Features of scientific papers and the relationships with their citation impact

Tian Yu and **Guang Yu**
School of Management,
Harbin Institute of Technology,
Harbin, 150001, PEOPLE'S REPUBLIC OF CHINA
e-mail: yutian.hit@gmail.com, yug@hit.edu.cn

ABSTRACT

It has been proven that several features of scientific papers are relevant to citation impact. The purpose of this paper is to evaluate the role of these features and unravel which features have greater influence on citation impact. A feature space is established to describe four types of scientific papers' features: features of a paper itself, features of authors, features of published journal, and features of citations. For a group of 676 articles published in 12 journals in the subject category of Information Science & Library Science (IS&LS) in 2007, we analyze quantitatively the difference among high-, medium-, and low-cited papers, and capture their influence on citation impact. The results make it clear that among these four feature types, the quality of a paper and the reputation of authors are the most and the least significant factor affecting the citation impact respectively, and a paper itself has greater influence than the published journal. The findings lay the foundation for the citation impact prediction by using the features of scientific papers.

Keywords: Feature space; Scientific papers; Journal articles; Citation studies; Citation impact.

INTRODUCTION

It is a known fact that citation distribution is highly skewed. For the huge majority of the scientific papers published, the number of citations to them is very low. Some even have never been cited; while some papers garner a lot of citations. Literature has pointed out that there are a few conditions that affect citation to a scientific paper. This study focuses on four conditions or type of factors; namely author characteristic, journal characteristic, research field characteristic and article characteristic.

***Author characteristic*:** A number of bibliometrics studies have shown that the number of citations to scientists' publications is correlated with scientists' impact or influence, such as scientists' prestige (Stewart 1983; Cole 1989; Simonton 1992; Smith and Eysenck 2002; Aksnes and Taxt 2004; Bornmann and Daniel 2005) and academic rank (Cole and Cole 1972;

Danell 2011). A scientist who has a significant influence in his scientific community easily receives more citations than the early career scientist. Furthermore, some interesting author characteristics are assumed to be the determinants of the allocation of citations. Some researchers have found that female scientists' publications receive the citations, on an average, less than that of male scientists (Prpić 2002; Penas and Willett 2006). Some researchers have shown that internationally co-authored papers receive more citations than domestic papers (Narin et al. 1991; Katz and Hicks 1997; Glänzel 2001)

Characteristic of journal in which an article is published: The prestige and influence of journals (and very often their editors) would affect the citations to papers published in them. Some researchers have proven that articles published in core journals receive considerably more citations than articles in lower-tier journals, and the majority of articles in the low-tier journals remain uncited in the five years following their publication (Moed et al. 1985; Van Dalen and Henkens 1999, 2001; Bornmann and Daniel 2005; Boyack and Klavans 2005). Furthermore, journal accessibility and visibility may influence the probability of citations (Vinkler 1987; Yue and Wilson 2004; Xia et al. 2011).

***Research field characteristic*:** Garfield (1979) underlined that "citation potential" can vary significantly from one research field to another. In some fields, the researchers cite recent literature more frequently than in others. Some studies have proven that the citation potential between different research fields using statistical analysis (Guerrero-Bote et al. 2007; Radicchi et al. 2008; Lillquist and Green 2010; Radicchi and Castellano 2012). Recently citation potential has shown to vary not only between research fields or disciplines, but also between the subfields within the same field (Klamer and Van Dalen 2002; Moed 2010).

***Characteristic of the article itself*:** Researchers have explored whether the external characteristics of scientific papers such as the language (Portes 1998; Van Dalen and Henkens 2001), the length (Stewart 1990; Baldi 1998), the level of inter-disciplinarity (Larivière and Gingras 2010), and the article type (Bott and Hargens 1991; MacRoberts and MacRoberts 1996; Costas et al. 2010) could affect citation counts. It has also been found that articles cited in patents are more likely to be cited by other papers (Meyer et al. 2010). Our previous studies have found that publication delay of scientific papers in the publication process will reduce the probability of citations (Yu, Wang and Yu 2004; Yu, Yu and Li 2005). Besides the external characteristics of scientific paper, it is noted that the quality of scientific paper is one of the most important factor for citation impact (Glänzel et al. 2003; Van Dalen and Henkens 2005; Wang et al. 2011).

A quantitative evaluation of the contribution of various features to citation impact is without a doubt important at a number of levels. For the authors it may be important because the evaluation could guide the authors to select the appropriate journal for publishing their research outputs. Whereas publishers and journal editors could discern reliably a trend of the influence of their journals, as well as guide the journal management.
It has been proven that the four types of characteristics, the author, the published journal,

the research field and the article itself, are relevant to citation impact. But which of these characteristics of scientific articles have greater influence on citation impact? In this paper we will evaluate the role of the four features of scientific papers and unravel which features have greater influence on citation impact. We will address this research question in detail, focusing on the scientific papers published in the subject of Information Science and Library Science (IS&LS).

