



Universiteit Leiden

ICT in Business

Identification of Essential References Based on the Full Text of Scientific Papers and Its Application in Scientometrics

Name: **Xi Cui**
Student-no: **s1242156**
Date: **25/08/2014**

1st supervisor: **Dr. Nees Jan van Eck (CWTS)**
2nd supervisor: **Dr. Hans Le Fever (LIACS)**

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my supervisors for their guidance and critical view on my thesis. Special thanks should be given to dr. Nees Jan van Eck, who gave me quite a lot of guidance and useful advices through the whole process of my research. With his help and patience, I have learned how to carry out a complex project in a structured and professional way. I also would like to thank dr. Hans Le Fever, who not only supported me with my thesis but also helped me during the two years of study in ICT in Business.

I would also like to thank the Centre for Science and Technology Studies (CWTS) at Leiden University, for providing me this research opportunity and all the necessary technical support. Especially thanks to Henri de Winter for developing the web survey for this study. I am also grateful to the participants of this survey, who identified essential references in their own publication. I would also like to extend my thanks to Elsevier BV for providing the full text data used in this research.

Finally, I really need to thank my friends Fei Liu, Ran An, and Yu Long. During these two years' time in Leiden, we have had quite a lot of interesting and useful discussions about study and, more importantly, about life. You give me a "home" in the Netherlands. I also thank my parents for their unconditional support to me. With their best love, I can go through all the challenges in my life.

ABSTRACT

Citation analysis is the quantitative study of science and technology based on publication-reference relationships. Currently, all references are assumed to make equal contribution to the citing publication, but as we all know this is not the case. To qualify this difference, the term “reference importance” is used to represent the amount of contribution that the reference makes to the citing publication. According to the previous studies, some citation features can be used to estimate the importance of references. In this thesis, the citation features that have been discussed in detail include: citation frequency, citing location, treatment, and self-citation. Based on these features, a model that can measure the importance of references was designed. This model takes the full text of scientific publications as input, and predicts the reference importance after examining citation frequency, citing location, treatment, and self-citation of each reference. The model has been validated by the author-rated importance of references which was collected through individualized web-based surveys. With the reference importance, the performance and accuracy of citation analysis can be improved. For example, it can be used to better analyze the structure and development of scientific fields, and to develop new citation impact indicators that more accurately evaluate scientific performance. In this thesis, we use the reference importance to reduce the size of citation networks. We expect that the reduced citation networks will contain less noise than the original one.

III

CONTENTS

ACKNOWLEDGEMENTS	I
ABSTRACT	II
CONTENTS	III
Chapter 1 INTRODUCTION.....	1
1.1 Research background.....	1
1.2 Research questions	2
1.3 Research contribution	2
1.4 Thesis outline	3
Chapter 2 SCIENTOMETRICS AND CITATION ANALYSIS	5
2.1 Scientometrics	5
2.2 Citation analysis	5
Chapter 3 INDICATORS OF REFERENCE IMPORTANCE	9
3.1 Importance of the reference	9
3.2 Frequency.....	9
3.3 Location	10
3.4 Treatment level.....	12
3.5 Self-citation.....	13
Chapter 4 A MULTIFACTOR MODEL FOR MEASURING THE IMPORTANCE OF REFERENCES	15
4.1 Overview of the model	15
4.2 Frequency score	16
4.3 Location score	17
4.4 Treatment score.....	18
4.5 Self-citation score.....	19

IV

4.6	Reference importance	20
Chapter 5 CALCULATION OF REFERENCE IMPORTANCE.....		23
5.1	Data extraction and storage.....	23
5.2	Datasets.....	27
5.3	Section classification method	29
5.4	Importance of references in the JOI dataset.....	33
Chapter 6 METHOD VALIDATION: AUTHOR-RATED IMPORTANCE OF CITED REFERENCES		35
6.1	Methodology	35
6.2	A web-based survey	36
6.2	Validation based on survey results.....	39
6.3	Optimization of the model using author-rated importance of references	41
Chapter 7 APPLICATION IN CITATION NETWORKS		47
7.1	Citation networks	47
7.2	Construction of reduced citation networks	48
7.3	Quantitative analysis of the reduced citation networks	50
Chapter 8 SUMMARY AND FUTURE RESEARCH		55
8.1	Summary of the thesis	55
8.2	Limitations and future research.....	56
References.....		59

Chapter 1

INTRODUCTION

1.1 Research background

Citation analysis is using a series of indicators to measure the output and impact of research entities and to analyze the relationship between for example scientific publications, journals, or researchers. Citation count, which is calculated by counting how many times a particular publication is cited by other publications (Yan, Tang, Liu, Shan, & Li, 2011), is one of the most basic measures used in citation analysis. Citation count can not only be used directly as an indicator of citation impact, but it is also the basis of other more complex measures, such as the h-index (Hirsch, 2005), the mean normalized citation score, and the percentage of highly cited publications (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). The overall quality and accuracy of citation analysis is therefore strongly dependent on the quality of the citation count measure.

The traditional citation count measure assumes that all references in a publication are equally important. However, as we all know, the contribution or importance of references in a publication may strongly vary. Therefore, it can be argued that references with a higher contribution level or references that are more important for a publication should get more credits in the calculation of the citation count measure. Therefore, one possible improvement is to measure the importance of references and then differentiate the references according to this value. From the literature it is known that the importance of references can be estimated from certain citation features, such as the citing location within the publication, the age of the cited reference, and the number of times a reference is cited within the publication (Voos & Dagaev, 1976). Based on this idea, some improved citation count methods have been introduced, but most of them only use single citation features to estimate the reference importance. However, to get a more accurate measurement, multiple features should be utilized. Although most of these features are not contained by the traditional bibliographic databases (e.g., Thomson Reuters' Web of Science and Elsevier's Scopus), they can be extracted from the full text of publications. Since academic publishers (e.g., Elsevier) are

more and more willing to make the full text of publications available in a structured and computer readable format (e.g., XML), it is possible to automatically identify these citation features using the computer. Therefore, the aim of this research is to design a methodology which automatically measures the importance of references based on information extracted from the full text of publications and then use it to improve the performance and accuracy of citation analysis.

1.2 Research questions

The main research question of this thesis is:

- MQ: How to measure the importance of references based on information extracted from the full text of scientific publications?

In order to answer the main research question, the following six sub questions will be investigated:

- RQ1: What is citation analysis and what is exactly its role in the field of scientometrics?
- RQ2: Which citation features can be used to identify the importance of references?
- RQ3: How to measure the importance of a reference based on multiple citation features extracted from the full text of a publication?
- RQ4: How to extract and store required citation features from the full text of publications?
- RQ5: How to evaluate the predicted importance of the cited reference?
- RQ6: How to reduce the noise in citation networks by using the reference importance model?

1.3 Research contribution

By answering the main research question, a model that can be used to estimate the importance of references will be introduced. The importance of references can for example be used to develop new citation impact indicators that more accurately evaluate scientific performance. In the calculation of citation impact, it is then possible to give more weight to important references and less weight to unimportant references. The importance of references can for

example also be used to better analyze the structure and development of scientific fields. To focus on the most important reference-publication relationships only, could help to identify more detailed subtopics within a field and how they are related to each other.

Compared with other attempts to measure the importance of reference, our methodology has the following distinguishing features:

- 1) Instead of a single feature (Ding, Liu, Guo, & Cronin, 2013; Hou, Li, & Niu, 2011), multiple citation features will be examined to estimate the importance of references. Specifically, four citation features will be included in this model: citation frequency, citing location, treatment level, and self-citation.
- 2) Because we will use the full text of publications as input material, the whole analysis process is more simplified and highly automated. During the earliest studies the citation features were extracted manually from the text (Bonzi, 1982; Herlach, 1978; Voos & Dagaev, 1976). Later some researchers processed the PDF version of publications to identify the target information (Zhu, Turney, Lemire, & Vellino, in press).

Our research will automatically extract information from the full text of publications, so compared with previous research our approach is easier and the extracted information will be more accurate.

- 3) Unlike most previous studies which only provide general qualitative results (such as “multiple mentioned references are more important than the references only mentioned once”), our model will quantify the level of importance. So it is more feasible to be applied in other citation analysis measures.

1.4 Thesis outline

This thesis consists of eight chapters. Chapters 2 to 7 roughly correspond to the six sub research questions proposed in Section 1.2. Table 1.1 briefly shows the connections between these research questions and the chapters. Chapter 2 is a literature review about scientometrics and citation analysis. This review provides a background for the limitations of current citation analysis and then leads to the necessity of our work. Chapter 3 describes the citation features which can be used to indicate the importance of cited references. Based on the indicators we have selected in Chapter 3, Chapter 4 introduces a multifactor model for measuring the

importance of the references. Chapter 5 applies this model on two datasets. One dataset contains publications from the *Journal of Informetrics* and another dataset contains publications in the field of renewable energy. Chapter 6 performs a validation for this reference importance measuring model. This validation is based on the author-rated importance of the references which is the result of an online survey. Chapter 7 presents an application. In this application, the importance of cited references is used to improve the structure of citation networks. Finally, Chapter 8 summarizes this thesis and proposes some directions for future research.

Table 1.1: The six sub research questions and their corresponding chapters in this thesis.

Research Question	Corresponding Chapter
RQ1: What is citation analysis and what is exactly its role in the field of scientometrics?	Chapter 2 Scientometrics and Citation Analysis
RQ2: Which citation features can be used to identify the importance of references?	Chapter 3 Indicators of Reference Importance
RQ3: How to measure the importance of a reference based on multiple citation features extracted from the full text of a publication?	Chapter 4 A Multifactor Model for Measuring the Importance of References
RQ4: How to extract and store required citation features from the full text of publications?	Chapter 5 Calculation of Reference Importance
RQ5: How to evaluate the predicted importance of the cited reference?	Chapter 6 Method Validation: Author-rated Importance of Cited References
RQ6: How to reduce the noise in citation networks by using the reference importance model?	Chapter 7 Application in Citation Networks

Chapter 2

SCIENTOMETRICS AND CITATION ANALYSIS

2.1 Scientometrics

In 1969, Nalimov and Mulchenko (1969) coined the term “scientometrics”. Now, after nearly 45 years of development, this term has already gained a wide recognition within the academic world. As it is implied by the name, scientometrics is mainly used to describe the “quantitative study of science and technology”. Tague-Sutcliffe (1992) provided a definition of scientometrics:

“Scientometrics is the study of the quantitative aspects of science as a discipline or economic activity. It is part of the sociology of science and has application to science policy-making. It involves quantitative studies of scientific activities, including, among others, publication, and so overlaps bibliometrics to some extent.”

To study the quantitative aspects of science, the scientific publications are important data sources. Citation analysis is the method that quantitatively studies the science and technology by using the information of publications. So citation analysis is a subfield of scientometrics.

2.2 Citation analysis

A scientific publication does not stand alone, but it is embedded in the network of all literatures through citation-reference relationships with other publications. According to Egghe and Rousseau (1990), the existence of a cited document in a reference list indicates the facts that there is a relationship between the cited and citing documents from the author’s point of view. Citation analysis is an area in the field of scientometrics that deals with the study of these relationships. By analyzing these relationships, it provides us a way to evaluate the academic or scientific performance from a quantitative perspective.

Before discussing citation analysis into more detail, it is necessary to distinguish between the two most frequently used notions: “reference” and “citation”. According to Ding et al. (2013), the term “reference” refers to a publication that is listed in the reference section of a citing

publication. A reference may be mentioned several times in a publication, and each occurrence is considered a citation. Although for the difference between these two notions, other researchers may hold different opinions, but within this research we will follow the rules given by Ding et al. (2013).

According to Zunde (1971), the applications of citation analysis can be classified into following three areas:

- 1) Qualitative and quantitative evaluation of scientists, publications and scientific institutions;
- 2) Modeling of the historical development of science and technology;
- 3) Information search and retrieval.

To better interpret and use the results of citation analysis, it is necessary to understand the nature of citation relations. However, this relationship is somewhat difficult to characterize as there are several reasons for citing a particular publication. For example, Garfield (1965) has identified the following fifteen reasons:

- 1) Paying homage to pioneers;
- 2) Giving credit for related work;
- 3) Identifying methodology, equipment, etc.;
- 4) Providing background reading;
- 5) Correcting one's own work;
- 6) Correcting the work of others;
- 7) Criticizing previous work;
- 8) Substantiating claims;
- 9) Alerting to forthcoming work;
- 10) Providing leads to poorly disseminated, poorly indexed, or uncited work;
- 11) Authenticating data and classes of fact – physical constants, etc.;
- 12) Identifying original publications in which an idea or concept was discussed;
- 13) Identifying original publications or other work describing an eponymic concept or term [...];
- 14) Disclaiming work or ideas of others;
- 15) Disputing priority claims of others.

As different references may be cited because of different reasons, the strength of the citation-reference relationship will also be varied. However within most of the current citation analysis methods (e.g., counting of citations, journal impact factor, and h-index) the references are only counted based on the reference list appearing at the end of the publication, so the strength or direction of the influence is not specified (Ding et al., 2013). All the references are assumed to make equal contributions to the citing publication, but as we all know in reality this is not the case. To account for this problem, the earliest work was done by Pinski and Narin (1976), who proposed to refine the citation analysis by taking into account the length of papers, the prestige of the citing journal, and the different referencing characteristics of different segments of the literature. Later more research has been done to investigate which citation features may indicate the contribution level of references and how to measure this influence. In general, this research was conducted at two main levels: the syntactic level and the semantic level.

On the syntactic level, the citations are differentiated according to the structural features of publications. The first feature is frequency, which represents how many times a reference is mentioned in the text of a publication. Both Virgo (1977) and Herlach (1978) have found a significant positive relationship between frequency and the importance of references. The second feature is citing location. The structure of academic papers is somewhat standardized, and typically it follows a structure like: introduction, materials and methods, results, discussion, and conclusions (Marshall, 2005). As we all know, different sections play different roles within a research paper. Therefore, citations that are mentioned in specific sections may also correspond to certain functions. Thirdly, treatment level, that is the amount of detail a reference is discussed in the text, may also influence the importance of a reference. Bonzi (1982) classified the reference into four treatment categories and Swales (1990) made a more straightforward framework with two categories: 1) Integral citation: the name of the researcher occurs in the actual citing sentence as some sentence-element; 2) Non-integral citation: the name of the researcher occurs either in parenthesis or is referred to elsewhere by a superscript number or via some other devices. Finally, whether one reference is self-citation or not may also influence its importance, because authors always rate self-citation references relatively more important (Tang & Safer, 2008).

On the semantic level, citations are analyzed based on the nature of the contributions they make to the citing publication by using text-mining techniques. At first, research on the semantic level of citations was limited to interviews and manual processing. Garfield (1974)

regarded the cited publications as subject headings of the citing publication. Based on this idea, Small (1978) analyzed the context of citations in the publications of chemistry, and has found that there were some standard functions and meanings. More recently, driven by the wide use of computer technology and the increasing availability of full text publications, the supervised machine-learning has become more popular. With the help of this technique, researchers such as Teufel, Siddharthan, and Tidhar (2006) were able to classify references according to their function in the citing publication and finally proposed a citation function annotation schema.

Based on these findings, some improvements of the traditional citation analysis method were proposed. For example, both Hou et al. (2011) and Ding et al. (2013) suggested to count how many times each reference has been mentioned in the full text instead of how many times it is listed in the reference list. To avoid the influence of self-citations, it is always possible to exclude self-citations in the counting process.

Although some improved citation analysis methods were introduced that include the importance of reference, most of them only use a single citation feature (e.g., citing frequency, self-citation) to measure the importance of references. Therefore, in this research, we plan to estimate the importance of references based on multiple citation features and finally use them together to improve the traditional citation analysis method.

Chapter 3

INDICATORS OF REFERENCE IMPORTANCE

3.1 Importance of the reference

As it has been discussed in the previous chapter, not all references are equally important to their citing publications. To qualify this difference, within this research the term “importance of reference” is employed to represent the amount of contribution that the reference makes to the publication. References that are more influential or inspirational for the core idea of a citing publication can be considered as more important than others. From the literature, a variety of properties of a reference-publication pair can be used to estimate the importance of a reference, such as the citing frequency, citing location within the publication, function of the reference, or self-citation (Ding et al., 2013; Hou et al., 2011; Tang & Safer, 2008; Zhu et al., in press). Here we call these properties the indicators of the reference importance. Our goal is to create a model that can quantify the reference importance based on a set of these indicators. However, before we step into this model, each indicator and its relationship with the importance of references will be elaborated in detail within this chapter.

