could be used for assessing the quality of a publication.

Bibliography

Aguillo, Isidro F. Is Google Scholar useful for bibliometrics? A webometric analysis. *Scientometrics*, 91(2):343–351, December 2011.

Almind, Tomas C. and Ingwersen, Peter. Informetric analyses on the world wide web: methodological approaches to webometrics. *Journal of Documentation*, 53(4):404–426, 1997.

Arnold, Douglas N and Fowler, Kristine K. Nefarious numbers. *Notices of the American Mathematical Society*, 58(3):434–437, 2010.

Bar-Ilan, Judit. Informetrics at the beginning of the 21st century A review. *Journal of Informetrics*, 2:1–52, 2008.

Bellis, Nicola De. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. The Scarecrow Press, Inc., Lanham, Maryland, USA, 2009.

Bergstrom, Carl. Eigenfactor: Measuring the value and prestige of scholarly journals. *C&RL News*, 68(5):314–316, 2007.

Berlin Declaration, . Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. `http://openaccess.mpg.de/286432/Berlin-Declaration`, 10 2003. Accessed: 24/10/2013.

Bertin, Marc; Atanassova, Iana; Lariviere, Vincent, and Gingras, Yves. The distribution of references in scientific papers: An analysis of the IMRaD structure. In *Proceedings of the 14th ISSI Conference*, pages 591–603, Vienna, Austria, 2013.

Bethesda Statement, . Bethesda Statement on Open Access Publishing. `http://legacy.earlham.edu/~peters/fos/bethesda.htm`, 6 2003. Accessed: 24/10/2013.

Björk, Bo-Christer; Welling, Patrik; Laakso, Mikael; Majlender, Peter; Hedlund, Turid, and Gudnason, Gudni. Open access to the scientific journal literature: situation 2009. *PloS one*, 5(6):e11273, January 2010.

Bjrneborn, Lennart and Ingwersen, Peter. Toward a Basic Framework for Webometrics. *Journal of the American Society for Information Science and Technology*, 55(14):1216–1227, December 2004.

Bollen, Johan; Van de Sompel, Herbert; Hagberg, Aric, and Chute, Ryan. A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE*, 4(6):e6022, 2009.

Bornmann, Lutz and Daniel, Hans-Dieter. The state of h index research. *EMBO reports*, 10(1):2–6, December 2009.

Bradford, Samuel C. Sources of information on specific subjects. *Engineering*, 137(3550):85–86, 1934.

Brody, Tim; Harnad, Stevan, and Carr, Leslie. Earlier Web Usage Statistics as Predictors of Later Citation Impact. *Journal of the American Society for Information Science & Technology*, 57(8):1060–1072, March 2006.

Brumback, Roger A. Impact factor wars: Episode V – The Empire Strikes Back. *Journal of child neurology*, 24(3):260–2, March 2009.

Budapest Open Access Initiative, . Read the Budapest Open Access Initiative. `www.budapestopenaccessinitiative.org/read`, 2 2002. Accessed: 24/10/2013.

Butler, Declan. Free journal-ranking tool enters citation market. *Nature*, 451 (7174):6, 2008.

Butler, Declan. Computing giants launch free science metrics. *Nature*, 476: 18, 2011.

Callaert, Julie; Looy, Bart Van; Verbeek, Arnold; Debackere, Koenraad, and Thijs, Bart. Traces of Prior Art: An analysis of non-patent references found in patent documents. *Scientometrics*, 69(1):3–20, 2006.

Campbell, Philip. Escape from the impact factor. *Ethics in Science and Environmental Politics*, 8:5–7, June 2008.

Cole, Francis J. and Eales, Nellie B. The history of comparative anatomy. part i: A statistical analysis of the literature. *Science Progress*, 11(43): 578–596, 1917.

Cui, Lei. Rating Health Web sites using the principles of Citation Analysis: A Bibliometric Approach. *Journal of Medical Internet Research*, 1(1):e4, 1999.

Dieks, Dennis and Chang, Hans. Differences in Impact of Scientific Publications: Some Indices Derived from a Citation Analysis. *Social Studies of Science*, 6(2):247–267, 1976.

Egghe, Leo. An improvement of the h-index: the g-index. *ISSI newsletter*, 2 (1):8–9, 2006.

Garfield, Eugene. Citation indexes for science. A new dimension in documentation through association of ideas [2006 reprint]. *Science*, 122(3159): 108–11, October 1955.

Garfield, Eugene. Citation analysis as a tool in journal evaluation. *Science*, 178(60):471–9, 1972.

Garfield, Eugene. The Agony and the Ecstasy The History and Meaning of the Journal Impact Factor, 2005.

Garfield, Eugene. The History and Meaning of the Journal Impact Factor. *JAMA: the journal of the American Medical Association*, 295(1):90–93, 2006.