METHOD

The Feature Space of Scientific Papers

Scientific papers can be described as a vector collection of multi-dimensional information such as references, authors and research field. In other words, this information is the multi-dimensional features of papers. The feature space X of scientific papers can be defined as:

$$X=\{x_0,x_1,x_2,x_3,...,x_n\}$$

where x_i ($i=0,1,2,...,n$) is the feature of papers. We can use the features to describe papers' author, references, published journal, publication date, and the institution, region and country the paper is affiliated to. This has laid the foundation for our research.

In this study, the features are divided into four types: features of the paper itself, features of the authors, features of the published journal, and features of the citations. Other external features, such as the paper type, the language, the publication date and the number of references are used to describe the paper itself. The information on the past performance of authors (i.e. their previous publication productivity and citation impact) is used to describe the authors' influence. Journal indicators in the Journal Citation Report (JCR) are used to describe the prestige and influence of the published journal. Furthermore, features of the citations are used to characterize the paper's quality. Previous studies have shown that a paper's quality could be approximated by the impact and speed with which knowledge is disseminated in the scientific community (Van Dalen and Henkens 2005). Citations could reveal the impact of a paper in the literature and the speed with which the paper is disseminated in the scientific community could be measured by the timing of the first citation.

The features listed in Table 1 are extracted to describe the scientific papers. Those features are simple indictors which are easily accessible from the source used in this study: the Web of Science (WoS) (developed by the Institute of Scientific Information (ISI) and maintained by Thomson Reuters) through its various citation indexes. Note that the title of the paper is only for sample labeling, and it is not practically significant.

Table 1: The Features of Scientific Papers

Features		Label
The total number of citations (it is used to represent the citation impact)		x_0
Features of the paper itself	The title (it is ignored for no practical significance)	
	The year when published (all were published in 2007)	
	The kind (the kind of each selected paper is article)	
	The number of references listed	x_1
Features of the authors	The number of authors	x_2
	The country of author's institution	x_3
	The *h* index of the first author before publication of this paper	x_4
	The number of papers published by the first author before this paper	x_5
	The total citations to the papers published by the first author before this paper	x_6
	The average citations to the paper published by the first author before this paper	x_7
	The maximum *h* index of the authors before publication of this paper	x_8
	The maximum number of papers published by the authors before this paper	x_9
	The maximum total citations to the papers published by the authors before this paper	x_{10}
	The maximum average citations to the paper published by the authors before this paper	x_{11}
Features of the citations	The *h* index of the citing articles	x_{12}
	The first-cited age of this paper	x_{13}
	The total citations to this paper in its first 2 years after publication	x_{14}
	The number of countries citing this paper in its first 5 years after publication	x_{15}
	The number of types of papers citing this paper in its first 5 years after publication	x_{16}
	The number of journals citing this paper in its first 5 years after publication	x_{17}
	The number of subjects citing this paper in its first 5 years after publication	x_{18}
Features of the published journal	The total citations	x_{19}
	The impact factor	x_{20}
	The 5-year impact factor	x_{21}
	The immediacy index	x_{22}
	The number of articles published	x_{23}
	The number of citations per paper	x_{24}
	The cited half-life	x_{25}
	The Eigenfactor score	x_{26}
	The article influence score	x_{27}

Data Preparation and Collection

By using WoS, we identified the features of 676 papers published in 2007 in 12 journals from IS & LS category (listed in Table 2). We selected one category only, taking into account the difference among topics on the probability of being cited. We confined the citation

data to only citations garnered up to 2011. For practical reason, the month when the paper is published is ignored. To make it more convenient for comparing, we only selected the papers whose type is 'article'. We used the country of the first corresponding author as the country of author's institution. It should be noted that this is the text feature, which could not be analysed directly, therefore we manually converted the text data into numerical data. Values of this feature are assigned based on the order of the presence of countries, and the same country has the same value on this feature. In addition, we need to exclude articles published in 2007-2011 (in journals apart from those listed in Table 2) from the author's all publications in order to identify the features of the author before publication of the paper.