3.2 Frequency

The frequency of a reference is the number of times this reference is cited within its citing publication. Compared with references that are only cited once within a given publication, references that are cited multiple times are more likely to have a close relationship with the citing publication.

Regarding to the pattern of reference frequency, Lievers and Pilkey (2012) have examined 104,561 references from 3,150 publications in three research areas: economics, computing, and medicine & biology. They found that 3.8% of the references are cited five or more times, 0.48% of the references are cited 10 or more times, and only 0.05% of the reference are cited 20 or more times. Beside of this, Lievers and Pilkey (2012) have also found that this pattern of repeated citations is consistent across the sampled journals and research disciplines.

The idea that uses frequency to assess the importance or influence of a reference is not new. Voos and Dagaev (1976) analyzed 1170 citations of four publications which are published in 1970 and found out that it is possible to measure the value of a reference using a function of frequency. They proposed the following hypothesis:

An author who is cited more than once in an article might have more relevance and/or importance than an author who is cited only once in an article.

This hypothesis has been tested by both Virgo (1977) and Herlach (1978), and they all found a significant positive relationship between the reference frequency and the reference importance. Hou et al. (2011) and Ding et al. (2013) proposed to count how many times a reference is cited in the text of the publication, instead of how many times it is mentioned in the reference list to improve the accuracy of assessing scientific contribution. By comparing these two counting results, they found that citation frequency of individual articles in other publications more fairly measures their scientific contributions than mere presence in reference lists. Tang and Safer (2008) and Zhu et al. (in press) systematically analyzed the quantitative relationship between several citation features and author-rated importance of each reference. One of their main results is that the frequency of a reference is one of the best predictors of how influential a reference is. In addition, Tang and Safer (2008) also indicated that this relationship is stronger in publications where the mean level of reference frequency is low.

Based on these findings, we can conclude that the value of a reference can be predicted by its frequency and the mean level of reference frequency of its citing publication. More specifically, the reference importance is positively correlated with the reference frequency, but negatively correlated with the mean level of reference frequency of its citing publication.

3.3 Location

The location of a citation indicates where the reference has been cited in the citing publication. Since a reference can be cited several times in a publication, this reference can have multiple locations and each of them corresponds to a citation of this reference.

According to Swales (1990), in earlier years references were only concentrated in the Introduction section, but nowadays they are distributed throughout the whole research paper. The structure of an academic publication is somewhat standardized, and typically it follows a

structure like: introduction, materials and methods, results, discussion, and conclusions (Marshall, 2005). As we all know, different sections play different roles within a publication. Therefore, citations that are mentioned in specific sections may also correspond to certain functions. Therefore it will be quite reasonable to expect that references which have relatively more important functions (such as providing a conceptual idea that is specifically relevant to the citing publication) may be more important than the references that only have less significant functions (such as providing general background of the research topic). Therefore, it becomes possible to analyze a citation's perceived level of importance based on its location.

However before we step into the detailed relationship of the importance of references and the citation location, it is necessary to make clear what the structure of a scientific publication is. Since its origin in 17th century, the layout of scientific publications has changed quite a lot. Nowadays the structure is fairly standardized. It follows a sequence like: introduction, theoretical background, experimental/observational techniques, samples, data analysis, results/observations, discussion, and summary/conclusions (Ding et al., 2013). However, within a publication not all the sections will be listed, and some of them are always combined together, such as introduction and background. Therefore, a simplified structure, IMRAD (Introduction, Methods, Results, and Discussion), may be more widely adopted by today's research publications. Sollaci and Pereira (2004) measured the number of publications written under the IMRAD structure from 1935 to 1985 in four leading internal medicine journals, and they found that from 1985 this structure has become the only pattern adopted in the selected sample of publications. More recently, Hu, Chen, and Liu (2013) analyzed 350 papers published in *Journal of Informetrics* from 2007 to 2013 and found most of them are organized in four to six sections (74.3%). More specifically, 26% have four sections, 28.6% have five sections, and 19.4% have six sections. The four-section publications are always made up of: introduction, method/data, results, and conclusions/discussion. They also indicate that the five-section and six-section structures can be considered as an elaboration of the original four-section structure.

Voos and Dagaev (1976) first noticed the relationship between reference importance and the citing locations. They analyzed the citation contribution based on its location and concluded that the importance of a reference should be based on both its frequency and its location within the citing publication. Later Herlach (1978) found that a reference cited in the introduction or literature review section and later again in the methodology or discussion section should be regarded as having a greater contribution to the citing publication. Maričić,

Spaventi, Pavičić, and Pifat-Mrzljak (1998) conducted an analysis for 357 scientific publications published between 1955 and 1964. Their result showed that citations in the method, result, and discussion sections are more meaningful than the citations in the introduction section. Similarly, Tang and Safer (2008) analyzed the correlation between citation location and the author-rated importance of the references. They found that the references cited in the method section were rated as more important by the citing author than the references cited in the other sections. References that are only cited in the introduction section were considered less useful by the authors.

3.4 Treatment level

Citation treatment indicates how citations are mentioned in the citing publications. Bonzi (1982) indicated in her research that the extent of treatment of the cited reference in the citing publication can be used as a measure of reference importance. This is based on the hypothesis that references that are discussed in more detail are more likely to have a closer relationship with the citing publication than references that are discussed in less detail. After analyzing nearly 500 references, she classified the treatment of reference into following four levels:

- 1) Not specifically mentioned in text (e.g., “Several studies have dealt with...”);
- 2) Barely mentioned in text (e.g., “Smith has studied the impact of ...”);
- 3) One quotation or discussion of one point in text (e.g., “Smith found that ...”);
- 4) Two or more quotations or points discussed in text.

Similar with Bonzi, Dubois (1988) examined the biomedical journal articles and classified the extent of citation treatment into four categories:

- 1) Direct quotation;
- 2) Paraphrase;
- 3) Summary;
- 4) Generalization.

Swales (1990) has made a more straightforward classification:

- 1) Integral citation: in which the name of the researcher occurs in the actual citing sentence as some sentence-element;
- 2) Non-integral citation: where the name of the researcher occurs either in parenthesis or is referred to elsewhere by a superscript number or via some other device.

Swales’ model can be interpreted as a simplified version of Bonzi’s model, which means that non-integral citation is equivalent to Bonzi’s category “not specifically mentioned” and

integral citation is for the remaining three categories “barely mentioned in text”, “one quotation or discussion of one point in text”, and “two or more quotations or points discussed in text”.

Based on Bonzi’s classification, Tang and Safer (2008) quantitatively investigated the correlation between the citation treatment level and citation importance. They found that there is a significant positive association between these two factors, which means the more deeply a reference is discussed in the citing publication, the more important it will be.

3.5 Self-citation

Self-citations, which is defined as a citation in which the citing and cited paper have at least one author in common, account for a significant proportion of all citations (Aksnes, 2003). According to Schreiber (2007), in general there are three reasons for self-citations:

- a. Self-citations are really needed in the manuscript in order to avoid repetition of previously described experimental setups, theoretical models, as well as results and conclusions [...];*
- b. An author knows his own previous manuscripts best and therefore it is easier to refer to these own papers when a citation is required in a given context for a certain argument;*
- c. Due to the ever-increasing number of evaluations which are based on citation counts, it is of course tempting to enhance one’s citation count by referring to the own papers for this very purpose.*

The first two reasons of self-citations are legitimate, but the third kind of self-citations may lead to a lot of criticism. For the third kind self-citations, no matter how frequently they are cited in the publication, which section they are cited and how detail they are discussed, they always make very small contribution to the citing publication. So the patterns we have found for other three features (frequency, location, and treatment) are not suitable for this kind of self-citations. If all three kinds of self-citations are used to identify the importance of references, it is reasonable to suspect that the third kind of self-citations may introduce some noise into the analysis. Since it is quite difficult to identify whether self-citation belongs to the third category, many scholars have suggested that self-citations should be removed from citation counts in citation analysis, at least at micro and meso levels (Aksnes, 2003; Fowler & Aksnes, 2007). Given the different application areas of citation analysis, Schreiber (2007)

suggested to include the self-citations when identifying hot fields of research, but exclude them when assessing the scientific achievement of an individual scientist.

Based on these findings, we can conclude that some self-citations (first and second type of self-citations) are really essential to the citing publication, but the others (third type of self-citations) are unimportant. Since it is difficult to distinguish these two groups of self-citations, it is probably best to give a small penalty to self-citations.

Chapter 4

A MULTIFACTOR MODEL FOR MEASURING THE IMPORTANCE OF REFERENCES

4.1 Overview of the model

Within the previous chapter, the indicators of reference importance were discussed in detail. These indicators are frequency, location, treatment level, and self-citation. In this chapter, our aim is to construct a suitable model that can predict the importance of references using these indicators.

In general, this model takes the full text of publications as input, and by calculating the indicator level scores (location score, frequency score, treatment score, and self-citation score) it finally generates the importance of references as output. Figure 4.1 is an overview of this model.

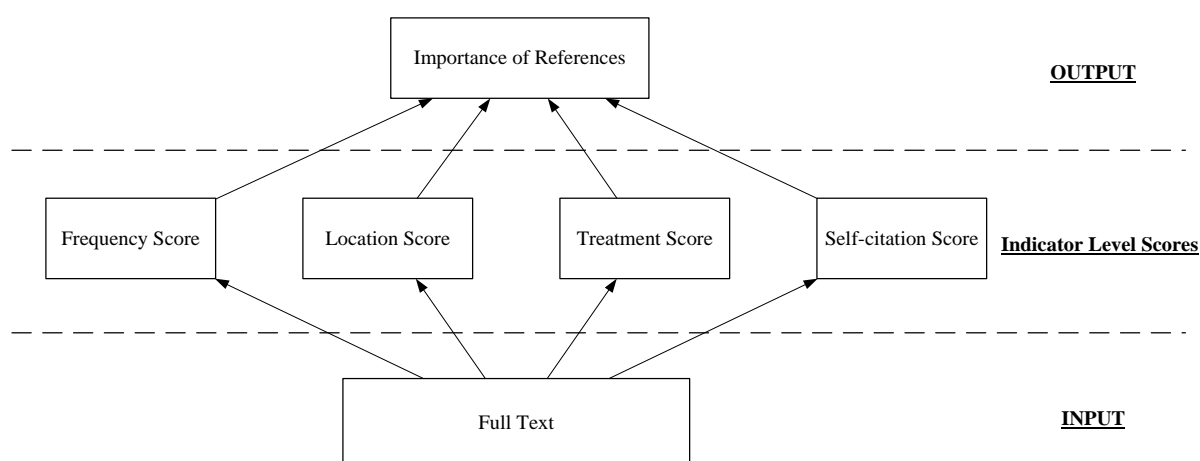
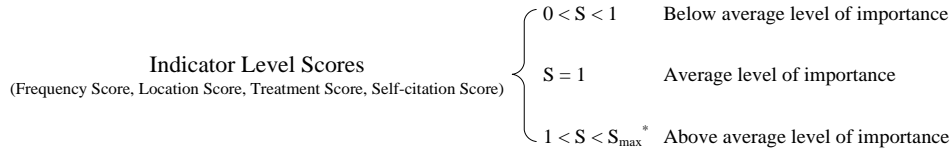


Figure 4.1: Structure of the reference importance model

The input data, citation features and other related properties, will be extracted directly from the full text of the publications. In Chapter 5, this extraction process will be discussed in detail.



*Maximum value of scores. Different scores have different maximum value, and they will be described in the following sections of this chapter. In general, S_{\max} is around 2.

Figure 4.2: Description of indicator level score

During the data processing process, four indicator level scores are calculated, and the greater a score is, the more important this reference will be (assessed by this indicator). The score is always positive, and 1 represents the average level of importance. This relationship is explained in Figure 4.2.

4.2 Frequency score

As has been discussed in Section 3.2, frequency of a reference is a good predictor of reference importance. The more frequently a reference is cited in the publication, the more influential this reference may be. The higher the average reference frequency of all the references in the given publication, the less essential this reference will be. So the reference importance is positively correlated with its frequency (F), but negatively correlated with the average frequency of all the references in the given publication (Af). It is quite reasonable to give the reference an average level of frequency score (1.00) if its frequency is equal to the average frequency of all the references in its citing publication. Therefore the frequency score can be calculated as:

$$\begin{cases} S_f = f(F, Af) = (1 - e^{-k \cdot \frac{F}{Af}}) \cdot (fh - fl) + fl \\ k = -\log\left(\frac{fh - 1}{fh - fl}\right) \end{cases} \quad (\text{Eq. 4.1})$$

where S_f is the frequency score, fh is the maximum value of frequency score and fl is its minimum value. Figure 4.3 is the plot of S_f . S_f has the following properties:

- 1) For the references whose citing publications have the same average frequency level, the more frequently the reference is mentioned, the higher its frequency score will be.
- 2) For the references that have the same frequency, the reference cited in a publication with a higher average frequency level will get a higher frequency score.

- 3) If the citing frequency of a reference in a publication equals the average citing frequency of all references, then its frequency score is 1.

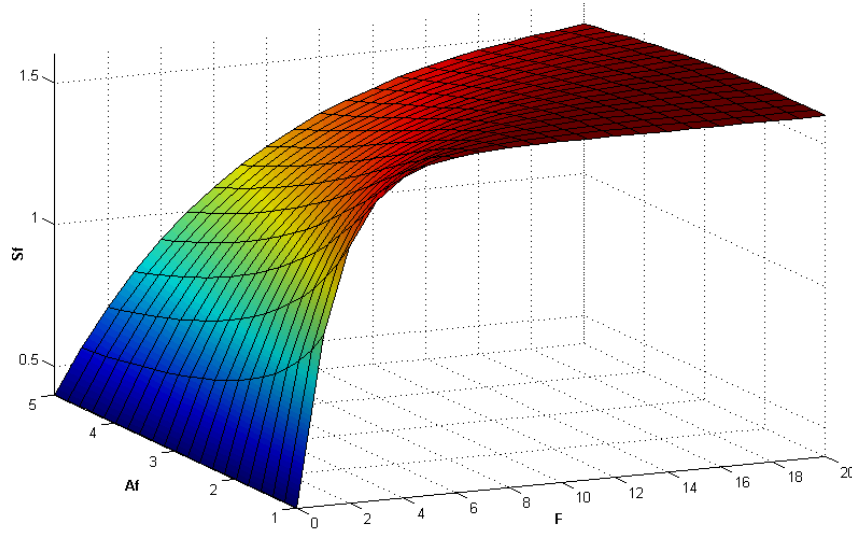


Figure 4.3: Plot of Eq.4.1: $S_f = f(F, Af) = (1 - e^{-k \frac{F}{Af}}) \cdot (fh - fl) + f$, $k = -\log(\frac{fh-1}{fh-fl})$, where $fh = 1.60$, $fl = 0.40$

4.3 Location score

According to Section 3.3, the citing location of a reference may be predictive of how influential this reference is. The location score is designed to qualify this level of influence. Based on their citing location, references are classified into following five types:

- 1) Introduction Only: references only cited in the introduction section
- 2) Method: references cited in the method section
- 3) Footnote Only: references only cited in the footnote
- 4) Appendices Only: references only cited in the appendices
- 5) Others: references that are not classified into above four types

The locations and their corresponding location scores are shown in Table 4.1. These scores are intuitively chosen based on the findings of Section 3.3. In general, *Introduction Only*, *Footnote*, and *Appendices* references are less influential than the others. Their location score is a fixed number that can be assigned by the analyst and this number is between 0 and 1. However, compared with the references cited in the appendices, references in the footnote always have more strong connection with the publication. Therefore, we decided to give more

credit to the *Footnote Only* references (0.50) compared with *Appendices Only* references (0.10). As is described before, references are cited because of different reasons, and instead of essential functions (definition, tool, starting point) references in the introduction section are more likely to be used for general purposes (background, avoid plagiarism). So it is reasonable to give this kind of references a slightly below average location score (0.90).

Table 4.1: Calculation of location score

Location	Location Score (S_l)*
Introduction Only	0.90
Method	1.50
Footnote Only	0.50
Appendices Only	0.10
Others	1.00

* Fixed value that can be assigned by the analysts. Here is the value we used in this research.

According to the literature, the *Method* references always play an essential role in the citing publication, so they are more likely to make a greater contribution to the publication. Taken this into consideration, a location score (1.50) that is greater than 1 is assigned to them.