Glenisson, Patrick; Glanzel, Wolfgang, and Persson, Olle. Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics*, 63(1):163–180, 2005.

Harries, Gareth; Wilkinson, David; Price, Liz; Fairclough, Ruth, and Thelwall, Mike. Hyperlinks as a data source for science mapping. *Journal of Information Science*, 30(5):436–447, October 2004.

Haustein, Stefanie; Peters, Isabella; Bar-Ilan, Judit; Priem, Jason; Shema, Hadas, and Terliesner, Jens. Coverage and adoption of altmetrics sources in the bibliometric community. In *Proceedings of 14th ISSI Conference*, pages 468–483, Vienna, Austria, 2013.

He, Qin. Knowledge Discovery Through Co-Word Analysis. *Library Trends*, 48(1):133–159, 1999.

Hirsch, Jorge E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–72, November 2005.

Holste, Dirk; Roche, Ivana; Hörlesberger, Marianne; Besagni, Dominique; Thomas Scherngell, ; Francois, Claire; Cuxac, Pascal, and Schiebel, Edgar. A concept for Inferring "Frontier Research" in Research Project Proposals. In *Proceedings of the 13th ISSI*, pages 315–326, Durban, South Africa, 2011.

Hood, William W. and Wilson, Concepcin S. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2):291–314, 2001.

Ingwersen, Peter. The Calculation of Web Impact Factors. *Journal of Documentation*, 54(2):236–243, 1998.

Jacso, Peter. The pros and cons of Microsoft Academic Search from a bibliometric perspective. *Online Information Review*, 35(6):983–997, 2011.

Kessler, M. M. Bibliographic Coupling Between Scientific Papers. *American Documentation*, 14(1):10–25, 1963.

Kyrillidou, Martha and Morris, Shaneka. ARL Statistics 20082009. Technical report, Association of Research Libraries, Washington, DC, 2011.

Ley, Michael. Dblp xml requests. `http://dblp.uni-trier.de/xml/docu/dblpxmlreq.pdf`, 2009.

Leydesdorff, Loet. Words and co-words as indicators of intellectual organization. *Research Policy*, 18(4):209–223, 1989.

Leydesdorff, Loet and Curran, Michael. Mapping University-Industry-Government Relations on the Internet: The Construction of Indicators for a Knowledge-Based Economy. *Cybermetrics*, 4(1):1–17, 2003.

Leydesdorff, Loet and Opthof, Tobias. Normalization at the field level: Fractional counting of citations. *Journal of Informetrics*, 4:644–646, 2010.

Li, Yunrong; Radicchi, Filippo; Castellano, Claudio, and Ruiz-castillo, Javier. Quantitative Evaluation of Alternative Field Normalization Procedures. In *Proceedings of the 14th ISSI Conference*, pages 1431–1444, Vienna, Austria, 2013.

Lotka, Alfred J. The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*, 16:317–323, 1926.

McCain, Katherine W. Mapping Economics through the Journal Literature: An Experiment in Journal Cocitation Anaiysis. *Journal of the American Society for Information Science*, 42(4):290–296, 1991.

Meho, Lokman I and Yang, Kiduk. Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13):2105–2125, 2007.

Merton, Robert K. The Matthew Effect in Science. *Science*, 159(3810):56–63, 1968.

Meyer, Martin. What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, 49(1):93–123, 2000.

Meyer, Martin. Academic patents as an indicator of useful research? A new approach to measure academic inventiveness. *Research Evaluation*, 12(1): 17–27, 2003.

Milanez, Douglas H.; Faria, Leandro I. L.; do Amaral, Roniberto M., and Gregolin, Jose A. R. Patents in nanotechnology: An analysis using macro-indicators and forecasting curves. In *Proceedings of the 14th ISSI Conference*, pages 1363–1378, Vienna, Austria, 2013.

Moed, Henk F. Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4(3):265–277, 2010.

Moed, Henk F. The Source Normalized Impact per Paper Is a Valid and Sophisticated Indicator of Journal Citation Impact Dear. *Journal of the American Society for Information Science and Technology*, 62(1):211–213, 2011.

Moed, Henk F and van Leeuwen, Thed. Impact factors can mislead. *Nature*, 381:186–186, 1996.

Nalimov, Vassily V. and Mulchenko, Z. M. Naukometriya. izuchenie razvitiya nauki kak informatsionnogo protsessa. [scientometrics. study of the development of science as an information process]. *Nauka. (English translation: 1971. Washington, DC: Foreign Technology Division. US Air Force Systems Command, Wright-Patterson AFB, Ohio. (NTIS Report No. AD735-634))*, 34:107–247, 1969.

Narin, Francis. Patent bibliometrics. *Scientometrics*, 30(1):147–155, 1994.

Neylon, Cameron and Wu, Shirley. Article-level metrics and the evolution of scientific impact. *PLoS Biology*, 7(11):1–6, November 2009.