Table 2: The 12 Journals from Information Science and Library Science category

Num.	Abbreviated Journal Title	ISSN	No of articles
1	INFORM SYST RES	1047-7047	21
2	INFORM MANAGE-AMSTER	0378-7206	52
3	INFORM SYST J	1350-1917	17
4	INFORM PROCESS MANAG	0306-4573	106
5	J AM SOC INF SCI TEC	1532-2882	176
6	INFORM RES	1368-1613	82
7	COLL RES LIBR	0010-0870	32
8	GOV INFORM Q	0740-624X	41
9	INFORM SOC	0197-2243	16
10	J ACAD LIBR	0099-1333	68
11	INT J INFORM MANAGE	0268-4012	31
12	ASLIB PROC	0001-253X	34

RESULTS AND DISCUSSIONS

The accumulated total number of citations to those 676 articles published in the 12 journals from 2007 to 2011 is 4173. Figure 1 shows that citations are skewed in the distribution of 676 articles on the total number of citations x_0, which conforms to the overall situation in the category of IS&LS. It implies that the data we selected are valid. Based on the number of citations we classify the articles into three types: high-, medium-, and low-cited articles. The judgment of the types of articles follows the 80/20 method (Bradford 1985). We obtained 13 high-cited articles which are cited more than 39 times, and 220 medium-cited articles which are cited more than 5 times, and 443 low-cited articles. Figure 2 illustrates the distribution of the three types of articles published in these 12 journals.

We also collected a number of explanatory variables and the descriptive statistics of these variables are presented in Table 3. We have calculated respectively the mean and standard deviation of 27 features of the 676 articles. Note that x_3 which is originally a text data, has been removed from the calculation of the mean and standard deviation. It is noted that the reciprocal of the first-cited age takes the place of the first-cited age in this study. The reason is that some papers have never been cited in WoS. In order to facilitate comparison,

the first-cited age of these papers could be defined as positive infinity. Therefore the reciprocal of the first-cited age could be in the range of 0 – 1.

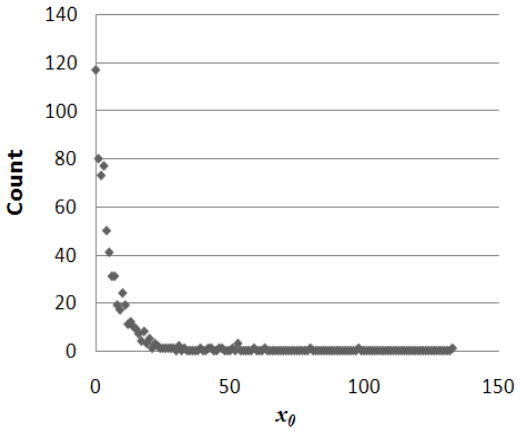

Figure 1: Distribution of 676 articles on the Total Number of Citations x_0

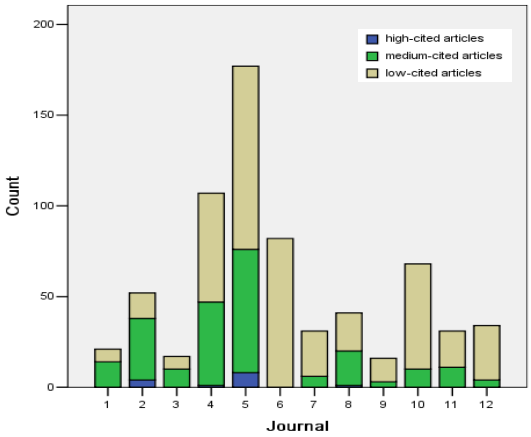

Figure 2: Distribution of Three Types of Articles Published in 12 Journals

Table 3: Mean and Std. Deviation of the Features for the Three Article Types

Features	Mean/Std. Deviation			
	High-cited articles	Medium-cited articles	Low-cited articles	All articles
x_1	45.2/22.5	44.0/21.3	30.3/20.1	35.2/21.6
x_2	2.0/0.8	2.6/1.3	2.1/1.3	2.3/1.3
x_4	5.0/5.8	2.4/3.7	1.8/2.9	2.1/3.3
x_5	11.6/16.3	5.2/10.0	5.0/11.5	5.2/11.1
x_6	216.2/376.1	87.6/233.2	45.0/151.5	62.7/191.0
x_7	24.6/22.2	9.7/15.6	5.2/12.1	7.1/14.0
x_8	9.1/7.2	4.8/5.1	3.0/4.1	3.7/4.7
x_9	22.1/22.9	11.3/14.5	8.9/15.9	10.0/15.7
x_{10}	554.1/761.2	209.6/404.8	104.3/281.5	148.6/350.2
x_{11}	29.2/24.4	17.9/34.2	8.6/22.3	12.1/27.3
x_{12}	8.9/4.6	2.4/1.4	0.7/0.7	1.4/1.8
$1/x_{13}$	0.8/0.3	0.6/0.3	0.3/0.3	0.4/0.3
x_{14}	12.5/7.6	2.3/1.9	0.5/0.7	1.3/2.5
x_{15}	18.4/7.4	6.9/3.0	1.6/1.5	3.7/4.0
x_{16}	3.7/1.3	2.3/0.8	1.0/0.8	1.5/1.1
x_{17}	36.2/16.0	9.6/4.2	1.9/1.6	5.1/6.8
x_{18}	23.4/11.9	8.2/4.2	2.2/2.2	4.6/5.2
x_{19}	2369.3/886.5	1704.5/1055.5	1169.6/1128.0	1372.2/1136.1
x_{20}	1.5/0.2	1.4/0.5	1.1/0.5	1.2/0.5
x_{21}	2.0/0.5	2.0/1.3	1.4/0.8	1.6/1.1
x_{22}	0.3/0.1	0.2/0.2	0.2/0.1	0.2/0.2
x_{23}	132.8/65.4	99.0/63.9	85.2/61.0	90.8/62.6
x_{24}	20.1/9.2	21.2/20.4	13.0/12.0	15.9/15.7
x_{25}	7.0/1.1	7.0/1.1	6.3/1.6	6.6/1.5
$x_{26}*10^3$	8.6/3.6	5.2/5.0	2.8/4.5	3.7/4.8
x_{27}	0.7/0.2	0.7/0.6	0.5/0.4	0.6/0.5