References that are not classified into *Introduction Only*, *Method*, *Footnote Only* or *Appendices Only* will be put into *Others*. For these references, no specific corresponding relationship between the location and their importance to the citing publication has been found, so the value that represents the average level of importance (1.00) is used.

4.4 Treatment score

As it has been mentioned in Section 3.4, Swales (1990) divided the citations into two groups:

- 1) Integral citation: author name of the reference is mentioned in the citing sentence;
- 2) Non-integral citation: author name of the reference is not mentioned in the citing sentence.

In this research, we will follow the same classification method. Reference may be cited several times in a publication. If the author name is mentioned in any of the citing sentences,

then the corresponding reference will be considered as an integral reference. But if none of the citing sentences include the author name, then the corresponding reference is considered as a non-integral reference.

According to Section 3.4, the relationship between reference treatment and the importance of the reference is: the more deeply a reference is discussed in the citing publication, the more important it will be. However, it is also reasonable to suppose that an integral reference ($T = 1$) is more influential in a publication where there are more non-integral references ($T = 0$), and vice versa. So besides the reference treatment level (T), the average treatment level of all the references in the given publication (At) is also used to predict the reference importance. Therefore we suggest to calculate the treatment score (S_t) as follows:

$$S_t = f(T, At) = \begin{cases} f(At|T=0) = 1 - (1-tl) \cdot At \\ f(At|T=1) = th - (th-1) \cdot At \end{cases} \quad (\text{Eq. 4.2})$$

where tl is the minimum value of treatment score and th is the maximum value of it.

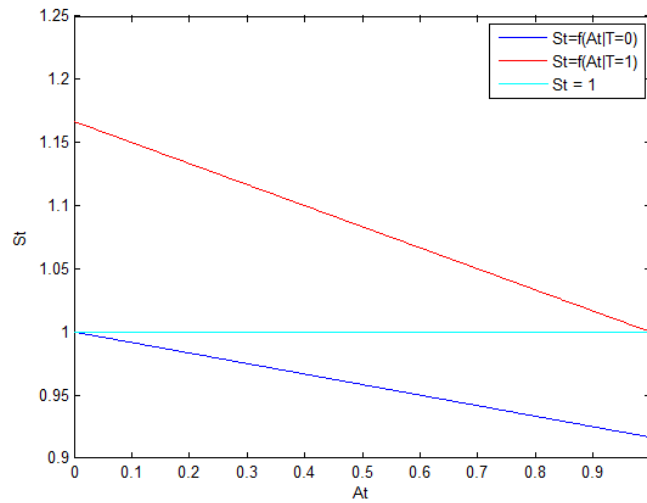


Figure 4.4: Plot of Eq. 4.2: $S_t = f(T, At) = \begin{cases} f(At|T=0) = 1 - (1-tl) \cdot At \\ f(At|T=1) = th - (th-1) \cdot At \end{cases}$, where $tl = \frac{11}{12}$, $th = \frac{7}{6}$

4.5 Self-citation score

Based on Section 3.5, to measure the reference importance, the self-citations need to be identified. Strictly speaking, the self-citation score doesn't represent the importance of

references, but it is used to identify whether a reference is self-citation or not. The rule of self-citation score is quite straightforward:

- 1) $S_s = 1$, self-citation;
- 2) $S_s = 0$, not self-citation.

4.6 Reference importance

After we have retrieved the four indicator-level scores (location score, frequency score, treatment score, and self-citation score), the importance of a reference (V) can be calculated as:

$$V = S_f \cdot p_f + S_l \cdot p_l + S_t \cdot p_t + S_s \cdot p_s + C \quad (\text{Eq. 4.3})$$

Here p_f, p_l, p_t, p_s are weights for frequency score (S_f), location score (S_l), treatment score (S_t), and self-citation score (S_s). They represent the percent of contributions each score made to the final importance of the reference. C is a constant that is used to make sure that the average reference importance of all references is around 1. The analyst can adjust these weights according to the characteristics of his research. For instance, if he thinks that self-citations have little influence in his dataset, he can give p_s a very small value or even remove this factor from the model by setting $p_s = 0$. As we all know, the patterns of reference value may slightly differ between disciplines, so by adjusting these weights this model can be tuned to different research requirements.

Previous research has shown that compared with the other citation features, citing frequency is the best predictor of the reference importance and self-citation has relatively limited impact to the importance of references (Tang & Safer, 2008; Zhu et al., in press). The performance of location and treatment are in between frequency and self-citation. According to the relative importance of these four features, we choose the weights in Table 4.2 for the scores. The constant C is used to make sure that the average reference importance of all references is closed to 1.00 (which represents the average level of importance).

Table 4.2: Weights for the indicator-level scores

Weight	Value
p_f : (frequency score weight)	0.70
p_l : (location score weight)	0.25
p_t : (treatment score weight)	0.25
p_s : (self-citation score weight)	0.05
C : (constant)	-0.20

Chapter 5

CALCULATION OF REFERENCE IMPORTANCE

5.1 Data extraction and storage

To calculate the importance of references using the model described in Chapter 4, certain citation features (e.g., citation frequency, location, citing sentence, etc.) need to be identified. Information on these features is not available in traditional bibliographic databases (e.g., Thomson Reuters' Web of Science and Elsevier's Scopus) which contain metadata about scientific publications and their cited references. Of course these features can be extracted from the full content of publications. Recently academic publishers are more and more willing to make the full text of publications available in a highly structured format. The text and data mining (TDM) tool of Elsevier can be used to retrieve the full text of publications that are published by Elsevier. In this research, we will use the online interface (API) of this TDM tool to batch-download the full text of publications in a computer-readable XML format.

The full text contains, for instance, publication metadata, reference information, citation information, publication structure, publication content, etc. All these data are clearly marked with XML tags (e.g., <dc:title>...</dc:title>) and corresponding IDs, so they can be easily matched with each other. Figure 5.1 gives an example about how reference information is linked with the citation data.

A custom program, written in VB.NET, was developed to download the XML files of publications published by Elsevier, to process these XML files, and to store the extracted data in a Microsoft SQL Server database. The structure of the database that is used to store the extracted data from the XML files is shown in Figure 5.2.

Based on our empirical results and in line with our earlier work (Waltman & Van Eck, 2013), we advise against using source normalization approaches that follow the fractional citation counting idea of Leydesdorff and Opthof (2010). The

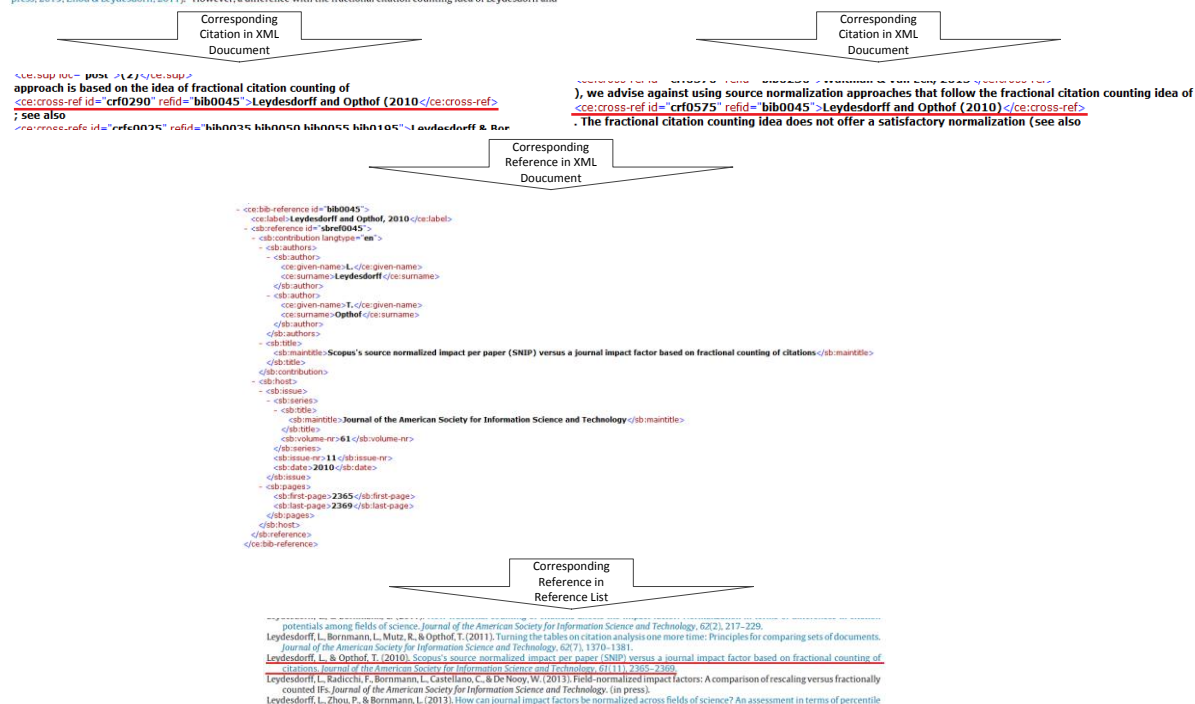


Figure 5.1: Reference and citation information in the full text of a publication
(Extracted from Waltman, L., & van Eck, N. J. (2013). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, 7(4), 833-849. <http://dx.doi.org/10.1016/j.joi.2013.08.002>)

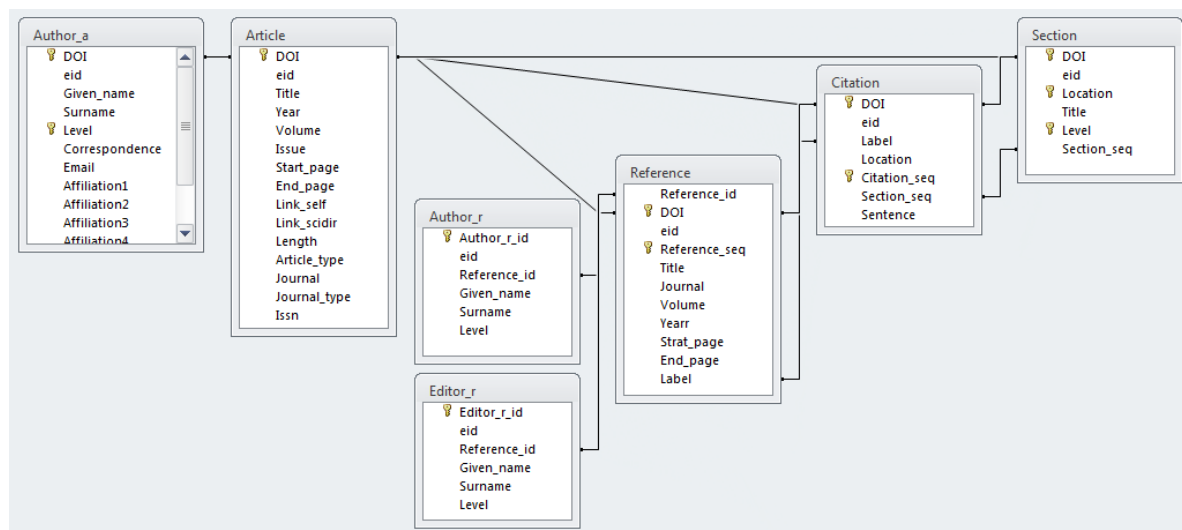


Figure 5.2: Structure of the database that stores the information extracted from the full texts

Each record in the *Article* table represents a publication and most of the metadata (such as *DOI*, *title*, *publication year*, *journal*, etc.) related to the source publication is stored in this table. Figure 5.3 shows some example rows from the *Article* table. In this research, *DOIs* are used to uniquely identify publications. The author information of publications is stored in the *Author_a* table. The *level* field in this table represents the order of the author in the author list.

By using the *DOI*, authors can be linked with the corresponding publication in the *Article* table. Since publications can have several authors, multiple records in *Author_a* table can link to the same publication. The structure of *Author_a* table is shown in Figure 5.4.

DOI	eid	title_m	year	volume	start_page	end_page	link_self	link_scidr	length	article_type	journal	journal_type	issn
10.1016/j.joi.2006.05.001	1+2.0-S1751157706000022	The influence of missing publications o...	2007-01-31	1	1	7	http://api.elsevier.com/cont...	http://www.sciencedir...	9520	Journal of Informetrics	Journal	17511577	
10.1016/j.joi.2006.06.001	1+2.0-S1751157706000034	Finding scientific gems with Google's P...	2007-01-31	1	1	15	http://api.elsevier.com/cont...	http://www.sciencedir...	18989	Journal of Informetrics	Journal	17511577	
10.1016/j.joi.2006.07.001	1+2.0-S1751157706000046	Hirsch's h-index: A stochastic model	2007-01-31	1	1	25	http://api.elsevier.com/cont...	http://www.sciencedir...	17653	Journal of Informetrics	Journal	17511577	
10.1016/j.joi.2006.08.001	1+2.0-S1751157706000058	Some measures for comparing citation ...	2007-01-31	1	1	34	http://api.elsevier.com/cont...	http://www.sciencedir...	19597	Journal of Informetrics	Journal	17511577	
10.1016/j.joi.2006.09.001	1+2.0-S175115770600006X	Measuring quality of similarity functions ...	2007-01-31	1	1	46	http://api.elsevier.com/cont...	http://www.sciencedir...	24441	Journal of Informetrics	Journal	17511577	
10.1016/j.joi.2006.09.002	1+2.0-S1751157706000071	Journal self-citations-Analyzing the JIF ...	2007-01-31	1	1	58	http://api.elsevier.com/cont...	http://www.sciencedir...	34737	Journal of Informetrics	Journal	17511577	

Figure 5.3: *Article* table

DOI	eid	given_name	surname	level	correspondence	email	affiliation1	affiliation2
10.1016/j.joi.2006.05.001	1+2.0-S1751157706000022	Ronald	Rousseau	1		ronald.rousseau@khbo.be	KHBO (Association K.U.Leuven), Department of ...	University of Antwerp (UA), IBW, I
10.1016/j.joi.2006.06.001	1+2.0-S1751157706000034	P.	Chen	1	Corresponding author.	patrick@bu.edu	Center for Polymer Studies and Department of P...	
10.1016/j.joi.2006.06.001	1+2.0-S1751157706000034	H.	Xie	2		hxie@bni.gov	New Media Lab, The Graduate Center, CUNY, ...	Department of Condensed Matter
10.1016/j.joi.2006.06.001	1+2.0-S1751157706000034	S.	Maslov	3		maslov@bni.gov	Department of Condensed Matter Physics and M...	
10.1016/j.joi.2006.06.001	1+2.0-S1751157706000034	S.	Redner	4		redner@bu.edu	Center for Polymer Studies and Department of P...	
10.1016/j.joi.2006.07.001	1+2.0-S1751157706000046	Quentin L.	Burell	1	Tel.: +44 1624693706; fax: +44 162466...	q.burell@lbs.ac.im		

Figure 5.4: *Author_a* table

In the *Section* table, the *location* of a section is measured in terms of the number of characters from the beginning of the publication to the beginning of the section. Most publications contain sections that are structured in a hierarchical way. Sections may contain subsections, and subsections may contain subsubsections. To describe this structure, *level* and section sequence (*section_seq*) fields are used in the *Section* table. Main sections are stored as level 1 sections, and subsections of the level 1 sections are stored as level 2 sections. The same principle is applied to sections of level 3, level 4, etc. For all level 1 sections, their sequence of appearance is stored in the *section_seq* field. For other level sections, their sequence information will not be used in the later analysis. So instead of the real sequence, we just assign 0 to their *section_seq* field. Figure 5.5 provides some example rows in the *Section* table.

DOI	eid	location	title	level	section_seq
10.1016/j.joi.2006.08.001	1+2.0-S1751157706000058	0	Introduction	1	1
10.1016/j.joi.2006.08.001	1+2.0-S1751157706000058	2063	The measures	1	2
10.1016/j.joi.2006.08.001	1+2.0-S1751157706000058	2416	Overlap and footnote	2	0
10.1016/j.joi.2006.08.001	1+2.0-S1751157706000058	4531	Fagin measure	2	0
10.1016/j.joi.2006.08.001	1+2.0-S1751157706000058	7263	Inverse rank measure	2	0
10.1016/j.joi.2006.08.001	1+2.0-S1751157706000058	8515	Data collection	1	3
10.1016/j.joi.2006.08.001	1+2.0-S1751157706000058	13989	Results	1	4
10.1016/j.joi.2006.08.001	1+2.0-S1751157706000058	18759	Conclusions	1	5
10.1016/j.joi.2006.05.001	1+2.0-S1751157706000022	0	Introduction	1	1
10.1016/j.joi.2006.05.001	1+2.0-S1751157706000022	1304	A simple discrete model	1	2
10.1016/j.joi.2006.05.001	1+2.0-S1751157706000022	4156	A first example: Citations follow a Zipf distrib...	1	3

Figure 5.5: *Section* table

Citation information can be extracted from the body section of the XML files and it is stored in the *Citation* table. The *location* for a citation is measured in terms of the number of characters from the beginning of the publication to the citing location. The section sequence (*section_seq*) of a citation is the sequence number of the level 1 section that contains the citation. By using the *DOI* and the *section_seq*, a citation can be located into a specific section

of a publication. To calculate the treatment level of a citation, the sentence that contains this citation is extracted and stored in the *sentence* field. See Figure 5.6 for some example rows from the *Citation* table.