Nicolaisen, Jeppe. Citation Analysis. *Annual Review of Information Science and Technology*, 41(1):609–641, 2007.

Panitch, Judith M. and Michalak, Sarah. The Serials Crisis. Technical report, A White Paper for the UNC-Chapel Hill Scholarly Communications Convocation, January 2003. Accessed: 2013-10-04.

Patashnik, Oren. *Bibtexing*, 1988.

Patterson, Mark. Article-level metrics at PLoS addition of usage data. http://blogs.plos.org/plos/2009/09/article-level-metrics-at-plos-addition-of-usage-data/, 2009. Accessed: 26/10/2013.

Price, Derek J. de Solla. Networks of Scientific Papers. *Science*, 149(3683): 510–515, 1965.

Priem, Jason and Hemminger, Bradely M. Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday*, 15(7), July 2010.

Priem, Jason; Taraborelli, Dario; Groth, Paul, and Neylon, Cameron. Altmetrics: A manifesto. http://altmetrics.org/manifesto/, October 2010.

Pritchard, Alan. Statistical Bibliography or Bibliometrics? *Journal of Documentation*, 25(4):348–349, December 1969.

Redner, Sidney. Citation Statistics from 110 Years of Physical Review. *Physics Today*, 58(June):49–54, 2005.

Research Councils UK, . RCUK Policy on Open Access. http://www.rcuk.ac.uk/research/Pages/outputs.aspx, 2012. Accessed: 24/10/2013.

Rossner, Mike; Epps, Heather Van, and Hill, Emma. Show me the data. *The Journal of Cell Biology*, 179(6):1091–1092, 2007.

Rousseau, Ronald. Sitations: an exploratory study. *Cybermetrics*, 1(1):1–7, 2003.

Rousseau, Ronald. Reflections on recent developments of the h-index and h-type indices. In *Proceedings of WIS 2008, Fourth International Conference on Webometrics, Informetrics and Scientometrics*, pages 1–8, Berlin, Germany, June 2008. Humboldt-Universitat zu Berlin.

Sample, Ian. Harvard University says it can't afford journal publishers' prices. *The Guardian*, 04 2012. Accessed: 2013-10-04.

Scimago Lab, . SJR SCImago Journal & Country Rank, 2007. Retrieved October 21, 2013 from http://www.scimagojr.com/.

Seglen, Per Ottar. The Skewness of Science. *Journal of the American Society for Information Science*, 43(9):628–638, October 1992.

Seglen, Per Ottar. Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal*, 314(February): 498–502, 1997.

Shi, Xiaolin; Leskovec, Jure, and Mcfarland, Daniel A. Citing for High Impact. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries - JCDL '10*, page 49, New York, New York, USA, 2010. ACM Press.

Small, Henry. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science*, 24(4):265–270, 1973.

Sutherland, William J.; Goulson, David; Potts, Simon G., and Dicks, Lynn V. Quantifying the Impact and Relevance of Scientific Research. *PLoS One*, 6(11):e27537, January 2011.

Taraborelli, Dario. Soft peer review: Social software and distributed scientific evaluation. In *Proceedings of the 8th International Conference on the Design of Cooperative Systems (COOP '08)*, Carry-le-Rouet, France, May 2008.

Thelwall, Mike. Bibliometrics to Webometrics. *Journal of Information Science*, 34(4):1–18, 2007.

Thelwall, Mike; Haustein, Stefanie; Larivière, Vincent, and Sugimoto, Cassidy R. Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE*, 8(5):e64841, May 2013.

The PLoS Medicine Editors, . The impact factor game. *PLoS medicine*, 3 (6), June 2006.

University of Washington, . Eigenfactor.org faq, 2012. Retrieved October 21, 2013 from http://www.eigenfactor.org/faq.php.

van Noorden, Richard. Half of 2011 papers now free to read. *Nature*, 500: 386–387, 2013.

Waltman, Ludo and van Eck, Nees Jan. A Systematic Empirical Comparison of Different Approaches for Normalizing Citation Impact Indicators. In *Proceedings of the 14th ISSI Conference*, pages 1649–1664, Vienna, Austria, 2013.

White, Howard D. and Griffith, Belver C. Author Cocitation: A Literature Measure of Intellectual Structure. *Journal of the American Society for Information Science*, 32(3):163–171, 1981.

Whittaker, John. Creativity and Conformity in Science: Titles, Keywords and Co-Word Analysis. *Social Studies of Science*, 19(3):473–496, 1989.

Yan, Rui; Huang, Congrui; Tang, Jie; Zhang, Yan, and Li, Xiaoming. To Better Stand on the Shoulder of Giants. In *Proceedings of the 12th Joint Conference on Digital Libraries*, pages 51–60, Washington, DC, 2012. ACM.

Zipf, George K. The psycho-biology of language: an introduction to dynamic philology, 1935.