Overall, the mean value of high-cited articles for most of these features is the highest in all three types, and for low-cited articles it is the lowest. However two features, namely x_2 (the number of authors) and x_{24} (The number of citations per paper), are not consistent with this situation. The mean value of x_2 for high-cited articles is the lowest in three types, and that for medium-cited articles is the highest (This result may be explained by the influence of the number of authors which does not seem to be significant). We also get the interesting result on another feature x_{24}, which is defined as x_{23} (the number of articles published) divided by x_{19} (the total citations). The mean value of x_{24} for medium-cited articles is slightly higher than that for the high-cited articles. It is probably due to the fact that the high-cited articles were published in only 4 journals in our dataset (INFORM MANAGE-AMSTER, INFORM PROCESS MANAG, J AM SOC INF SCI TEC, and GOV INFORM Q). In other words, the mean of x_{24} for medium-cited articles is similar with that for high-cited articles.

Relationships between Citation Impact x_0 and the other Features of Articles

In order to capture features of scientific articles in affecting the citation impact, correlation analysis between the features is used to describe the links between two variables, and it reflects how much one variable changes when the value of another variable is controlled. Here Spearman correlation coefficient is used to measure the relationship between citation impact x_0 and the other features of articles because of the distribution of these data in this research.

(a) Features of the paper itself

Table 4 illustrates Spearman correlation coefficient between x_0 and x_1.

Table 4: Correlation Matrix of the Features of Authors

	Correlation between Vectors of Values	
	x_0	x_1
x_0	1.000	0.406**
x_1	0.406**	1.000

This is a symmetric matrix. **Significant at the 0.01 level.

The number of references listed (x_1). Overall, the majority of the articles have a range of 10-50 references. From the viewpoint of the mean value, high-cited articles have more references. Surprisingly, based on the result of correlation analysis the correlation coefficient between the total number of references and citation impact is close to 0.5, which is a relatively high value. It seems to be that the number of references could influence the citation impact. This is probably a consequence of reading a lot of literature. The more literature a researcher reads, more deeply he understands the current situation and development trend of his research field. This is an effective method to enhance the quality of their research.

(b) Features of the Authors

Table 5 illustrates Spearman correlation coefficient between x_0 and the authors' features.

The result shows that a strong correlation is observed between any two of eight features describing authors' prestige and influence. But the correlations between the number of authors x_3 and the other features are very low.

Table 5: Correlation Matrix of the Features of Authors

	Correlation between Vectors of Values										
	x_0	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_0	1.000	0.212**	0.007	0.175**	0.082**	0.218**	0.258**	0.287**	0.190**	0.321**	0.369**
x_2	0.212**	1.000	0.081*	0.071	0.081*	0.096*	0.118**	0.369**	0.386**	0.406**	0.377**
x_3	0.007	0.081*	1.000	0.006	0.032	0.001	-0.034	0.081*	0.125**	0.060	-0.016
x_4	0.175**	0.071	0.006	1.000	0.881**	0.960**	0.885**	0.675**	0.596**	0.633**	0.573**
x_5	0.082**	0.081*	0.032	0.881**	1.000	0.838**	0.703**	0.586**	0.654**	0.542**	0.413**
x_6	0.218**	0.096*	0.001	0.960**	0.838**	1.000	0.958**	0.665**	0.579**	0.672**	0.651**
x_7	0.258**	0.118**	-0.034	0.885**	0.703**	0.958**	1.000	0.625**	0.490**	0.657**	0.710**
x_8	0.287**	0.369**	0.081*	0.675**	0.586**	0.665**	0.625**	1.000	0.900**	0.949**	0.794**
x_9	0.190**	0.386**	0.125**	0.596**	0.654**	0.579**	0.490**	0.900**	1.000	0.848**	0.603**
x_{10}	0.321**	0.406**	0.060	0.633**	0.542**	0.672**	0.657**	0.949**	0.848**	1.000	0.895**
x_{11}	0.369**	0.377**	-0.016	0.573**	0.413**	0.651**	0.710**	0.794**	0.603**	0.895**	1.000

This is a symmetric matrix.
*Significant at the 0.05 level. **Significant at the 0.01 level.