	DOI	eid	label	location	citation_seq	section_seq	sentence
1	10.1016/j.joi.2006.05.001	1-s2.0-S1751157706000022	bib1	108	1	1	Recently the Hirsch index, in short: h-index, has attracted a lot of attentio
2	10.1016/j.joi.2006.05.001	1-s2.0-S1751157706000022	bib7	108	2	1	Recently the Hirsch index, in short: h-index, has attracted a lot of attentio
3	10.1016/j.joi.2006.05.001	1-s2.0-S1751157706000022	bib9	108	3	1	Recently the Hirsch index, in short: h-index, has attracted a lot of attentio
4	10.1016/j.joi.2006.05.001	1-s2.0-S1751157706000022	bib11	108	4	1	Recently the Hirsch index, in short: h-index, has attracted a lot of attentio
5	10.1016/j.joi.2006.05.001	1-s2.0-S1751157706000022	bib10	137	5	1	This index, introduced by aaaaa137bbbbHirsch is calculated as follow
6	10.1016/j.joi.2006.05.001	1-s2.0-S1751157706000022	bib2	591	6	1	Clearly, this definition can also be applied to some other source-item pain
7	10.1016/j.joi.2006.05.001	1-s2.0-S1751157706000022	bib8	593	7	1	Clearly, this definition can also be applied to some other source-item pain
8	10.1016/j.joi.2006.05.001	1-s2.0-S1751157706000022	bib13	595	8	1	Clearly, this definition can also be applied to some other source-item pain
9	10.1016/j.joi.2006.05.001	1-s2.0-S1751157706000022	bib1	858	9	1	Yet, it is also possible to collect citations from the Web via Google Schol

Figure 5.6: *Citation* table

Most of the reference metadata that is available from the reference list is stored in the *Reference* table. The *label* of a reference is a string that uniquely identifies the reference within its citing publication. The *reference_id*, which is the combination of *DOI* and *label*, uniquely identifies the reference within the entire database. See Figure 5.7 for some example rows from the *Reference* table. References can also have multiple authors or editors. So, similar with the *Author_a* table, the author and editor information of references is stored separately in an *Author_r* and *Editor_r* table. Figure 5.8 shows some example rows from these two tables. Both these tables can be linked with the *Reference* table by making use of the *reference_id* field. The *level* field in these tables represents the order of the author or editor in the author or editor list of the publication.

	DOI	reference_id	eid	reference_seq	title	journal	volume	year	start_page	end_page	label
1	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib1	1-s2.0-S1751157706000022	1	H-index for Price medalists revisited	ISSI Newsletter	2	2006	3	5	bib1
2	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib2	1-s2.0-S1751157706000022	2	A Hirsch-type index for journals	The Scientist	19	2005	8		bib2
3	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib3	1-s2.0-S1751157706000022	3	Power Laws in the Information Production Process: ...			2005			bib3
4	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib4	1-s2.0-S1751157706000022	4	How to improve the h-index	The Scientist	20	2006	14		bib4
5	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib5	1-s2.0-S1751157706000022	5	An improvement of the Hindex: the Gindex	ISSI Newsletter	2	2006	8	9	bib5
6	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib6	1-s2.0-S1751157706000022	6	Theory and practice of the g-index	Scientometrics	69	2006	131	152	bib6
7	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib7	1-s2.0-S1751157706000022	7	Egghe, L. Dynamic h-index: the Hirsch index in funct...						bib7
8	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib8	1-s2.0-S1751157706000022	8	An informetric model for the Hirsch index	Scientometrics	69	2006	121	129	bib8
9	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib9	1-s2.0-S1751157706000022	9	On the h-index—a mathematical approach to a new ...	Scientometrics	67	2006	315	321	bib9
10	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib10	1-s2.0-S1751157706000022	10	An index to quantify an individual's scientific researc...	Proceedings of the National Academy of Sciences ...	102	2005	16569	16572	bib10
11	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib11	1-s2.0-S1751157706000022	11	H-index sequence and h-index matrix: constructions ...	Scientometrics	69	2006	153	159	bib11
12	10.1016/j.joi.2006.05.001	10.1016/j.joi.2006.05.001/bib12	1-s2.0-S1751157706000022	12	Conglomerates as a general framework for informetr...	Information Processing and Management	41	2005	1360	1368	bib12

Figure 5.7: *Reference* table

	author_r_id	reference_id	eid	given_name	sumame	level
1	10.1016/j.joi.2006.05.001/bib1/1	10.1016/j.joi.2006.05.001/bib1	1-s2.0-S1751157706000022	J.	Bar-Ilan	1
2	10.1016/j.joi.2006.05.001/bib10/1	10.1016/j.joi.2006.05.001/bib10	1-s2.0-S1751157706000022	J.E.	Hirsch	1
3	10.1016/j.joi.2006.05.001/bib11/1	10.1016/j.joi.2006.05.001/bib11	1-s2.0-S1751157706000022	L.	Liang	1
4	10.1016/j.joi.2006.05.001/bib12/1	10.1016/j.joi.2006.05.001/bib12	1-s2.0-S1751157706000022	R.	Rousseau	1
5	10.1016/j.joi.2006.05.001/bib2/1	10.1016/j.joi.2006.05.001/bib2	1-s2.0-S1751157706000022	T.	Braun	1
6	10.1016/j.joi.2006.05.001/bib2/2	10.1016/j.joi.2006.05.001/bib2	1-s2.0-S1751157706000022	W.	Glänzel	2

	editor_r_id	reference_id	eid	given_name	sumame	level
1	10.1016/j.joi.2006.09.005/bib10/1	10.1016/j.joi.2006.09.005/bib10	1-s2.0-S1751157706000113	H.	Zuckerman	1
2	10.1016/j.joi.2006.09.005/bib10/2	10.1016/j.joi.2006.09.005/bib10	1-s2.0-S1751157706000113	J.R.	Cole	2
3	10.1016/j.joi.2006.09.005/bib10/3	10.1016/j.joi.2006.09.005/bib10	1-s2.0-S1751157706000113	J.T.	Bruer	3
4	10.1016/j.joi.2006.10.001/bib17/1	10.1016/j.joi.2006.10.001/bib17	1-s2.0-S1751157706000125	L.	Egghe	1
5	10.1016/j.joi.2006.10.001/bib17/2	10.1016/j.joi.2006.10.001/bib17	1-s2.0-S1751157706000125	R.	Rousseau	2
6	10.1016/j.joi.2006.10.002/bib18/1	10.1016/j.joi.2006.10.002/bib18	1-s2.0-S175115770600023X	D.	Yarovsky	1

Figure 5.8: *Author_r* table and *Editor_r* table

5.2 Datasets

Two datasets have been used in this research:

- 1) A *Journal of Informetrics* (JOI) dataset¹: contains all the 420 publications from *Journal of Informetrics* related to the period 2007-2013.
- 2) A Renewable Energy (RE) dataset²: contain 15684 publications from 9 journals in the field of Renewable Energy. These publications cover the period 2001-2010.

Table 5.1 lists the 9 journals that are included in the RE dataset. Two criteria were used to select journals: 1) focus on the research area of renewable energy; 2) can be retrieved using Elsevier's text and data mining service (published by Elsevier).

Table 5.1: Journals included in the RE dataset.

Journal	No. of Publications
Biomass and Bioenergy	1265
Energy for Sustainable Development	340
Geothermics	360
International Journal of Hydrogen Energy	5310
Journal of Wind Engineering and Industrial Aerodynamics	893
Renewable and Sustainable Energy Reviews	954
Renewable Energy	2185
Solar Energy	1492
Solar Energy Materials and Solar Cells	2885

The number of publications, citations, references, and sections that is contained by both datasets is summarized in Table 5.2.

¹ Data collection took place on 8 April 2014.

² Data collection took place on 31 July 2014.

Table 5.2: Summary statistics of the JOI and RE datasets

Dataset	No. of Publications	No. of References	No. of Citations	No. of Sections	No. of Journals	Time Period
JOI	420	13,486	20,207	3,985	1	2007-2013
RE	15,684	394,577	513,482	166,616	9	2001-2010

Most publications contain several references. Figures 5.9 and 5.10 show the distribution of the number of references per publication in our two datasets. In these figures, the horizontal axis represents the number of references a publication has, and the vertical axis shows how many publications have the corresponding number of references. Figure 5.9 shows that the distribution of the number of references per publication in the JOI dataset approximately follows the normal distribution. The number of references per publication, except one outlier with 622 references, ranges between 0 and 111. Most of the publications (82%) have 6 to 50 references. The distribution of the RE dataset, which is shown in Figure 5.10, is more close to the normal distribution. The maximum number of references in one publication is 303. However, in Figure 5.10, we only plot the reference number that is less than 150. Most publications (93.57%) are located in the head of the distribution ($[0, 50]$), and the tail part ($[51, 303]$) only covers 6.43% of the publications.

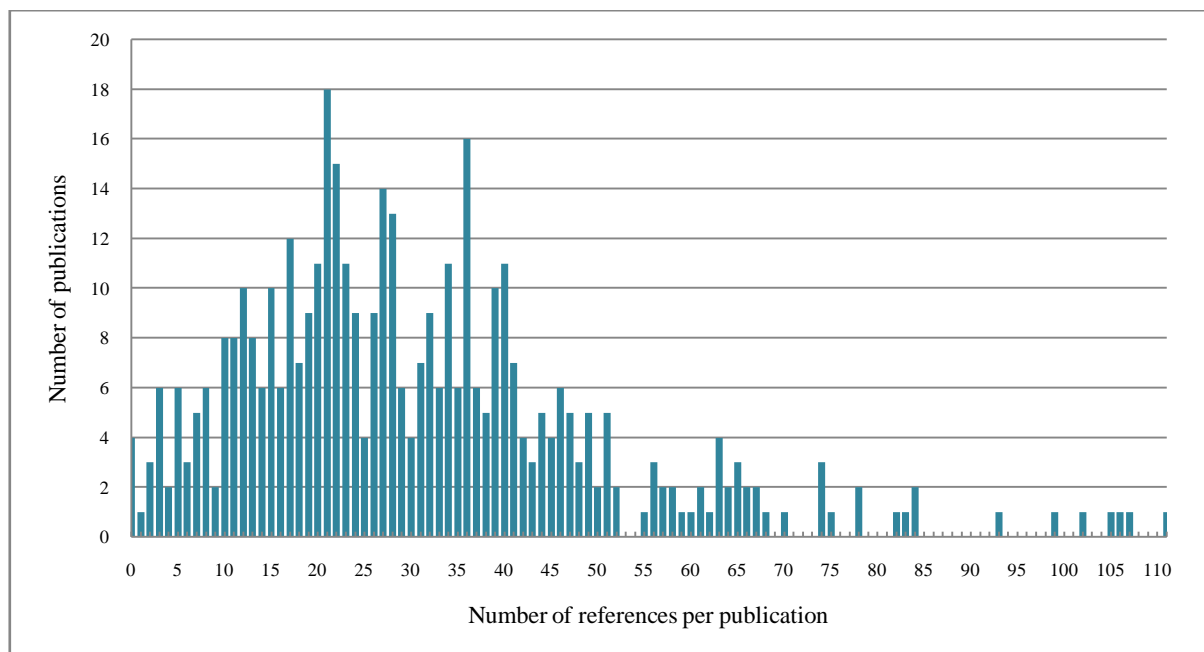


Figure 5.9: Distribution of the number of references per publication in the JOI dataset

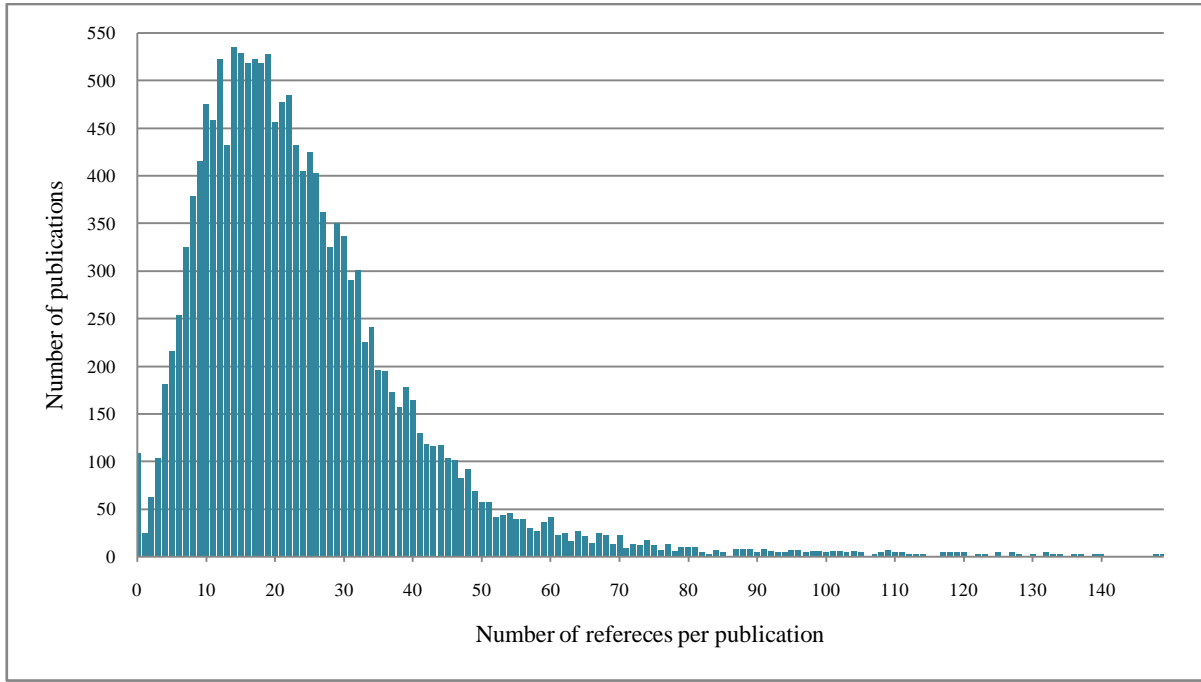


Figure 5.10: Distribution of the number of references per publication in the RE dataset

5.3 Section classification method

As is described in Chapter 3, to calculate the location score for references, references have to be assigned to the following five types of locations: introduction only, method, footnote only, appendix only, and others. The location types footnote only and appendix only can be directly identified from the structure of the full text. So no more processing is required. To identify the other three location types (introduction only, method, and others), some additional processing is needed. To properly identify these location types, the structure of publications in the JOI dataset have been analyzed.

According to Hu et al. (2013), a scientific publication is typically organized in four to six sections. This conclusion has been confirmed by our findings. Figure 5.11 shows the distribution of the number of sections per publication in the JOI dataset. Out of the 420 articles, 123 (29.29%) have 4 sections, 137 (32.62%) have 5 sections, and 76 (18.10%) have 6 sections. Therefore publications with four to six sections make up nearly 80% of the total publications. Here the number of sections is counted based on the level 1 sections. So subsections of the level 1 sections are not taken into account.

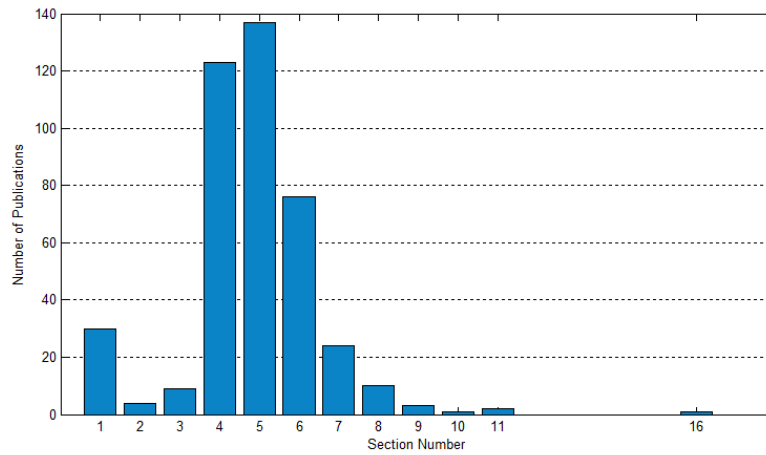


Figure 5.11: Distribution of the number of sections per publication in the JOI dataset

Figure 5.12 presents the words that are extracted from the title of each section and the size of the words represent their frequency of occurrence. Words that share the same stem were combined together. For example, “concluding”, “conclusion”, “conclusions”, “conclude” were combined to “conclu%”. If we look at the results, then we see that “introduction” is identified as the most commonly used word in the title of the first section. This observation is independent of the number of sections a publication has. In the title of the last section, the word “conclu%” (which represents “conclusion”, “conclusions”, “conclude”, and “concluding”) appears most frequently. Furthermore we can see that 4-section publications in most of the cases contain the sections *Introduction*, *Method*, *Result*, and *Conclusion*. In the case of 5-section publications, the second and third sections are likely to be *Data* and *Method*, but in some cases they also can be *Literature Review* and *Result*. The last two sections of 5-section publications are normally *Result/Discussion* and *Conclusion*. The 6-section publications are often organized in terms of *Introduction*, *Data*, *Method*, *Result*, *Discussion*, and *Conclusion*. However, the function of their second section is sometimes more ambiguous. Besides *Data* it also can be a description of related works.

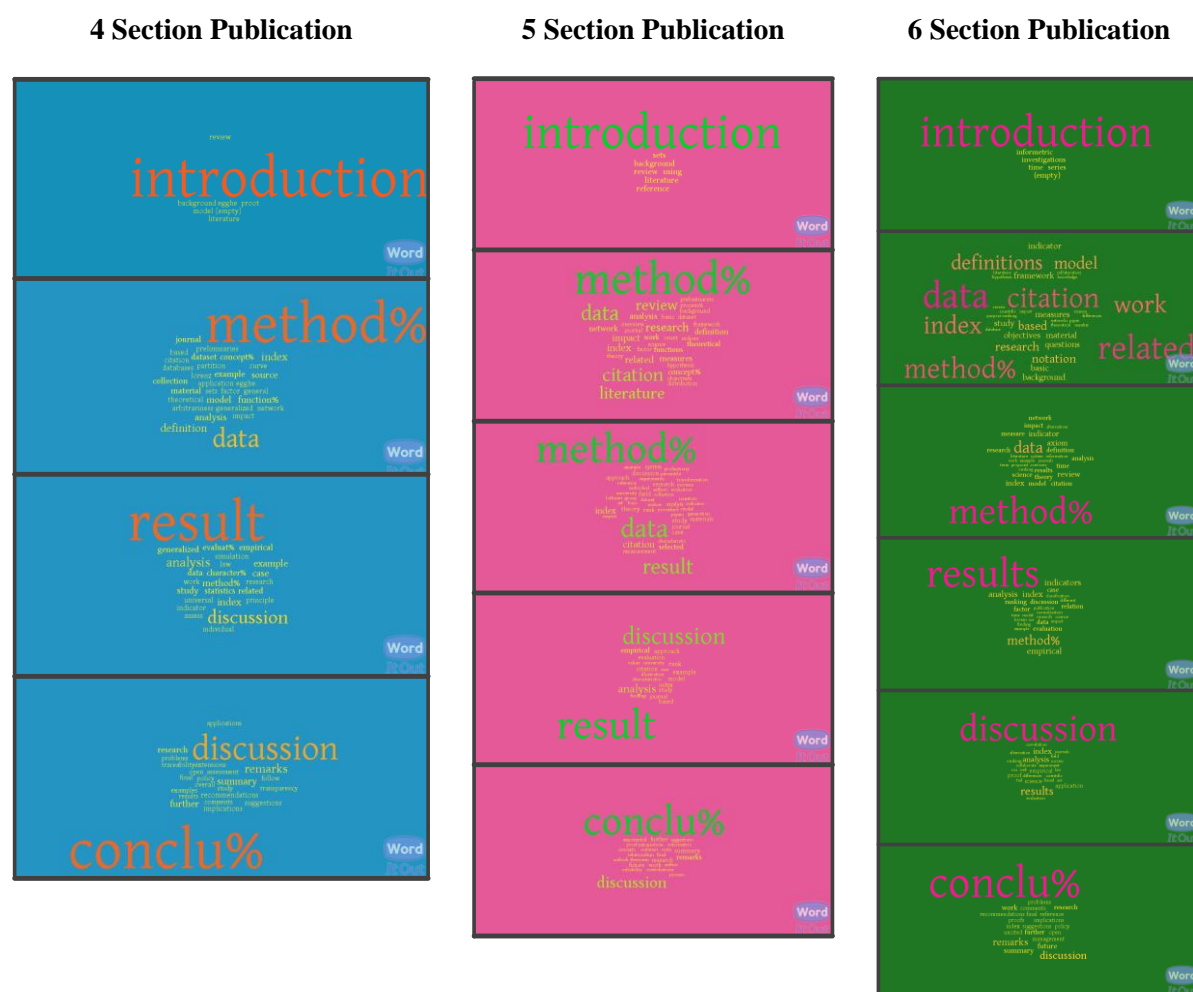


Figure 5.12: A word cloud visualization of section titles extracted from 4-section publications, 5-section publications, and 6-section publications. The word clouds are created using WordItOut (<http://worditout.com/>).

Within the JOI dataset, there are 34 publications that have only one or two sections and most of these publications are letters, editorials, or corrections.

The structure of these publications is different compared with other scientific publications. In most of cases, they don't have *Introduction*, *Method*, *Result*, and *Discussion* sections. So it is unnecessary and not possible to classify their sections according to the IMRAD framework. There are 9 publications that contain 3 sections. All their first sections are *Introduction* and the last sections are *Conclusion/Result*. But in most of cases, the second section is a combination of the *Review* section, the *Method* section, and the *Result* section.

Based on our findings, we manually created rules to automatically classify sections into the following four types: *Introduction*, *Method*, *Result+*, and *Others*. *Result+* is a combination of

Result, *Discussion*, and *Conclusion*. *Others* are sections that cannot be classified into the other three types. The rules to automatically classify sections are as follows:

- Rule 1: If a publication only has one or two sections, all its sections are classified as *Others*;
- Rule 2: If a publication has three sections, the 1st section is classified as *Introduction*, the 2nd section is classified as *Others*, and the 3rd section is classified as *Result+*;
- Rule 3: If a publication has more than three sections, the 1st section is classified as *Introduction* and the last section is classified as *Result+*;
- Rule 4: Sections that cannot be classified based on rules 1, 2, and 3 will be classified based on the word stems contained in their title. The word stems and their corresponding section type are listed in Table 1.1. If the title contains word stems that are related to certain section type, this section is classified as that type. However, if the title contains word stems that are related to multiple section types, this section is classified as *Others*.

Table 5.3: Word stems for each section type

Section Type	Word Stems
Introduction	introduction, background, review
Method	method, data, material
Result+	result, discussion, conclu, summary, remark

By applying the above presented rules, we ended up with 417 *Introduction* sections, 208 *Method* sections, 654 *Result+* sections, 41 *Others* sections, and 2665 unknown sections. To improve the accuracy of our classification, more rules are created based on the section sequence and the number of sections per publication.

For 4-section publications:

- Rule 5: If the 2nd section is identified as *Result+* and the 3rd section as *Method*, then this classification is probably wrong. Therefore, in this case the 2nd section will be classified as *Method*, and the 3rd section as *Others*.
- Rule 6: If there is no section identified as *Method*, the second section will be classified as *Method* section;

For 5-section publications:

- Rule 7: If there is no section identified as *Method* and the 3rd and/or 4th section is identified as *Result+*, then the section before the first *Result+* section is classified as *Method*;
- Rule 8: If there is no section identified as *Method* and neither the 3rd nor the 4th section is identified as *Result+*, then the section after the last *Introduction* section is classified as *Method*.

For 6-section publications:

- Rule 9: If there is no section identified as *Method* and among the 3rd, 4th, and 5th sections at least one is identified as *Result+*, then the section before the first *Result+* section is classified as *Method*;
- Rule 10: If there is no section identified as *Method* and all the 3rd, 4th, and 5th sections are not identified as *Result+*, then the section after the last *Introduction* section is classified as *Method*.

Based on these 10 rules, finally we identified 417 *Introduction* sections, 382 *Method* sections, 652 *Result+* sections, 41 *Others* sections, and 2493 unknown sections.

Finally, to calculate the location score, all the references that are only cited in the *Introduction* section will have the reference location “Introduction Only”. References that are cited at least once in the *Method* section will be assigned the reference location “Method”. The references that are only cited in the footnote section are “Footnote Only”. The references that are only cited in the appendix section are “Appendices Only”. All the other references that are not covered by the above four situations will have the reference location “Others”.

5.4 Importance of references in the JOI dataset

To get the importance of references, first we download the full text files for the JOI dataset, then extract and store the data into the database that are described in Section 5.1. Next we classify the sections in the database using the rules that are created in Section 5.3. Finally, the importance of references is calculated based on the model which is developed in Chapter 4. Figure 5.13 shows the distribution of the importance of references.

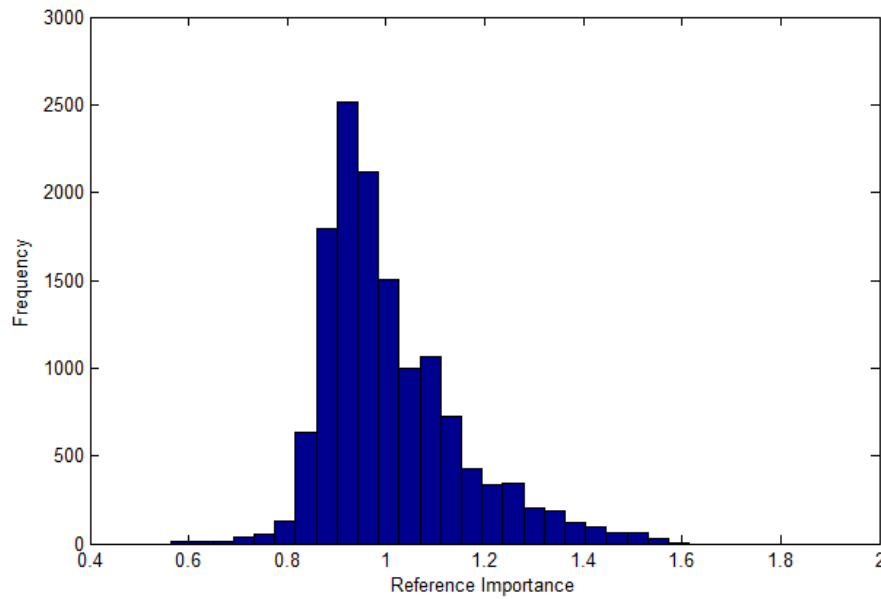


Figure 5.13: Distribution of reference importance

The histogram plot above provides an overview of the distribution of the reference importance of the 13486 references contained in the 420 publications of the JOI dataset. The reference values are distributed within the range $[0.5646, 1.6146]$, and 85% of reference values are between 0.82 and 1.17. From this result we can see that in general the reference importance follow the normal distribution. So for most of the references their importance is closely concentrated around the mean value (1.0080). We also notice that the distribution is slightly positively skew. This means that for more than half of the references, their importance is below average.

Chapter 6

METHOD VALIDATION: AUTHOR-RATED IMPORTANCE OF CITED REFERENCES

6.1 Methodology

In the beginning of Chapter 3, we defined the reference importance as the amount of contribution that the reference makes to the citing publication. In Chapter 4 and Chapter 5, we measured the reference value based on multiple citation features (frequency, location, treatment, and self-citation). However, for the question “how important a reference is”, we still believe that it could be best answered by the authors of the citing publications themselves. By comparing the reference importance given by the authors with the value calculated by our model, we can evaluate the performance of our model.


However, sometimes the authors may be wrong about how much contribution a reference makes to its citing publication. According to Zhu et al. (in press), there are two types of situations where the authors’ judgment may be biased. In the first situation, the author may say a reference is important because this reference is very authoritative or very popular. In the second situation, a reference may influence the authors’ opinion at the subconscious level or the authors don’t want to admit that they were influenced by this reference. So even if the reference contributed a lot to the publication, the authors may say it is not important. Although the authors’ feeling may be inaccurate, this is the most reliable way to measure the importance of the references. Therefore, within this chapter the model we developed to calculate the reference value is validated based on author-rated data.

Dietz, Bickel, and Scheffer (2007) asked the authors of 22 publications to manually label the strength of influence of references they cited on a Likert scale. Zhu et al. (in press) collected an important reference dataset by guiding the authors to provide a list of essential references of their paper. Tang and Safer (2008) asked the participating authors to rate the importance of the references on a seven-point scale from “slightly important” to “extremely important”. In

our research, we asked the authors to first identify the essential references and then rank them according to their importance.

6.2 A web-based survey

A web survey is sent to the corresponding authors of publications in our JOI dataset, so that they can help us to identify the essential references in their publication. In the survey, for each publication of an author we list all its references and the author can identify about five of them as essential references. As we all know, not all references are equally important of a citing publication, and to keep the survey easy for the authors, we only asked them to identify the five most essential references for each publication. Then based on how many contributions the reference makes to its citing paper, these five essential references are ranked by the author from 1 to 5. Figure 6.1 is an example of the web survey.



Leiden University
CWTS
CWTS B.V.
Other CWTS sites

Survey

Identification of essential references in scientific publications

Dear N.J. van Eck,

The Centre for Science and Technology Studies (CWTS) of Leiden University is working on a research project aimed at developing an algorithm for identifying the most essential references in scientific publications. To measure the accuracy of such an algorithm, we are building a test set of publications for which we know which references are considered most essential by the authors. Essential references are references that are highly influential or inspirational for the core ideas in a publication, for instance references that inspired or strongly influenced a new algorithm, an experimental design, or the choice of a research problem.

We would like to ask for your help. You have been identified as the corresponding author of a publication that has appeared in the *Journal of Informetrics* in the period 2007–2013. Could you please help us by indicating the most essential references in this publication?

Your publication is listed below. The full list of references is provided. Please indicate the five most essential references. First identify the reference that you consider most essential and assign it to rank 1, then identify the second most essential reference and assign it to rank 2, and so on. If you consider two or more references equally essential, you may assign them the same rank. If you find it difficult to identify the five most essential references, you may choose to identify only the one, two, three, or four most essential references.

We plan to make the results of this survey publicly available so that the results can be used not only by ourselves but also by other researchers.

In case of questions, please do not hesitate to contact us by sending an e-mail to [Xi Cui](mailto:Xi.Cui).

We thank you for your cooperation.