The number of authors (x_2). We find that more than 80% of all articles have between 1-3 authors in the histogram of x_2. In Table 3, the mean value of x_2 for high-cited articles is the lowest in all three types, and x_2's mean value for medium-cited articles is higher than the low-cited articles. It seems that the number of authors for the IS&LS category sampled is not an important factor to increase the probability of being cited based on the correlation matrix. Leimu and Koricheva (2005) found that ecological papers with four or more authors received more citations than those with fewer authors. It may be due to the difference between these two disciplines, ecology and IS&LS.

The country of author's institution (x_3). The authors of 284 articles, about 42% of all articles, come from American institutions. The authors from England, South Korea, Canada and Spain also published a great number of articles in 2007. Among 13 high-cited articles, the authors of two articles come from institutions in America, four from England, and two from South Korea. The distribution of x_3 for high-cited articles is basically similar to that for all articles.

The first author's h index (x_4), the number of papers (x_5), the total citations (x_6), and the average citations per article (x_7) before publication of the paper: These four features indicate the prestige and influence of the first author. In our data, the value of x_4 for about 40% of all articles are zero; the value of x_5 for about 60% of articles are not more than 2; the value of x_6 for about 80% of articles are lower than 50; the value of x_7 for over 70% of articles are lower than 7. It implies that about half of all researchers in the field are new and their prestige is very low. Similarly, Levitt and Thelwall (2009) found a high percentage of the first authors with relative lower *h* index in the field of IS&LS. With increase of citation counts, the mean and standard deviation of x_4, x_5, x_6 and x_7 have a substantial rise. Further, Spearman correlation coefficients between citation impact and these four

features are only about 0.2, which indicates the weak relationship between the first author's prestige and citation impact. The result confirms that the effect of author reputation is quite small (Van Dalen and Henkens 2005). In addition, among these features, x_5 has the lowest influence on citation counts. It means that increasing the number of papers published is not enough to improve citation impact, and it is significant to enhance the quality of their papers.

The authors' maximum h index (x_8), the maximum number of papers (x_9), the maximum total citations (x_{10}), and the maximum average citations per article (x_{11}) before publication of the paper: The four features are similar with the features of the first author, but there is still little difference among them. These features indicate the maximum prestige and influence of the authors. Based on the correlation matrix, the effect of these four features is higher than that of the features indicated the first author's reputation, which means the author with the highest reputation could influence the probability of citations. This may be the reason why one wants to collaborate with researchers having high reputation.

(c) Features of the Published Journal

The result of Spearman correlation coefficient between x_0 and the features of published journal is shown in Table 6. The correlations between the nine features of published journal are all high.

Table 6: Correlation Matrix of the Features of Published Journal

	Correlation between Vectors of Values									
	x_0	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}
x_0	1.000	0.353**	0.366**	0.366**	0.281**	0.184**	0.365**	0.334**	0.318**	0.346**
x_{19}	0.353**	1.000	0.644**	0.807**	0.787**	0.747**	0.790**	0.649**	0.847**	0.889**
x_{20}	0.366**	0.644**	1.000	0.887**	0.670**	0.337**	0.726**	0.537**	0.499**	0.801**
x_{21}	0.366**	0.807**	0.887**	1.000	0.779**	0.416**	0.888**	0.439**	0.799**	0.942**
x_{22}	0.281**	0.787**	0.670**	0.779**	1.000	0.684**	0.640**	0.421**	0.639**	0.729**
x_{23}	0.184**	0.747**	0.337**	0.416**	0.684**	1.000	0.227**	0.358**	0.571**	0.456**
x_{24}	0.365**	0.790**	0.726**	0.888**	0.640**	0.227**	1.000	0.535**	0.754**	0.942**
x_{25}	0.334**	0.649**	0.537**	0.439**	0.421**	0.358**	0.535**	1.000	0.296**	0.512**
x_{26}	0.318**	0.847**	0.499**	0.799**	0.639**	0.571**	0.754**	0.296**	1.000	0.847**
x_{27}	0.346**	0.889**	0.801**	0.942**	0.729**	0.456**	0.942**	0.512**	0.847**	1.000

This is a symmetric matrix. **Significant at the 0.01 level.