Xi Cui and Nees Jan van Eck

Publication 1

Author: van Eck, N.J., Waltman, L.
Title: [Generalizing the h- and g-indices](#)
Source: *Journal of Informetrics*, 2(4), 263-271, 2008

Rank	Reference
<input type="text"/>	Anderson, T.R., Hankin, R.K.S., & Killworth, P.D. (2008). Beyond the Durfee square: Enhancing the h-index to score total publication output. <i>Scientometrics</i> , 76(3), 577-588.
<input type="text"/>	Baill, P. (2005). Index aims for fair ranking of scientists. <i>Nature</i> , 436, 900
<input type="text"/>	Bar-Ilan, J. (2006). h-Index for Price medalists revisited. <i>ISSI Newsletter</i> , 2, 3-5
<input type="text"/>	Batista, P.D., Campitelli, M.G., Kinouchi, O., Martinez, A.S. (2006). Is it possible to compare researchers with different scientific interests?. <i>Scientometrics</i> , 68, 179-189
<input type="text"/>	Bornmann, L., Daniel, H.-D. (2005). Does the h-index for ranking of scientists really work?. <i>Scientometrics</i> , 65, 391-392
<input type="text"/>	Bornmann, L., Daniel, H.-D. (2007). What do we know about the h index?. <i>Journal of the American Society for Information Science and Technology</i> , 58, 1381-1385
<input type="text"/>	Bornmann, L., Mutz, R., Daniel, H.-D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. <i>Journal of the American Society for Information Science and Technology</i> , 59, 830-837
<input type="text"/>	Costas, R., Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. <i>Journal of Informetrics</i> , 1, 193-203
<input type="text"/>	Costas, R., & Bordons, M. (in press). Is g-index better than h-index? An exploratory study at the individual level. <i>Scientometrics</i> .
<input type="text"/>	Egghe, L. (2006). An improvement of the h-index: The g-index. <i>ISSI Newsletter</i> , 2, 8-9
<input type="text"/>	Egghe, L. (2006). Theory and practise of the g-index. <i>Scientometrics</i> , 69, 131-152
<input type="text"/>	Egghe, L., Rousseau, R. (2006). An informetric model for the Hirsch-index. <i>Scientometrics</i> , 69, 121-129
<input type="text"/>	Egghe, L., Rousseau, R. (2008). An h-index weighted by citation impact. <i>Information Processing and Management</i> , 44, 770-780
<input type="text"/>	Glänzel, W., Persson, O. (2005). h-Index for Price medalists. <i>ISSI Newsletter</i> , 1, 15-18
<input type="text"/>	Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. <i>Proceedings of the National Academy of Sciences</i> , 102, 16569-16572
<input type="text"/>	Hirsch, J.E. (2007). Does the h index have predictive power?. <i>Proceedings of the National Academy of Sciences</i> , 104, 19193-19198
<input type="text"/>	Iglesias, J.E., Pecharrromán, C. (2007). Scaling the h-index for different scientific ISI fields. <i>Scientometrics</i> , 73, 303-320
<input type="text"/>	Jin, B. (2007). The AR-index: Complementing the h-index. <i>ISSI Newsletter</i> , 3, 6
<input type="text"/>	Jin, B., Liang, L., Rousseau, R., Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. <i>Chinese Science Bulletin</i> , 52, 855-863
<input type="text"/>	Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original h-index. <i>ISSI Newsletter</i> , 2, 4-6
<input type="text"/>	Lehmann, S., Jackson, A.D., Lautrup, B.E. (2006). Measures for measures. <i>Nature</i> , 444, 1003-1004
<input type="text"/>	Lehmann, S., Jackson, A.D., Lautrup, B.E. (2008). A quantitative analysis of indicators of scientific performance. <i>Scientometrics</i> , 76, 369-390

Figure 6.1: A sample page of the web survey

We sent this web survey to 205 corresponding authors of 420 publications in the JOI dataset at 12 June 2014. Most of the authors only have one publication in our dataset, but for some of them they have multiple publications. So the number of authors is less than the number of publications. Until 16 July 2014, 65 authors had finished the survey and 111 publications had been reviewed. Within these 111 publications, there are 3410 references and 648 of them are labeled as essential references by the authors (ranked as 1, 2, 3, 4 or 5). Although we asked the author to identify five most essential references per publication, some of them selected more or less than this number. So the number of essential references (648) is not exactly five times the publication number (111). Figure 6.2 shows the distribution of the number of publications with a certain number of essential references.

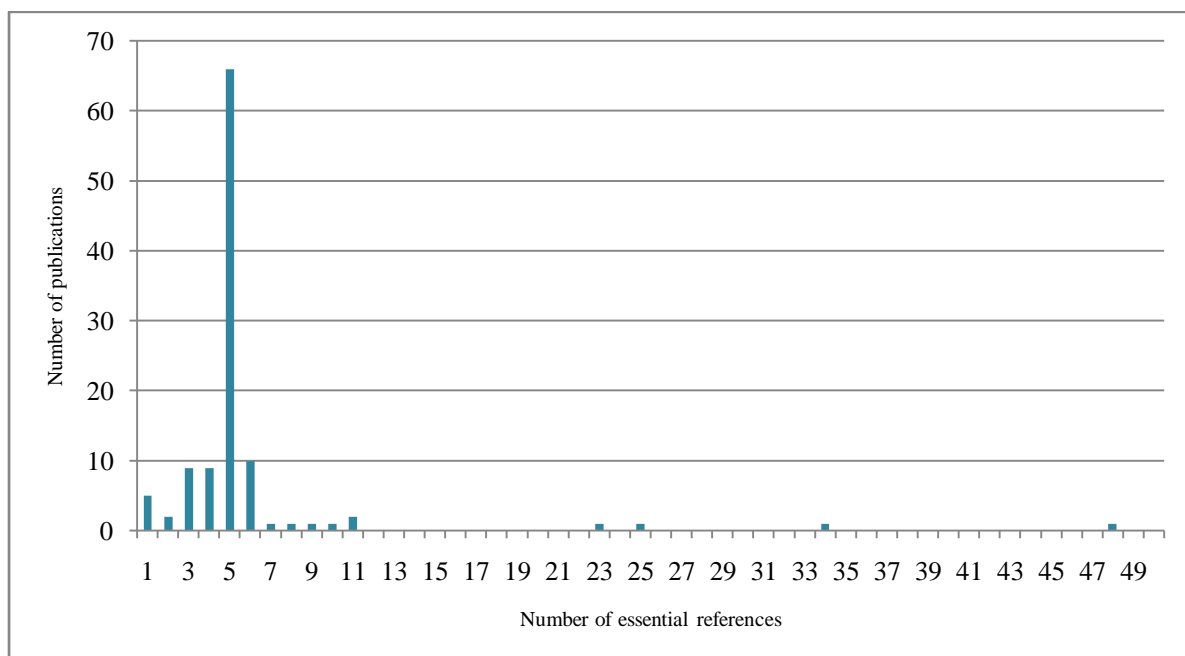


Figure 6.2: Distribution of the number of publications with a certain number of essential references.

In Table 6.1 we can see that most of the publications have 3 to 5 essential references. But when we investigated the four publications that contain more than 20 essential references into more detail, we noticed that in these publications all the references were identified as important by the author. This situation may influence the reliability of the following analysis, so we manually adjusted the rank of these publications. To reduce the number of essential references to about 5, only the references that were ranked as 1 or 2 were remained, and all the others were deleted from the essential references collection. So after this change, 556 references were labeled as essential reference and the average number of essential references per publication was 5.009. Table 6.1 is a summary of statistics of the final survey results.

Table 6.1: Summary statistics of the survey results

# of publication	# of reference	# of essential reference	Average number of essential references per publication
111	3410	556	5.009

6.2 Validation based on survey results

Based on the data we collected from the web survey, in this section a commonly used measure from information retrieval, the so-called F-measure, is used to evaluate the performance of the newly developed reference importance model. This F-measure indicates the accuracy of the model by considering both the precision and recall at the same time.

Precision and recall measure a model's performance from two different directions. Precision is the percentage of important references predicted by the model that are also identified as important by the author. Recall is the percentage of important references identified by the author that are also predicted as important by the model. Based on the precision and recall, the F-measure is defined as: $F = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$.

The meaning of precision and recall can be intuitively explained using Figure 6.3.

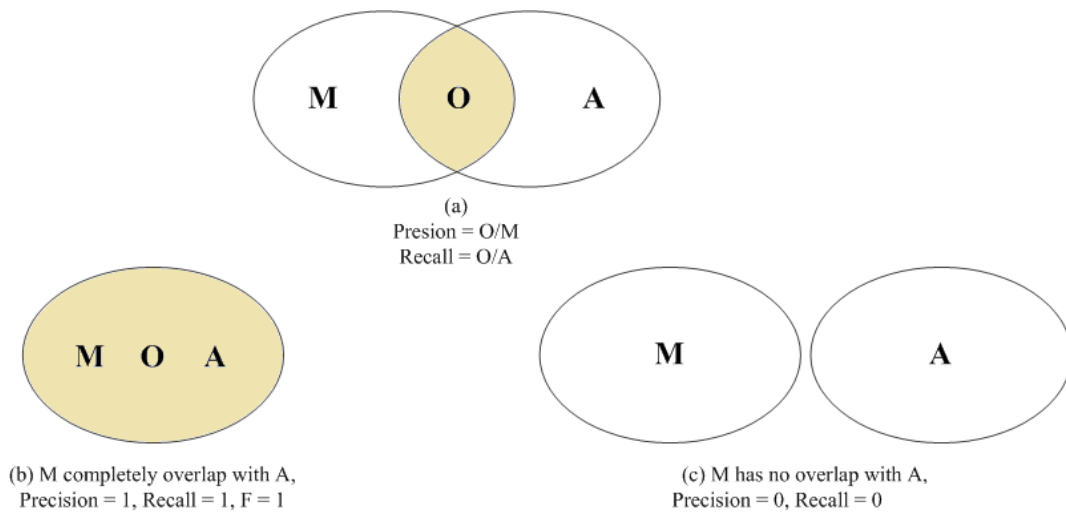


Figure 6.3: Relationship between precision and recall

As it is shown in Figure 6.3 (a), the left circle (M) represents the set of references that are predicted as important by the model and the right circle (A) represents the set of references

that are identified as important by the authors. The overlapping part (O) of these two circles indicates the references that are assessed as important by both our model and the author at the same time. The precision of the model is the percentage that the overlapping part (O) is accounted in the model predicted set (M). The recall of the model is the percentage that the overlapping part (O) is accounted in the author identified set (A). If the model predicts exactly the same important references as the author, the model predicted set (M) and the author identified set (A) will completely overlap (Figure 6.3 (b)). In this case, the precision and recall are both equal to 1, so the F-measure of the model is also 1. However, if the model predicts totally different with the author's result, there will be no overlap between the model predicted set (M) and the author identified set (A) (Figure 6.3 (c)). In this case, both precision and recall are 0.

To calculate the F-measure, we need to create a set of important references that is predicted by the model. According to the survey results, there are about five important references per publication (Table 6.1). So to give the model-identified important reference set the similar size as the author-identified one, we decided to label the top five references in each publication, with the highest reference important value, as important.

Table 6.2 shows the evaluation result of the model. 37.1% of important references that are predicted by the model are also identified as important by the author. 44.8% of important references that are identified by the author are also predicted as important by the model. The F-measure of the model is 0.4059. To more clearly show the performance of the model, we introduced a baseline which randomly labels five references as important in each publication. The F-measure for this random model is 0.1783. Because the F-measure, precision, and recall of the reference importance model are all much higher than that of the baseline, we conclude that this model indeed can predict the importance of references.

Table 6.2: Evaluation results of the reference importance model

Model	F-measure	Precision	Recall
Reference Importance Model	0.4059	0.3711	0.4478
Random	0.1783	0.1630	0.1968

During the calculation of F-measures, we didn't differentiate the author's rank of the important references. If this model can measure the importance of references, the references

with a higher author rank should be easier to be identified by the model. So to test this hypothesis, we calculate the percentage of references with a certain rank that are also retrieved by the model. The result shown in Table 6.3 supports this hypothesis. 51% of the references that are ranked as 1 by the author are retrieved by the model, and 33% of the references with rank 5 are identified by the model. Except for a slightly increase in the probability of author rank 2 references, in general the higher ranked references are more likely to be retrieved by our model. This phenomenon also indicates that the model can predict the importance of reference.

Table 6.3: The probability that references with a certain author rank are identified as important by the model

Author's Rank	1	2	3	4	5
Probability	50.99%	58.88%	42.00%	39.39%	33.33%

6.3 Optimization of the model using author-rated importance of references

As it has been described in Section 4.6, the reference importance model use four weights (p_f , p_l , p_t , and p_s) to control the contributions each indicator-level score makes to the final reference importance value. Within Section 4.6, these weights were decided based on the conclusions of previous studies. However, that is only a rough estimation and is also doubtful that these weights are suitable enough for our research. Since we have already got the author-ranked importance of the references, our idea is to use the ranks to optimize these weights.

The author-ranked reference importance result is a combination of binary data and ordered data. For its binary part, the references are classified into two types: important reference and unimportant reference. For the ordered part, the important references are ranked by the authors into 5 scales (1 to 5), and 1 means most important and 5 means least important. So to perform the optimization, we decided to eliminate the ordered aspect of the results and only use its binary aspect to do the analysis.

Here we choose a logistic regression model to get the new weights for each indicator-level score. Logistic regression represents the probability of the binary variable Y in the form of:

$$P(Y = 1 | \bar{X}) = \frac{1}{1 + e^{-(\beta_0 + \sum_i \beta_i X_i)}}$$

where $\bar{X} = (X_1, X_2, \dots)$ are feature values and $\beta_0, \beta_1, \beta_2, \dots$ are weights for each feature. For our optimization problem, the binary variable Y is a vector representing if a reference is important (1) or unimportant (0) according to the author. Each \bar{X} has four feature values: $\bar{X} = (S_f, S_l, S_t, S_s)$.

The *mnrfit* command in Matlab is used to perform the logistic regression, and the 3410 references in 111 publications that were examined by the authors are the training instances. Finally we get the weights assigned to each features. Table 6.4 shows the weights:

Table 6.4: Optimized weights for the model (before rescaling)

Weight	Value
p_f : (frequency score weight)	3.5017
p_l : (location score weight)	0.7238
p_t : (treatment score weight)	1.2263
p_s : (self-citation score weight)	0.2774

The absolute values of the weights indicate the relative importance of the corresponding features in the model (Zhu et al., in press). So the most important citation feature is citing frequency and the least important one is self-citation. The performance of location and treatment as an indicator of reference importance is in between that of frequency and self-citation. This sequence of the relative importance is the same with what we have found in the previous studies (Section 4.6).

Figure 6.4 is the distribution of reference importance calculated using the weights in Table 6.4 (here the constant C in Equation 4.3 is 0). In general, the reference importance value follows the normal distribution. Most of the records are located around the mean value (5.4401). All the values range between 3.8487 and 7.9967. Similar with the reference important that was calculated using the simple weights (Section 5.4), this distribution is also a slightly positively skewed one and it has a relatively long tail at the right side.

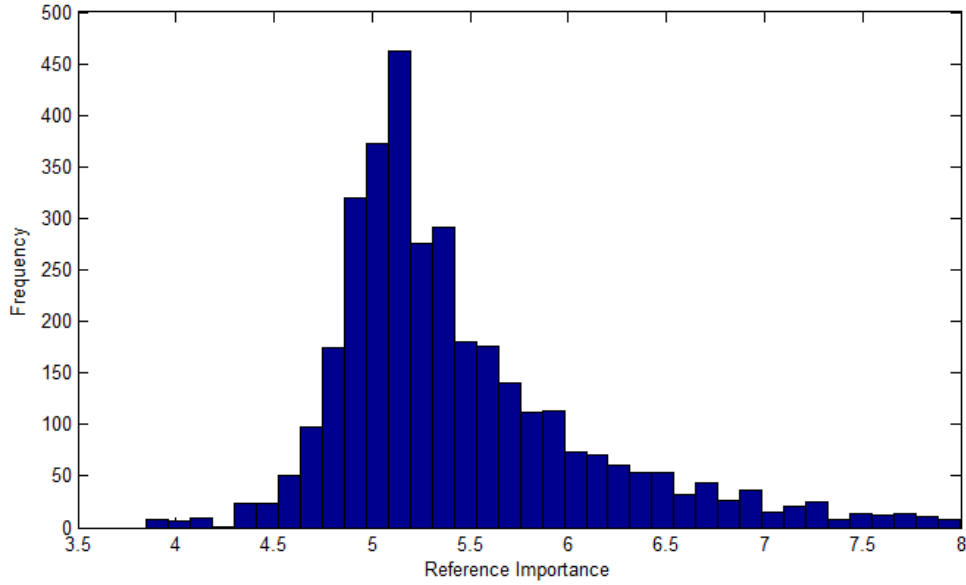


Figure 6.4: Distribution of reference importance (using the weights in Table 6.5)

To make the results easier to interpret, we decided to rescale the values to (0, 2) and adjust the mean value to around 1. In this way, the reference importance results can be explained in the same way as the indicator-level scores (Figure 4.2). So for this purpose, the reference importance R is processed as: $R' = R \times 0.3 - 0.6320$. Here R' is the new reference important value. Reflecting this rescale process on the weights, weights in Table 6.4 should be multiply by 0.3 and the constant (C) is assigned -0.6320. Table 6.5 shows the new weights for Equation 4.3.