The number of articles (x_{19}), the impact factor (x_{20}), the 5-year impact factor (x_{21}), and the immediacy index (x_{22}), the total citations (x_{23}), the number of citations for each paper (x_{24}), and the cited half-life (x_{25}): These features are the important indicators to evaluate the size and influence of the journal in a certain period. The result of mean analysis shows that the highly cited articles are usually published in journals which have high value of the features. In addition, Spearman correlation coefficients between citation impact and these features are all about 0.35 (except for x_{19}), which have higher influence on citation impact than the features of authors. In other words, compared with author reputation, the

published journal reputation makes more significant contributions to citation impact. For authors, it is able to explain their strong wish to publish in core journals. Whereas for publishers, expanding journals' scale is not enough to improve their reputation.

The Eigenfactor score (x_{26}) and the article influence score (x_{27}): The two features evaluate the importance of the journal based on the whole citation network. The result shows that the high-cited articles are usually published in the journals which have a high value of x_{26} in Table 6. But the mean of x_{27} for medium-cited articles is significantly higher than that for high-cited articles. It means that when studying the whole citation network, we can not only focus on the high-cited articles. And it may be of great significance to properly study the medium-cited articles.

(d) Features of the Citations

We analyze Spearman correlation coefficient between the features of citations. It is observed that these features of citations are significantly associated with the number of citations in Table 7.

Table 7: Correlation Matrix of the Features of Citations

	Correlation between Vectors of Values							
	x_0	x_{12}	$1/x_{13}$	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_0	1.000	0.821**	0.744**	0.716**	0.932**	0.834**	0.974**	0.868**
x_{12}	0.821**	1.000	0.667**	0.648**	0.770**	0.685**	0.784**	0.690**
$1/x_{13}$	0.744**	0.667**	1.000	0.870**	0.699**	0.708**	0.744**	0.678**
x_{14}	0.716**	0.648**	0.870**	1.000	0.657**	0.636**	0.717**	0.638**
x_{15}	0.932**	0.770**	0.699**	0.657**	1.000	0.813**	0.927**	0.861**
x_{16}	0.834**	0.685**	0.708**	0.636**	0.813**	1.000	0.840**	0.817**
x_{17}	0.974**	0.784**	0.744**	0.717**	0.927**	0.840**	1.000	0.898**
x_{18}	0.868**	0.690**	0.678**	0.638**	0.861**	0.817**	0.898**	1.000

*This is a symmetric matrix. **Significant at the 0.01 level.*

The h index of the citing articles (x_{12}): The *h* index of the citing articles means that the *h* number of the citing articles which received at least *h* citations (Araújo and Sardinha 2011). It has been recognized as an indicator to measure the impact of scientific papers. Generally, x_{12} is closely related with x_0. The higher the total number of citations, the bigger the *h* index of the citing articles.

The first-cited age of the paper (x_{13}), and the total citations to the paper in its first 2 years after publication (x_{14}): More than 50% of all articles were firstly cited in their first 2 years after publication, and about 75% were firstly cited in their first 3 years. Eight of 13 high-cited articles were firstly cited in their first year after publication, and the other 6 were firstly cited in their first 2 years (most of these articles were published at the end of the first year). It means that high-cited articles have strong capability to be cited in their first 2 years after publication. The result of Spearman correlation also implies the strong relationship between x_{13} and citation impact. In addition, the mean of x_{14} for high-, medium-, and low-cited articles are respectively 12.5, 2.3, and 0.5 in Table 3. The Spearman correlation coefficient between citation impact and the feature is also over 0.7.

Consequently it indicates that x_{14} could accurately reflect the citation impact. Previous study has also shown that the accepted high-quality papers have good capacity of knowledge diffusion in the period of the first-cited year after publication (Glänzel et al. 2003).

***The number of countries (x_{15}), the number of types of papers (x_{16}), the number of journals (x_{17}), and the number of subjects (x_{18}) citing the paper in its first 5 years after publication*:** The result of correlation analysis shows that these four features are closely related with the citation impact x_0. The higher the total numbers of citations is, the bigger the value of these four features are. Previous studies have shown that a five-year interval after publication is a sufficient term to distinguish highly cited papers from the other papers (Glänzel et al 2003; Aksnes 2003; Wang et al 2011). Our result also confirms that the features of citation are the best predictors of citation frequency.

CONCLUSIONS

In this paper, we have established the feature space of scientific papers with a mathematical description and analyzed quantitatively the features of scientific papers. We have also obtained the result of the correlations between citation impact and these features. We have examined the role played by four types of features in assessing the influence on citation impact: features of the paper itself, features of the authors, features of the published journal, and features of the citations.