Table 6.5: Optimized weights for the reference important model (after rescaling)

Weight	Value
p_f : (frequency score weight)	1.0505
p_l : (location score weight)	0.2171
p_t : (treatment score weight)	0.3679
p_s : (self-citation score weight)	0.0832
C : (constant)	-0.6320

If we compare the optimized weights in Table 6.5 with the original weights in Table 4.2, we can find that the weights optimized by the survey data have the same pattern with the weights

selected based on the literature review. In both sets of weights, the value of p_f is significantly higher than the others. The weight for the self-citation score (p_s) is quite low compared with other score weights (p_f , p_l , and p_t). The location score weight (p_l) and the treatment score weight (p_t) have similar values. The only negative weight is the constant used in the model.

Using the weights in Table 6.5, we recalculated the reference importance values for the JOI dataset. Figure 6.5 is the distribution of the results. Compared with the distribution in Figure 6.4, this one has a smaller range ([0.5226, 1.7670]) and its mean is equal to 1.00.

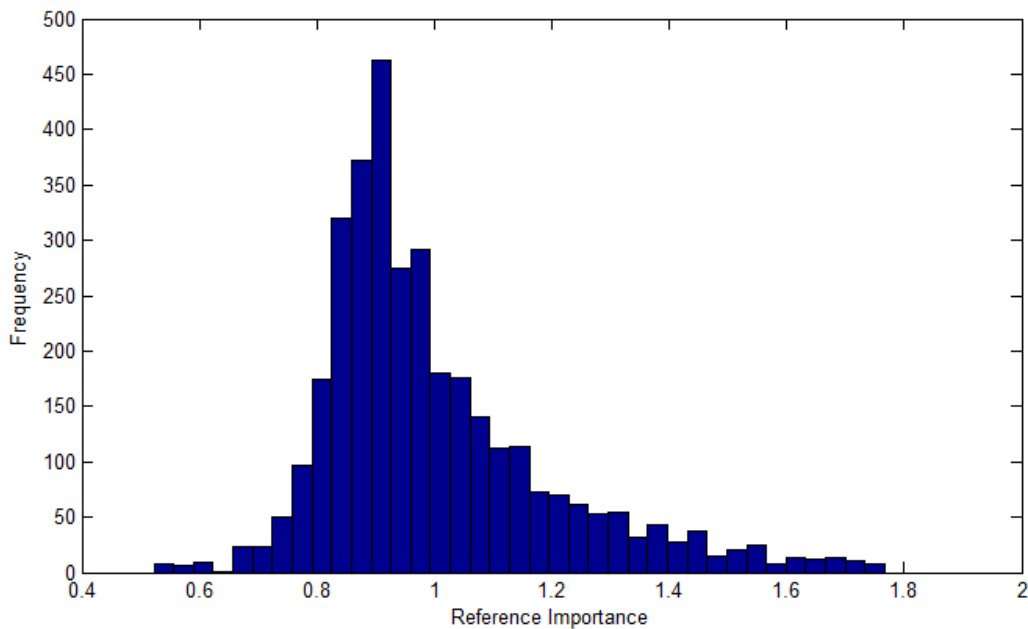


Figure 6.5: Distribution of reference important (using the weights in Table 6.6)

Again the F-measure, precision, and recall are calculated for this optimized model. The result is shown in Table 6.6. Compared with the results for the original model (Table 6.2), all the measures for this new model are slightly better than the old ones. Now 37.3% of the important references that are predicted by the model are also identified as important by the author. 45.7% of the important references that are identified by the author are also predicted as important by the model. The F-measure is improved from 0.4059 to 0.4107. Therefore, by optimizing the weights based on the survey data, the performance of the model is improved.

Table 6.6: Evaluation of the optimized model performance

F-measure	Precision	Recall
0.4107	0.3730	0.4568

In Chapter 7, the reference important model will be applied to improve the structure of citation networks. Instead of the original model (Table 4.2), this optimized model (Table 6.5) will be used.

Chapter 7

APPLICATION IN CITATION NETWORKS

7.1 Citation networks

In this chapter we are going to reduce the noise in citation networks by removing unimportant citation relations. We expect that in the new citation networks, the main structure of scientific fields will become clearer.

Bibliometric networks are used to quantitatively analyze and visualize scientific literature based on the bibliographic data. There are various types of bibliometric networks and each of them provides somewhat different information and can be used for different purposes. In general, bibliometric networks can be used to get an overview of the structure of the scientific literature in a certain domain or on a certain topic (van Eck, 2011). Bibliometric networks are constructed based on different types of relations between bibliographic entities. These relations include co-authorship relations between researchers, co-occurrence relations between keywords extracted from title and abstract, and, most frequently, citation relations between publications. Bibliometric networks constructed based on citation relations are also called citation networks. Examples of citation relations include co-citation, bibliographic coupling, and direct citation (van Eck & Waltman, in press). Within this chapter, we will focus on how to reduce the noise in citation networks that are constructed based on the direct citation relations.

In direct citation networks, each node represents a publication and each edge represents a citation relation between two publications. According to Egghe and Rousseau (1990), the existence of a citation relation indicates that there is a relationship between the cited reference and citing publication from the author's point of view. Therefore, by analyzing citation networks we can get an intuitive understanding of how publications are related with each other and furthermore the main structure of scientific fields. Based on this idea, the quality of citation networks depends on whether citation relations can truly indicate the relatedness of publications. Since references are sometimes chosen arbitrarily by the authors, not all references can represent a strong and clear relatedness between cited and citing publications

(van Eck, 2011). Therefore, edges that are produced by arbitrarily chosen references will introduce a certain amount of noise into citation networks. One possible way to reduce this noise in citation networks is to remove the edges that are produced by weak citation relations. The reference importance model introduced in Chapter 4 can be used to measure the importance of references. Important references indicate a strong relatedness between two publications and unimportant references indicate a weak relatedness. Therefore, based on the importance of references calculated by the model, we can filter out the less important references from citation networks. In this way, we want to reduce the noise in citation networks and make the structure of scientific fields easier to extract.

7.2 Construction of reduced citation networks

Our idea is to filter out the least important citation relations in citation networks, and that by doing this, the main structure of scientific fields will become easier to extract and further analyze. To test this idea, first we calculated the reference importance for all the references in the JOI dataset using the optimized model which is developed at the end of Chapter 6. Then to filter out the less essential references, all the references are sorted by their reference importance in descending order. The higher the importance of a reference is, the more essential the reference might be. Then instead of all the references, we only use a certain percentage of references (such as top 40%) to construct the citation network. Based on this idea, citation networks for the JOI dataset are constructed and then visualized using CitNetExplorer (van Eck & Waltman, in press). Figure 7.1 is the visualization of the original citation network of publications in the JOI dataset, Figures 7.2 and 7.3 are the reduced citation networks. In these visualizations, each circle represents a publication. A publication is labeled by the last name of the first author. For easy interpretation, only the 40 publications that have been most frequently cited are included in the visualization. The vertical location of a publication represents its publication year and the horizontal distance between two publications is determined by their citation relations. Although all the papers published in *Journal of Informetrics* are related to the quantitative analysis of science, they still can be assigned to several sub-groups. Therefore based on their citation relations, CitNetExplorer clusters the publications into different groups, and the color of a publication in the visualizations indicates the group to which the publication is assigned.

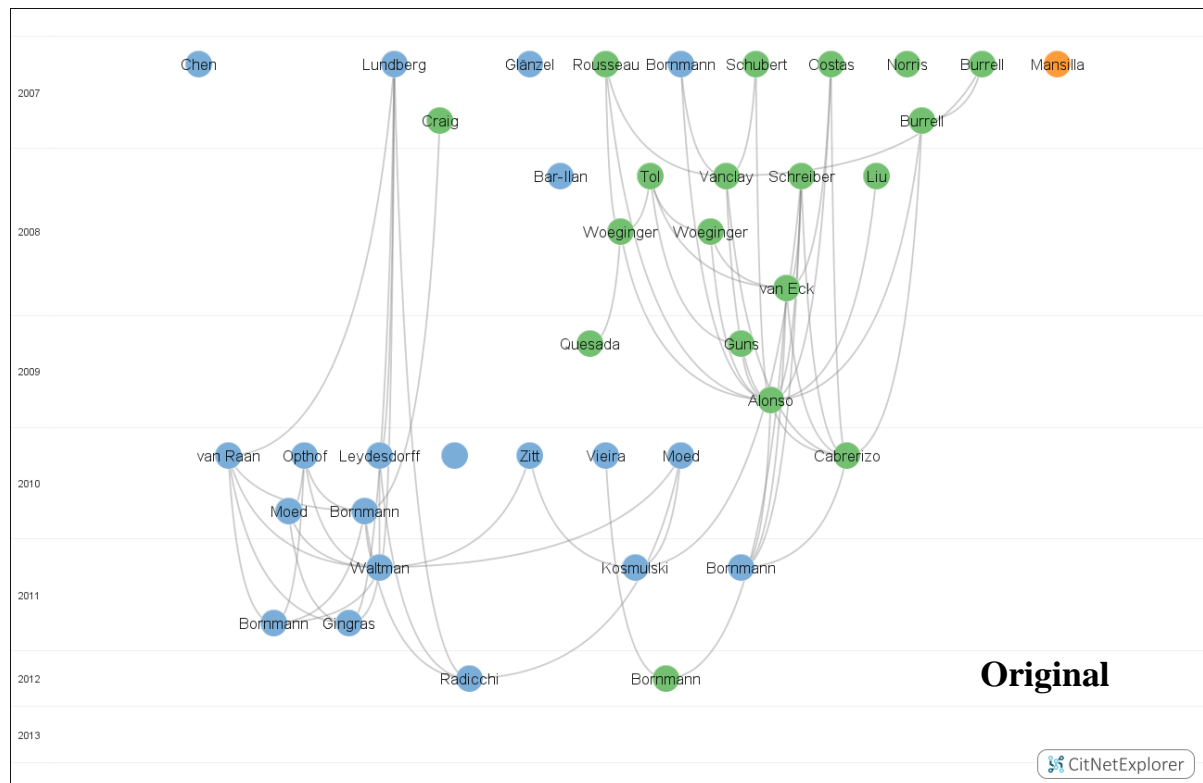


Figure 7.1: Visualization of the original citation network of publications in the JOI dataset

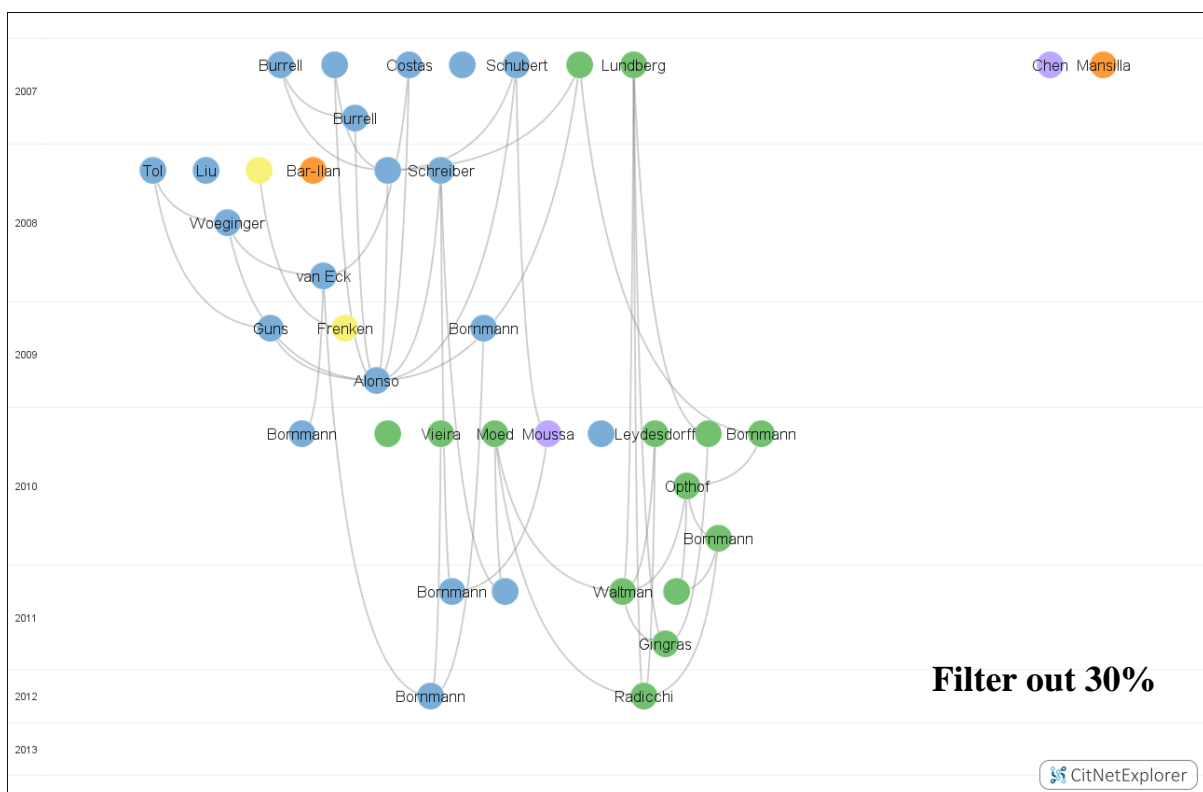


Figure 7.2: Visualization of the reduced citation network in which 30% of less important references is filtered out

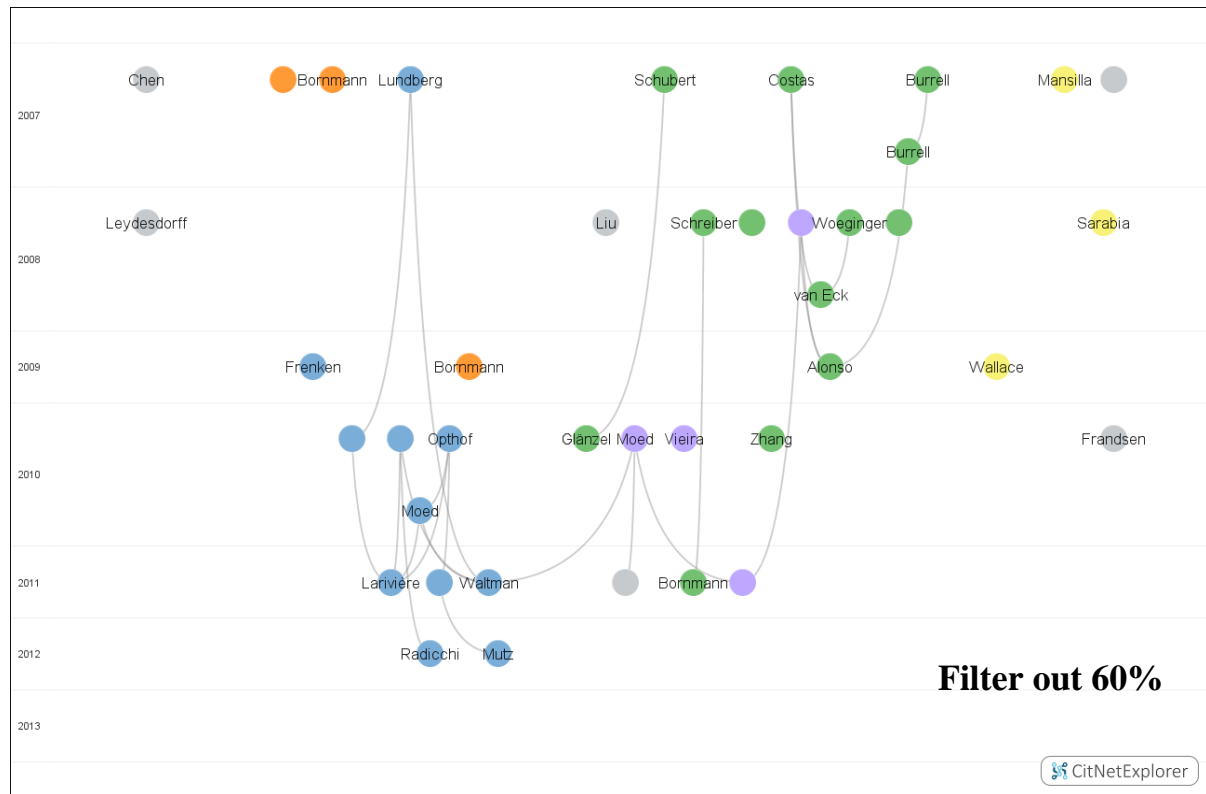


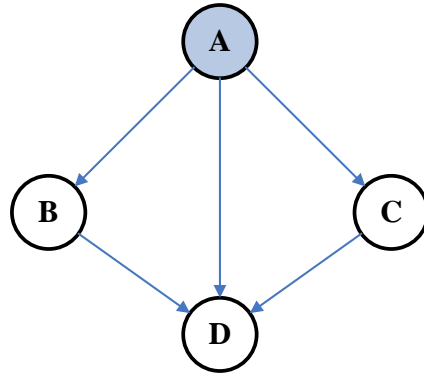
Figure 7.3: Visualization of the reduced citation network in which 60% of less important references is filtered out

7.3 Quantitative analysis of the reduced citation networks

In the previous section, we have generated reduced citation networks in which less important citation relations are removed. We expect that the reduced citation networks contain less noise and the main structure is also clearer. However based on the visualizations of the reduced citation networks (Figures 7.1, 7.2 and 7.3), it is difficult to observe the improvement intuitively. Therefore in this section, we are going to analyze the citation networks using a quantitative method. Instead of the JOI dataset, in this section the RE dataset will be used to do the analysis. The RE dataset contains more publications than the JOI dataset, so it has a better coverage of publications in its scientific field (renewable energy). The main structure of citation networks constructed based on the RE dataset will be easier to extract by the quantitative measures.