To summarize our findings succinctly we can state the following four conclusions. First, the quality of scientific papers that could be approximated by the features of citations is the most significant factor affecting the citation impact. Similar conclusions have also been suggested by Van Dalen and Henkens (2001, 2005). Second, external features of a paper itself are the important factors that affect the citation impact. In this study, some features of a paper itself which several studies have discussed are eliminated, and the number of references is the only one feature that we choose. We have found that the number of references could exert a great influence on the citation impact. Third, the features of authors and published journal are able to affect the citation impact to some extent, but the features of a paper itself have more influence than the authors and journal features. It suggests that for increasing the citation impact of a scientific paper, the author need to not only improve its quality but also offer a better path for its knowledge diffusion. And finally, compared with the features of authors, the features of published journal make more significant contributions to improve the citation impact. Thus it is a good choice for authors to select journals with higher reputation for the submission of their manuscripts.

Several important caveats should tamper these conclusions. Most importantly, the sample of scientific papers included in this analysis is quite limited. It includes 12 journals in one subject and covers the articles published in 2007 only. The analyses assume that this limited sample is a representative of the publication and citation of scientific articles in the

IS&LS category. The data obtained are limited to the articles covered in the ISI database. Some of the limitations of the ISI database itself, such as incompleteness, are bound to be brought into the study. However, it is undeniable that the ISI is the largest comprehensive academic information resource database in the world which covers the most subjects. It is the reason for selecting this database. In addition, we believe that the scientific paper has a multidimensional complex features. In the paper, we only selected the features which is considered available and could be obtained in a relatively convenient and practical manner. That may cause the omissions of some features, and make the results not ideal.

Even with these caveats, the findings of this study still show that there are interesting relationships between the features and citation impact of scientific papers. Based on the effective method for description of scientific papers, we need to further consider the comprehensiveness and effectiveness of the features, involving many aspects of the quality of the paper itself, scientific innovation capability, and acceptable level of the audiences. And the dataset used needs to be larger and more comprehensive.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 70973031).

REFERENCES

Aksnes, D. W. 2003. Characteristics of highly cited papers. *Research Evaluation*, Vol.12, no.3: 159-170.

Aksnes, D. W. and Taxt, R. E. 2004. Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. *Research Evaluation*, Vol.13, no. 1: 33-41.

Araújo, C.G.S. and Sardinha, A. 2011. Index of articles H-citing: a contribution to the evaluation of scientific production of experienced researchers. *Revista Brasileira de Medicina do Esporte*, Vol.17, no.5: 358-362.

Baldi, S. 1998. Normative versus social constructivist processes in the allocation of citations: a network-analytic model. *American Sociological Review*, Vol.63, no. 6: 829-46.

Bornmann, L. and Daniel, H.-D. 2005. Selection of research fellowship recipients by committee peer review. Analysis of reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics*, Vol. 63, no. 2: 297-320

Bott, D.M. and Hargens, L.L. 1991. Are sociologists' publications uncited? Citation rates of journal articles, chapters, and books. *The American Sociologist*, Vol. 22, no. 2: 147-158.

Boyack, K.W. and Klavans, R. 2005. Predicting the importance of current papers, in Ingwersen, P. and Larsen, B. (Eds). *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, Karolinska University Press, Stockholm.

Bradford, S. C. 1985. Sources of information on specific subjects. *Journal of Information*

Science, Vol. 10, no. 4:173-180.

Cole, J.R. and Cole, S. 1972. The Ortega hypothesis. *Science*, Vol. 178, no. 4056: 368-75.

Cole, S. 1989. Citations and the evaluation of individual scientists. *Trends in Biochemical Sciences*, Vol. 14, no. 1: 9-13.

Costas, R., Van Leeuwen, T.N., and Van Raan, A.F.J. 2010. Is scientific literature subject to a 'Sell-By-Date'? A general methodology to analyze the 'durability' of scientific documents. *Journal of the American Society for Information Science and Technology*, Vol. 61, no.2: 329–339.

Danell, R. 2011. Can the quality of scientific work be predicted using information on the author's track record? *Journal of the American Society for Information Science and Technology*, Vol. 61, no.1: 50-60.

Garfield, E. 1979. *Citation indexing. Its theory and application in science, technology and humanities.* New York: Wiley.

Glänzel, W. 2001. National characteristics in International Scientific Co-authorship. *Scientometrics*, Vol. 51, no.1: 69-115

Glänzel, W., Schlemmer, B. and Thijs, B. 2003. Better later than never? On the chance to become highly cited only beyond the standard bibliometric time horizon, *Scientometrics*, Vol. 58, no. 3: 571–586.

Guerrero-Bote, V. P., Zapico-Alonso, F., Espinosa-Calvo, M. E., Go´mez-Criso´stomo, R., and De Moya-Anego´n, F. 2007. Import-export of knowledge between scientific subject categories: The iceberg hypothesis. *Scientometrics*, Vol. 71, no. 3: 423-441.

Katz, J. S., and Hicks, D. 1997. How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, Vol. 40, no. 3: 541-554.

Klamer, A. and Van Dalen, H. P. 2002. Attention and the art of scientific publishing. *Journal of Economic Methodology*, Vol. 9, no. 3: 289-315.

Larivière, V., Gingras, Y. 2010. On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology*, Vol. 61, no. 1: 126-131.

Leimu, R., and Koricheva, G. 2005. What determines the citation frequency of ecological papers?. *Trends in Ecology and Evolution*, Vol. 20, no. 1: 28-32.

Levitt, J. M., and Thelwall, M. 2009. The most highly cited Library and Information Science articles: Interdisciplinarity, first authors and citation patterns. *Scientometrics*, Vol. 78, no. 1: 45-67.

Lillquist, E. and Green, S. 2010. The discipline dependence of citation statistics. *Scientometrics*, Vol. 84, no. 3: 749-762.

MacRoberts, M. H. and MacRoberts, B. R. 1996. Problems of citation analysis. *Scientometrics*, Vol. 36, no. 3: 435-44.

Meyer, M., Debackere, K. and Glänzel, W. 2010. Can applied science be 'good science'? Exploring the relationship between patent citations and citation impact in nanoscience. *Scientometrics*, Vol. 85, no. 2: 527-539.

Moed, H. F., Burger, W. J. M., Frankfort, J. G. and Van Raan, A. F. J. 1985. The use of bibliometric data for the measurement of university research performance. *Research Policy*, Vol. 14, no. 3: 131-49.

Moed, H. F. 2010. Measuring contextual citation impact of scientific journals. *Journal of*

Informetrics, Vol. 4, no. 3: 265-277.

Narin, F., Stevens, K., and Whitlow, E. S. 1991. Scientific cooperation in Europe and the citation of multinationally authored papers. *Scientometrics* Vol. 21, no. 3: 313-323.

Penas, C. S., and Willett, P. 2006. Gender differences in publication and citation counts in librarianship and information science research. *Journal of Information Science*, Vol. 32, no. 5: 480-485.

Portes, A. 1998. Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology*, Vol. 24: 1-24.

Prpić, K. 2002. Gender and productivity differentials in science. *Scientometrics*, Vol. 55, no. 1: 27-58.

Radicchi F and Castellano C. 2012. Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics*, Vol. 6 no. 1: 121-130.

Radicchi, F., Fortunato, S. and Castellano, C. 2008. Universality of citation distributions: toward an objective measure of scientific impact. *PNAS*, Vol. 105, no. 45: 17268-17272.

Simonton, D.K. 1992. Leaders of American psychology, 1879-1967: career development, creative output, and professional achievement. *Journal of Personality and Social Psychology*, Vol. 62, no. 1: 5-17.

Smith, A. and Eysenck, M. 2002. *The Correlation between RAE Ratings and Citation Counts in Psychology.* Department of Psychology, Royal Holloway, University of London, London.

Stewart, J. A. 1983. Achievement and ascriptive processes in the recognition of scientific articles. *Social Forces*, Vol. 62: 166-184.

Stewart, J.A. 1990. *Drifting continents and colliding paradigms: Perspectives on the geoscience revolution*. Indiana University Press, Bloomington, IN.

Van Dalen, H. P., and Henkens, K. 1999. How influential are demography journals?. *Population and Development Review*, Vol. 25, no. 2: 229-251.

Van Dalen, H. P., and Henkens, K. 2001. What makes a scientific article influential? The case of demographers. *Scientometrics*, Vol. 50, no. 3: 455-482.

Van Dalen, H. P., and Henkens, K. 2005. Signals in science-on the importance of signaling in gaining attention in science. *Scientometrics*, Vol. 64, no. 2: 209-233.

Vinkler, P. 1987. A quasi-quantitative citation model. *Scientometrics*, Vol. 12, no. 1-2: 47-72.

Wang, M.Y., Yu, G., and Yu, D.R. 2011. Mining typical features for highly cited papers. *Scientometrics*, Vol. 87, no. 3: 695-706.

Xia, J. F., Myers, R. L., and Wihoite, S. K. 2011. Multiple open access availability and citation impact. *Journal of Information Science*, Vol. 37, no. 1: 19–28.

Yu, G., Wang, X.H., and Yu, D.R. 2004. The influence of publication delays on impact factors. *Scientometrics*, Vol. 64, no. 2: 235-246.

Yu, G., Yu, D. R., and Li, Y. J. 2004. The universal expression of periodical average publication delay at steady state. *Scientometrics*, Vol. 60, no. 2: 121-129.

Yue, W. and Wilson, C.S. 2004. Measuring the citation impact of research journals in clinical neurology: a structural equation modelling analysis. *Scientometrics*, Vol. 60, no. 3: 317-332.