In citation networks, each node represents a publication, and each edge represents a citation relation between two publications. The direction of the edge is from the citing publication to the cited publication. So citation networks can be seen as a kind of directed graph. In graph

theory, the local clustering coefficient of a node measures how close its neighbors are connected with each other (Watts & Strogatz, 1998). Given a graph $G=(N,E)$ which includes a set of nodes N and a set of edges E . Neighbors of node $n \in N$ are the nodes that are directly connected with n . The clustering coefficient of n is defined as the number of edges between its neighbors divided by the maximum number of edges that could exist between them. The calculation of clustering coefficient also can be explained by Figure 7.4.



Neighbors of node A: B, C, D;

Number of edges that are within the neighbors of node A: 2 ($B \rightarrow D, C \rightarrow D$);

Maximum number of edges that could exist between the neighbors of node A: $3 \times (3-1) = 6$;

Clustering coefficient of node A: $2/6$

Figure 7.4: Clustering coefficient example

Figure 7.4 shows a graph that is made up of four nodes (A, B, C, and D). The clustering coefficient of node A will be calculated. Within this graph, node A has 3 neighbors (B, C, and D). There are two edges within the neighbors of node A ($B \rightarrow D, C \rightarrow D$). Because this is a directed graph, so the maximum number of edges that could exist between the neighbors of node A is $3 \times (3-1) = 6$. Then the clustering coefficient for node A is $2 \div 6 = \frac{1}{3}$.

To measure the overall level of clustering in our reduced citation networks, for each network we calculate the average clustering coefficient which is the arithmetic average of the local clustering coefficients of all the publications in the network (Kemper, 2010).

To evaluate the structure of the reduced citation networks, the average clustering coefficient for the following two groups of citation networks are calculated:

- Group 1: Citation networks in which n% of less important references are filtered out based on their reference importance;

- Group 2: Citation networks in which n% of references are filtered out randomly.

Here group 2, the random group, is used as a baseline. In group 1 and group 2, citation networks are reduced to the same size. So the new citation networks of group 1 and group 2 have the same number of nodes and edges. The difference between these two groups is that in group 1 only the least important citation relations are removed. But in group 2, the important citation relations and unimportant citation relations are equally likely to be removed. To get more reliable results of group 2, we will repeat the random sampling process for 100 times and use the average number as the final result.

The average clustering coefficients (CC) for group 1 and group 2 citation networks are calculated. Table 7.1 shows the results.

Table 7.1: Average clustering coefficient for group 1 and group 2 citation networks

% of References Removed from the Original Citation Network	Group 1	Group 2 [*]	Group 1 - Group 2 ^{**}
10%	0.017988	0.018569	-0.000580
20%	0.016527	0.015521	0.001005
30%	0.015614	0.012651	0.002963
40%	0.013555	0.009825	0.003729
50%	0.012059	0.007226	0.004834
60%	0.010862	0.004886	0.005976
70%	0.008283	0.002809	0.005474
80%	0.005107	0.001230	0.003876
90%	0.002363	0.000247	0.002116

^{*} To get a more reliable result, the random sampling is repeated for 100 times. CCs for group 2 are the average of these 100 tests.

^{**} This number is calculated as the CC of group 1 minus the CC of group 2.

In Table 7.1, for both group 1 and group 2 the more references are removed, the smaller the CC is. This result is reasonable because for the reduced citation networks, the number of nodes (the number of publications in the dataset) is unchanged, but by filter out the less essential references the number of edges (citation relations between publications) is reduced.

So if more references are removed, the reduced citation networks will be more loosely connected.

The last column of Table 7.1 is calculated as CC of group 1 minus CC of group 2. So from this column we can know if CC of group 1 is greater than that of group 2 and how big this difference is. The result shows that beside of the first pair of citation networks in which only 10% of less important references are filtered out, for all the other pairs of citation networks the CC of group 1 citation networks is always greater than that of group 2. The difference between group 1 and group 2 is increasing as more references are filtered out. When 60% of references are removed, the difference reaches its peak.

The result means that in most of the cases, citation networks in group 1 are more closely connected than citation networks in group 2. The number of nodes and the number of edges are the same for citation networks in group 1 and citation networks in group 2. But the citation networks in group 1 are always more closely connected than that in group 2. Although this result cannot directly indicate that subgroups are more distinct in group 1 citation networks, this is the most reasonable way to explain this phenomenon.

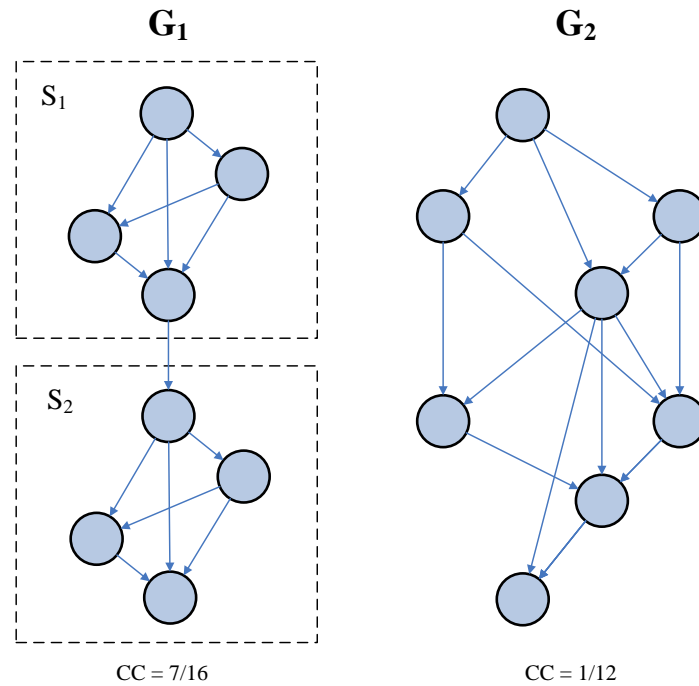


Figure 7.5: CC of graphs with subgroups and without subgroups

As it is shown in Figure 7.5, these two directed graphs, G_1 and G_2 , have the same size: 8 nodes and 13 edges. In G_1 , there are two distinct subgroups: S_1 and S_2 . In G_2 , no subgroup can

be identified. When the CC of G_1 and G_2 is calculated, we can see that the CC of G_1 (7/16) is significantly larger than that of G_2 (1/12). This is due to the fact that in G_2 the clustering coefficient for each node is very small. So the CC of the entire graph, which is calculated as the arithmetic average of the clustering coefficients of all the nodes in the graph, is also small. But in G_1 , the CC of each subgroup is very high. Although there is only one link between the subgroups, the CC for the entire graph is still high.

Based on this reason, we infer that subgroups in group 1 citation networks are more distinct than that in group 2 citation networks. This result can, to certain extent, indicates that the main structure of group 1 citation networks is easier to extract and the amount of noise is also reduced.

However, the quality of citation networks is difficult to measure. There are no measures that can directly indicate the quality of a citation network or the amount of noise in the citation network. In our analysis, we only used the clustering coefficient to measure the quality of citation networks, but it only reflects one aspect of the citation network: how closely it is constructed. So for the future improvement, more measurements should be used.

Chapter 8

SUMMARY AND FUTURE RESEARCH

8.1 Summary of the thesis

Scientometrics is the quantitative study of science, which describes the growth, structure, interrelationships and productivity of science. As the output of science, scientific publications link with each other through publication-reference relationships. Citation analysis is an important area of scientometrics which deals with these relationships. Currently in most cases of citation analysis, all the references are treated equally. However as we all know, this is not true because of the commonly existed arbitrary usage of references. Therefore if we come up with a method to identify the most essential references within a publication, then we can eliminate the less essential references from the analysis or give different weights to the references according to their level of importance. To measure the importance of the references, data in the traditional bibliographic database (e.g., Web of Science, Scopus) is not enough but various features that can be used to estimate the reference value are contained by the full text of the scientific papers. Therefore, in this thesis a model for measuring the importance of references based on citation frequency, citing location, treatment, and self-citation, has been presented.

Given the main research question of this thesis “how to measure the importance of references based on the full text of scientific publications”, a general introduction into the topic of this thesis, the field of scientometrics, and citation analysis were provided in Chapters 1 and 2. The research justification and objective were also presented.

In Chapter 3, after we defined the term “reference importance”, a literature review of the citing features that can be extracted from the full text of publications has been presented. The citing features that have been discussed in detail include citation frequency, location, treatment level, and self-citation. Compared with other features, the citing frequency has the strongest relationship with the reference value.

Based on these citing features and their relationships with the reference value, in Chapter 4 a model that can estimate the importance of references has been introduced. This model takes

the full text of publications as input, and predicts the importance of references after examining the four features of each reference.

Chapter 5 described how to use this model to calculate the reference importance. Firstly, the structure of the full text of publications was introduced. After that, two datasets, a dataset containing publications from the *Journal of Informetrics* (JOI) and a dataset containing publications from the field of renewable energy, have been described. A section classification method has also been introduced in this chapter. Finally, the reference importance of all the references in the JOI dataset has been presented.

In Chapter 6, the reference importance model was validated by the author-rated importance of references. To collect the authors' rank of each reference, an online survey was employed. Based on the 111 responses from authors, the F-measure, precision, and recall of this new model were calculated. The validation result shows that the model we proposed in Chapter 4 can indeed predict the importance of references to a certain extent. Then by making use of this survey result, weights of the model were optimized using logistic regression.

As it is mentioned before, the reference value can be used to improve citation analysis, so in Chapter 7 we try to reduce the noise in citation networks by removing the less essential references based on their reference importance. Besides constructing and analyzing visualizations, an experiment, which was based on the renewable energy dataset, was designed to evaluate the quality of the reduced citation networks based on the average clustering coefficient. These citation networks which are reduced based on the reference importance, are always more closely connected than the citation networks which are reduced randomly.

8.2 Limitations and future research

In this section we discuss some limitations of the research presented in this thesis and we provide some suggestions for further research.

Firstly, the overall quality of the reference importance model that has been described in Chapter 4 is largely depended on the patterns between individual citing features (e.g., frequency, location) and the importance of the reference. These patterns have been summarized in Chapter 3. But currently there are only a few studies about these patterns. Most of the studies that have been done only reported the qualitative results, and from the

previous studies we can hardly find enough quantitative descriptions about these patterns. Because of this situation, when we transformed these relationships into a mathematical model in Chapter 4, it unavoidably introduced some assumptions and inaccuracies. Therefore, future studies are advocated to look at the more detailed quantitative relationship between various citing features and the importance of references.

Secondly, in this research all the citing features that were selected as the indicators of reference importance are at the syntactic level. As has been described before, all the citing features can be classified into two levels: the syntactic level and the semantic level. The features that are used in this model (frequency, location, treatment level, and self-citation) are concerned with the structural aspect of publications, so they belong to the syntactic level. Because compared with the syntactic level features, the semantic level features (e.g., motivation and function of citation) are much more difficult to process with computers, so in this research only syntactic level features were selected. However, the semantic level features are also good predictors of the reference importance, so for further improving this model the semantic level features should be included.

Thirdly, this study only uses the data from the full text of publications to calculate the reference value. Future work can try to integrate this internal data (full text) with external data (e.g., citation score, reputation of the author), to see whether a more comprehensive and accurate estimation of the importance of references could be made.

Additionally, in this research only articles published by Elsevier have been analyzed. This is because the full text files were downloaded through the API of Elsevier. This led to an uneven coverage of journals in certain subject. For example, in Chapter 7 we try to generate the citation network for the major journals about renewable energy, but actually we only constructed the networks for major renewable energy journals that are published by Elsevier. This uneven coverage of journals possibly affected the evaluation of the results. So if it is possible, the future studies should combine data from different publishers together.

Finally, during this research the new proposed reference importance model has been applied to the *Journal of Informetrics* dataset and the Renewable Energy dataset. After analyzing the results, we concluded that this model can predict the importance of references to a certain extent. Different disciplines may have different characteristics, so we are not sure if this conclusion can be extended to other disciplines. There is not sufficient proof to show this

model is flexible enough to handle all these difference. Therefore, future research is needed to test the model's applicability in different disciplines.

References

- Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics*, 56(2), 235-246.
- Bonzi, S. (1982). Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science*, 33(4), 208-216.
- Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, 233-240.
- Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, 7(3), 583-592.
- Dubois, B. L. (1988). Citation in biomedical journal articles. *English for Specific Purposes*, 7(3), 181-193.
- Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*: Elsevier Science Publishers.
- Fowler, J. H., & Aksnes, D. W. (2007). Does self-citation pay? *Scientometrics*, 72(3), 427-437.
- Garfield, E. (1965). Can citation indexing be automated. In *Statistical association methods for mechanized documentation*, 189-192.
- Garfield, E. (1974). Citation Index As A Subject Index. *Current Contents*, 18, 5-7.
- Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science*, 29(6), 308-310.
- Hirsch, J. E. (2005). 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), 16569-16572.
- Hou, W. R., Li, M., & Niu, D. K. (2011). Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. *BioEssays*, 33(10), 724-727.
- Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics*, 7(4), 887-896.
- Kemper, A. (2010). *Valuation of network effects in software markets: A complex networks approach*: Springer.
- Lievers, W., & Pilkey, A. (2012). Characterizing the frequency of repeated citations: The effects of journal, subject area, and self-citation. *Information Processing & Management*, 48(6), 1116-1123.
- Maričić, S., Spaventi, J., Pavičić, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. *Journal of the American Society for Information Science*, 49(6), 530-540.

Marshall, G. (2005). Critiquing a research article. *Radiography*, 11(1), 55-59.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social studies of science*, 5(1), 86-92.

Nalimov, V. V., & Mulchenko, Z. (1969). Naukometriya. Izuchenie Razvitiya Nauki kak Informatsionnogo Protsessa.[Scientometrics. Study of the Development of Science as an Information Process]. Moscow: Nauka.(English translation: 1971. Washington, DC: Foreign Technology Division. US Air Force Systems Command, Wright-Patterson AFB, Ohio.(NTIS Report No. AD735-634)) cited by: Wilson, Conception S.(1999). *Informetrics. Annual Review of Information Science and Technology*, 34, 107-247.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5), 297-312.

Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *EPL (Europhysics Letters)*, 78(3), 30002.

Small, H. G. (1978). Cited documents as concept symbols. *Social studies of science*, 8(3), 327-340.

Sollaci, L. B., & Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the medical library association*, 92(3), 364.

Swales, J. (1990). *Genre analysis: English in academic and research settings*: Cambridge University Press.

Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing & Management*, 28(1), 1-3.

Tang, R., & Safer, M. A. (2008). Author-rated importance of cited references in biology and psychology publications. *Journal of Documentation*, 64(2), 246-272.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 103-110.

van Eck, N. J. (2011). *Methodological advances in bibliometric mapping of science*. (Doctoral dissertation), Erasmus Research Institute of Management (ERIM). (EPS-2011-247-LIS)

van Eck, N. J., & Waltman, L. (in press). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*. arXiv:1404.5322.

Virgo, J. A. (1977). A statistical procedure for evaluating the importance of scientific papers. *The Library Quarterly*, 415-430.

Voos, H., & Dagaev, K. S. (1976). Are All Citations Equal? Or, Did We Op. Cit. Your Idem? *Journal of Academic Librarianship*, 1(6), 19-21.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5(1), 37-47.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440-442.

Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011). Citation count prediction: Learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 1247-1252.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (in press). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*.

Zunde, P. (1971). Structural models of complex information sources. *Information storage and retrieval*, 7(1), 1-18.